Temporally Consistent Depth Map Estimation Using Motion Estimation for 3DTV

Sang-Beom Lee and Yo-Sung Ho

Department of Information and Communications
Gwangju Institute of Science and Technology
261 Cheomdan-gwagiro, Buk-gu, Gwangju, 500-712 Korea
E-mail: {sblee, hoyo}@gist.ac.kr

ABSTRACT

In this paper, we propose a new algorithm to estimate temporally consistent depth sequence. Our algorithm first calculates the matching cost using left and right views. In order to enhance the temporal consistency, we modify the matching function by adding the temporal weighting function and we perform the motion estimation technique to refer to the previous depths of moving objects. Experimental results have showed that the proposed algorithm improved the temporal consistency of the depth sequence and reduced flickering artifacts in the virtual view while maintaining visual quality.

Keywords: Three-dimensional television, Multiview video, Depth estimation, Temporal consistency

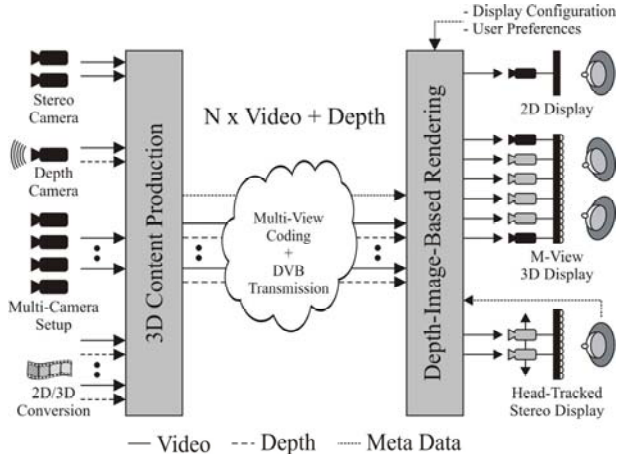


Figure 1. Three-dimensional television system

1. INTRODUCTION

Owing to great advancements in computing power, interactive computer graphics, digital transmission, and immersive displays, we experience and reproduce simulations of reality. In other words, the gap between the real world and the virtual environment is getting closer. When users are exposed to such immersive, interactive, and perceptually realistic media, they report a sense of presence in the mediated environment [1][2]. Especially, technological advances in displays have been aimed at improving the range of vision, such as a wide-screen, high-definition, immersive, and 3D displays.

Recently, a three-dimensional television (3DTV) using multiview video is in the spotlight as one of the next-generation broadcasting services [3]. In order to acquire multiview video, we utilize multiple cameras with parallel or convergent configuration to capture a 3D scene with wide-viewing angle. We believe that the 3DTV is the next-generation broadcasting system in the history of TV. By aiding of advances in display devices, such as stereoscopic displays, the 3DTV provides 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 including content generation, coding, transmission, and display. Figure 1 shows the conceptual framework of the 3DTV system [4].

In general, there are two major problems in multiview video. The first problem is the reliability of distance between cameras. The other problem is the sudden viewpoint change. When users change their views while watching contents through the 3D display, they feel unnatural viewpoint change on the display if the distance between adjacent cameras is large and the scene is changed suddenly. It causes a visual discomfort to viewers' eyes.

In order to reconstruct intermediate views of virtual viewpoints, we need depth information. Many works have been carried out for the acquisition of 3D depth information [5][6]. Recently, 3D video coding subgroup members in Moving Picture Experts Group (MPEG) recognized the importance of multiview video and the corresponding depth sequence, and they tried to develop the depth estimation and the view synthesis tools [7]. As a result, they implemented and distributed the graph cut-based depth estimation software [8].

However, the software includes several problems such as boundary mismatch, textureless regions, or wide baseline. Especially, since the software estimates depth sequence for each frame separately, the depth sequence is temporally inconsistent. In other words, we notice the inconsistent depth values at the same background but in a different time. This problem causes flickering artifacts and it discomforts the users.

Therefore, we propose a new algorithm for enhancing temporal consistency of the depth sequence. The main contribution of this paper is that we add a temporal weighting function to the conventional matching function for consistent depth map estimation. The temporal weighting function refers to the previous depth when estimating the current depth. In order to exploit the temporal weighting function for the moving objects, we reconstruct the depth map from the previous depth map by using motion estimation technique. In addition, we prevent the border error propagation to reduce the depth errors.

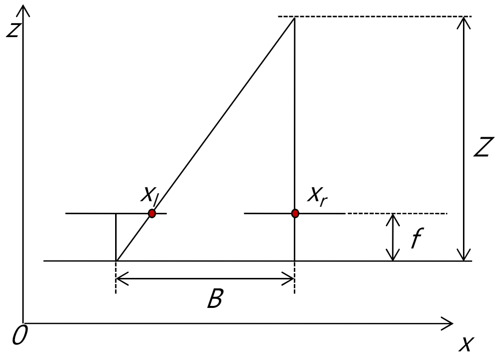


Figure 2. Relationship between disparity and depth

2. MULTIVIEW DEPTH ESTIMATION

2.1 Disparity and Depth

Figure 2 illustrates the relationship between disparity and depth. Suppose that a certain 3D point is projected onto the right image plane and it is located at (x_r,y). This 3D point is also projected onto the left image plane and it is located at (x_l,y). Then, the relationship between disparity d and depth Z can be defined by

$$Z = \frac{Bf}{d} = \frac{Bf}{x_l - x_r} \tag{1}$$

where B represents the camera distance and f represents the focal length of each camera. This equation proves that we can find the depth if we estimate the disparity by using the correspondence of multiview video.

2.2 Disparity Computation

As mentioned before, 3D video coding subgroup members in MPEG tried to develop the depth estimation and the view synthesis tools. As a result, they implemented and distributed the depth estimation and view synthesis softwares. The software is categorized by three parts: disparity computation, graph cut-based error minimization, and disparity-to-depth conversion.

The first step is to aggregate the matching costs for whole pixels of the center view. Since we have three and more views of multiview video, we can compare center view to the left and right view simultaneously. Thus, we can easily handle the occlusion region that is visible in the center view but become invisible in the left or right views. The matching function is defined by

$$E_{sim}(x,y,d) = \min\{E_L(x,y,d), E_R(x,y,d)\} \tag{2}$$

$$E_L(x,y,d) = |I_C(x,y) - I_L(x+d,y)| \tag{3}$$

$$E_R(x,y,d) = |I_C(x,y) - I_R(x-d,y)| \tag{4}$$

where $I(x,y)$ indicates the intensity at the point (x,y).

The second step is graph cut-based error minimization. The optimum disparity value is determined in this step by comparing matching costs of neighbor pixels.

The third step is disparity-to-depth conversion. The depth map can be represented by 8-bit grayscale image with the gray level 0 specifying the farthest value and the gray level 255 defining the nearest value. The metric space between the near clipping plane and the far clipping plane is divided into the same 256 spaces. Then, the depth value Z which corresponds to the pixel (x,y) is transformed into the 8-bit gray value v as follows:

$$v = \left\lfloor 255 - \frac{255(Z - Z_{near})}{Z_{far} - Z_{near}} + 0.5 \right\rfloor \tag{5}$$

where Z_{far} and Z_{near} represent the farthest and nearest depth values. Figure 3 shows the 3D scene limited by two clipping planes, far and near clipping planes.

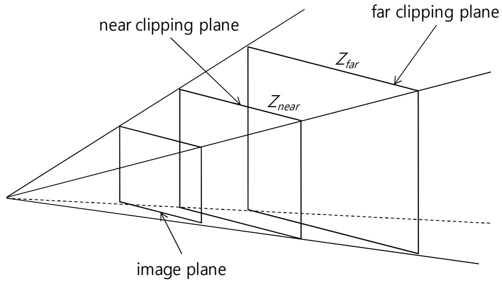


Figure 3. 3D scene limited by two clipping planes

2.3 Sub-pixel Precision

In order to improve the accuracy of the depth map, sub-pixel precision depth estimation is introduced. The left or right views are upsampled by various interpolation techniques, such as bi-linear or bi-cubic filter. Then, we can find the disparity more precisely. The upsampling can be done in half-pixel or quarter-pixel level. Figure 4 shows the sub-pixel precision of depth estimation.

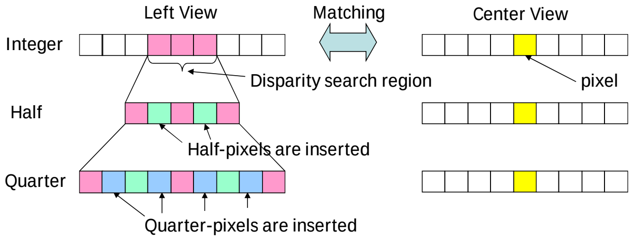


Figure 4. Sub-pel precision of depth estimation

3. TEMPORAL CONSISTENCY ENHANCEMENT

3.1 Temporal Weighting Function using Motion Estimation

As mentioned before, since previous works estimate the depth map for each frame separately, the depth sequence is temporally inconsistent. Therefore, we add a temporal weighting function that refers to the previous depth when estimating the current depth [9]. The temporal weighting function is defined by

$$E_{new}(x,y,d) = E_{sim}(x,y,d) + E_{temp}(x,y,d) \quad (6)$$

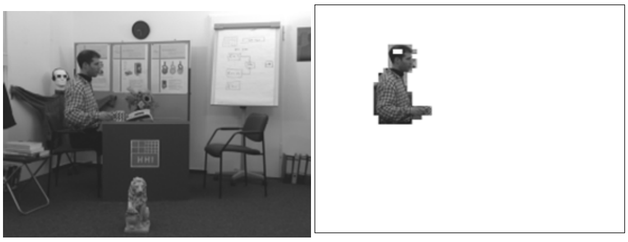
$$E_{temp}(x,y,d) = \lambda \,|\, d - D_{prev}(x+\Delta x, y+\Delta y)\,| \quad (7)$$

where λ represents the slope of the function and $D_{prev}(x,y)$ represents the previous disparity.

Since viewers mostly feel the flickering artifacts at the background, we first apply the weighting function to the background so as to reduce flickering artifacts. In order to separate the moving object, we calculate mean absolute difference (MAD) for each block and distinguish by threshold whether the block is background or not. Therefore, λ can be defined by

$$\lambda = \begin{cases} 1 & \text{if } MAD_k < Th \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where MAD_k represents the MAD of the k-th block including the position (x,y) and Th represents the threshold. The threshold is set by the experiments. Figure 5 shows the result of the moving object detection for 'Book Arrival'.



(a) Original image (b) Detected moving object

Figure 5. Moving object detection for 'Book Arrival'

In order to exploit the temporal weighting function for moving objects, we perform the motion estimation technique. Notice that the block size for motion estimation is smaller than that for the moving object detection for more accurate and reliable motion search. Figure 6 shows 80th frame of the reconstructed depth map resulting from motion estimation.

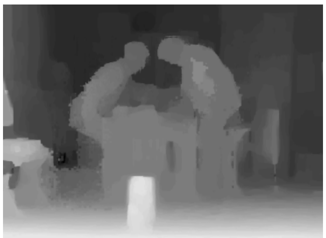
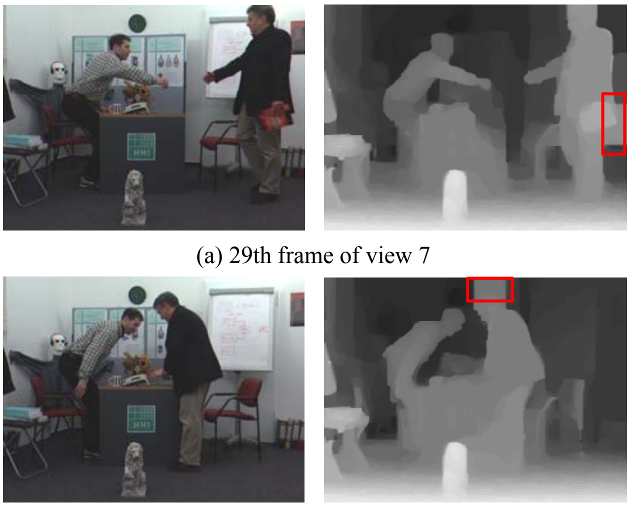


Figure 6. Reconstructed depth map for 'Book Arrival'

3.2 Prevention of Border Error Propagation

One of the noticeable errors of the graph cut-based depth estimation exist near the border of the depth map. If the foreground object suddenly appears from outside of the scene to the inside or approaches near the border, the background near the object has depth values of the object. Figure 7 shows the depth errors near the borders.



(a) 29th frame of view 7

(b) 67th frame of view 7

Figure 7. Depth errors near the border for 'Book Arrival'

If the area near the border is detected as the background, the temporal weighting function refers the wrong depth values and the errors are propagated. It leads to the serious problem for temporal consistency enhancement.

Therefore, we need to prevent the border error propagation. If λ is equal to 0, we refer to the depth values in the previous frame only when λ is equal to 0. Figure 8 shows how to refer the previous frame when applying the temporal weighting function. As shown in Fig. 8, the bold line represents the reference method for the border areas, whereas the dotted line represents the typical reference.

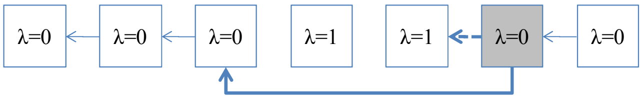
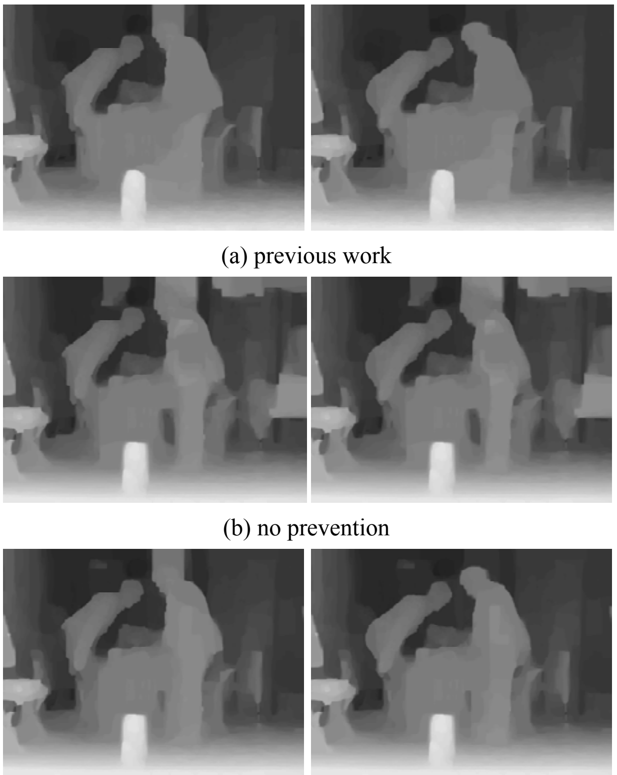


Figure 8. Prevention method for previous frame reference

Figure 9 shows the results of prevention of border error propagation. As shown in Fig. 9(b), if we do not prevent the border error propagation, the depth sequence has the depth errors near the border and those errors propagate. However, as shown in Fig. 9(c), we can enhance the temporal consistency without the border error propagation.

4. EXPERIMENTAL RESULTS

In our experiments, we used four test sequences: 'Alt Moabit', 'Book Arrival' provided by Heinrich-Hertz-Institut (HHI) [10], 'Lovebird1' provided by Electronics and Telecommunications Research Institute (ETRI) and MPEG-Korea Forum [11], and 'Newspaper' provided by Gwangju Institute of Science and Technology (GIST) [12]. These sequences are distributed for the purpose of 3D video coding and the resolution of them is 1024x768.



(a) previous work

(b) no prevention

(c) prevention results

Figure 9. Results of prevention of border error propagation

4.1 Depth Estimation and View Synthesis

In order to obtain the depth sequence, we used Depth Estimation Reference Software (DERS) provided by 3D video coding subgroup in MPEG. In our experiments, the block size for the background separation is set to be 32 and the block size for the motion estimation is 8. The search range is set to be from -16 to +16. The threshold for moving object detection is represented in Table 1.

Table 1. Threshold for moving object detection

Sequence	Threshold
Alt Moabit	2.50
Book Arrival	3.00
Lovebird1	1.50
Newspaper	1.50

Figure 10 through Figure 12 show the depth sequences for 'Alt Moabit', 'Book Arrival' and 'Newspaper'. As shown in Fig. 10(a), through Fig. 12(a), we noticed that the depth sequences have inconsistent depth near the background, whereas the depth sequences in Fig. 10(b) through Fig. 12(b) are temporally consistent.

In order to synthesize the virtual view, we used View Synthesis Reference Software (VSRS) provided by 3D video coding subgroup in MPEG [10]. Figure 13 through Figure 15 show the virtual views for 'Alt Moabit', 'Book Arrival' and 'Newspaper'. The right figures shows the difference images between the left and the center figures. We also noticed that the virtual views have errors at the background as shown in Fig. 13(a) through Fig. 15(a), whereas the virtual views in Fig. 13(b) through Fig. 15(b) have less errors than those without temporal consistency enhancement.

Table 2 shows the average PSNR using the original view and the virtual view. As shown in the results, PSNR of the proposed algorithm was almost the same as that of the previous work. It means that the flickering artifacts of synthesized views were reduced without any degradation of objective quality.

Table 2. Average PSNR of virtual views

Sequence	View	DERS	Proposed method	Difference (ΔdB)
Alt Moabit	8	35.1473	35.3298	0.1825
	9	35.4158	35.4689	0.0532
Book Arrival	8	34.3986	34.4529	0.0542
	9	35.5694	35.5472	-0.0222
Lovebird1	6	30.9870	30.9866	-0.0004
	7	30.4126	30.4214	0.0087
Newspaper	4	24.3732	24.3695	-0.0037
	5	25.4442	25.4389	-0.0054
Average		31.4685	31.5019	0.0334

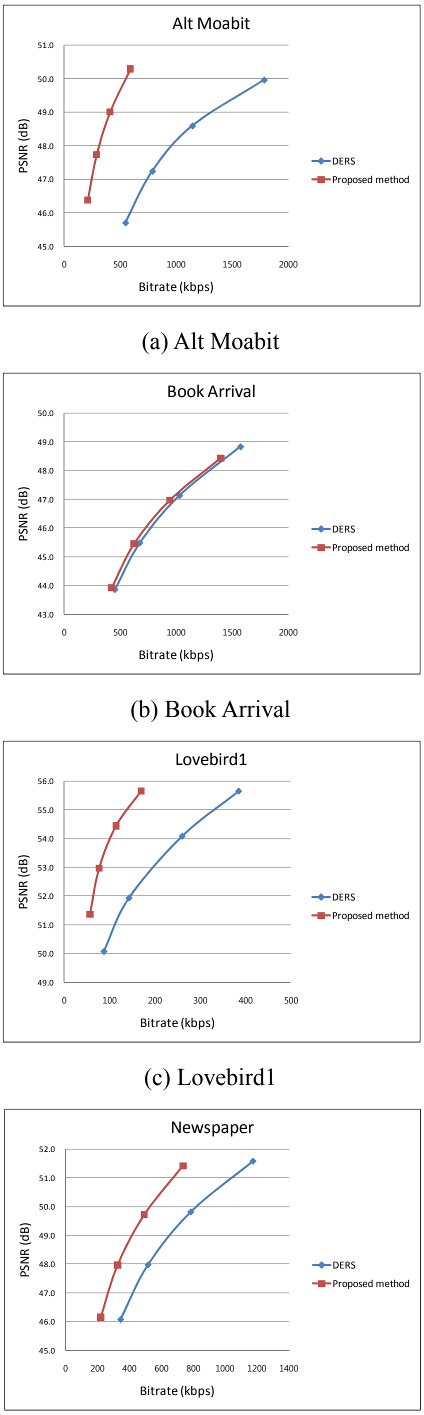
4.2 Depth Map Coding

For the evaluation of the effects on the video coding of our proposed method, we performed depth map coding for each sequence by using the H.264/AVC reference software version JM 14.0. We tested 100 frames for each sequence and the coding structure was IPPP...P. Table 3 represents the depth map coding results and Figure 16 shows the rate-distortion curves for each sequence.

Table 3. Depth map coding results

Sequence	QP	DERS		Proposed method	
		PSNR (dB)	Bitrate (kbps)	PSNR (dB)	Bitrate (kbps)
Alt Moabit (view 7)	22	49.96	1785.46	50.29	592.22
	25	48.59	1144.74	49.00	410.21
	28	47.24	788.39	47.73	290.33
	31	45.70	547.95	46.38	211.50
Book Arrival (view 7)	22	48.83	1573.54	48.43	1396.91
	25	47.13	1032.20	46.96	941.86
	28	45.48	677.70	45.46	625.13
	31	43.86	456.22	43.92	423.30
Lovebird1 (view 5)	22	55.64	384.62	55.66	170.26
	25	54.07	260.00	54.44	114.63
	28	51.92	142.81	52.96	77.36
	31	50.06	87.58	51.37	57.33
Newspaper (view 3)	22	51.47	1229.67	51.25	730.68
	25	49.67	829.03	49.53	488.31
	28	47.83	556.19	47.78	323.41
	31	45.95	383.23	45.97	221.28

From the depth map coding results, we noticed that the the visual quality of the proposed algorithm was improved about 2.56dB or the bitrate was reduced about 42.55% in Bjontegaard measure [13]. Since the proposed algorithm enhanced the temporal consistency of depth sequences without any degradation of virtual views, the accurate inter prediction of H.264/AVC was possible and the coding efficiency was improved.



(a) Alt Moabit

(b) Book Arrival

(c) Lovebird1

(d) Newspaper

Figure 16. Rate-distortion curves

5. CONCLUSIONS

In this paper we have proposed the temporal consistency enhancement algorithm for multiview depth estimation. We used the block-based moving object detection to separate the moving object from the background and applied the temporal weighting function only to the background. As a result, we reduced the flickering artifacts of virtual views without any degradation of visual quality from the experimental results.

ACKNOWLEDGEMENT

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 NIPA(National IT Industry Promotion Agency" NIPA-2009-(C1090-0902-0017)).

REFERENCES

[1] J. Freeman, S.E. Avons, "Focus Group Exploration of Presence through Advanced Broadcast Services," in Proc. of the SPIE, Human Vision and Electronic Imaging, pp. 3959-3976, 2000.

[2] A. Smolic, K. Müller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, T. Wiegand, "3D Video and Free Viewpoint Video – Technologies, Applications and MPEG Standards," in Proc. of IEEE International Conference on Multimedia and Expo, pp. 2161-2164, 2006.

[3] A. Redert, M. O. Beeck, C. Fehn, W. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, I. Sexton, P. Surman, "ATTEST: Advanced Three-dimensional Television System Techniques," in Proc. of International Symposium on 3D Data Processing, pp. 313-319, 2002.

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

[5] D. Sharstein, R. Szeliski, "A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms," in Proc. of IEEE Workshop on Stereo and Multi-Baseline Vision, pp. 131-140, 2001.

[6] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, "High-quality Video View Interpolation Using a Layered Representation," in Proc. of SIGGRAPH'04, pp. 600-608, 2004.

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

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

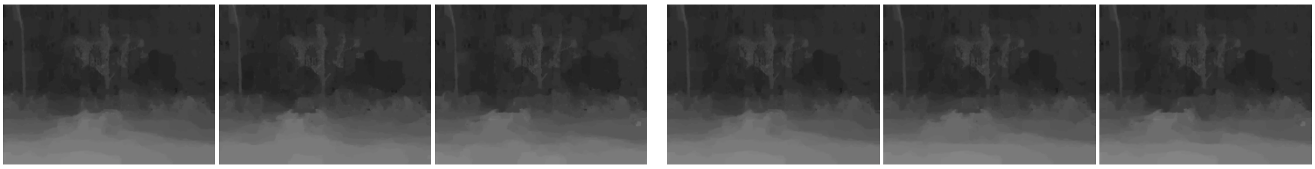
[9] ISO/IEC JTC1/SC29/WG11, "Experiment on Temporal Enhancement for Depth Estimation," M15852, Oct. 2008.

[10] ISO/IEC JTC1/SC29/WG11, "HHI Test Material for 3D Video," M15413, April 2008.

[11] ISO/IEC JTC1/SC29/WG11, "Contribution for 3D Video Test Material of Outdoor Scene," M15371, April 2008.

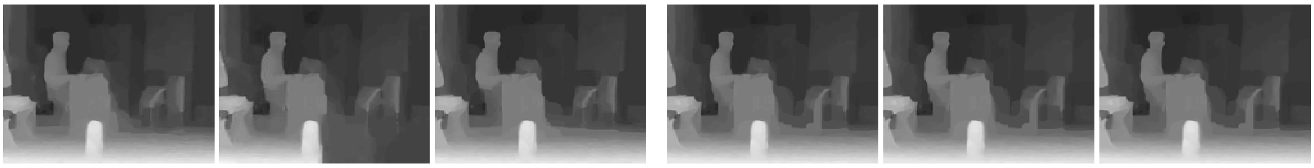
[12] ISO/IEC JTC1/SC29/WG11, "Multiview Video Test Sequence and Camera Parameters," M15419, April 2008.

[13] ITU-T SG16 Q.6, "An Excel Add-in for Computing Bjontegaard Metric and Its Evolution," VCEG-AE07, Marrakech, MA, Jan. 2007.



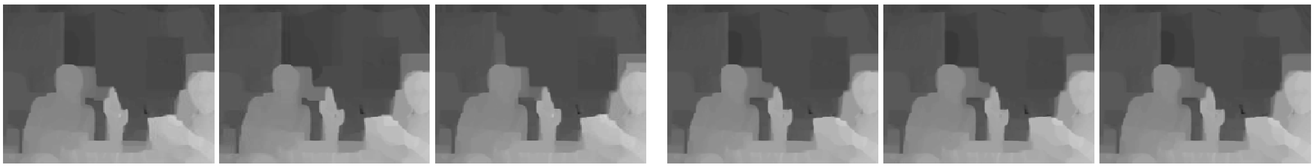
(a) Depth map without temporal consistency enhancement (b) Depth map with temporal consistency enhancement

Figure 10. Depth estimation results for 'Alt Moabit'



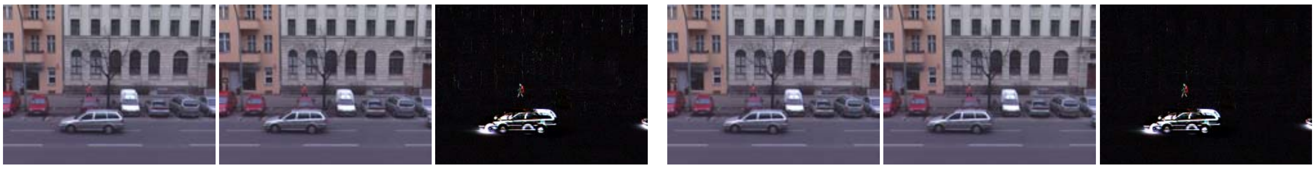
(a) Depth map without temporal consistency enhancement (b) Depth map with temporal consistency enhancement

Figure 11. Depth estimation results for 'Book Arrival'



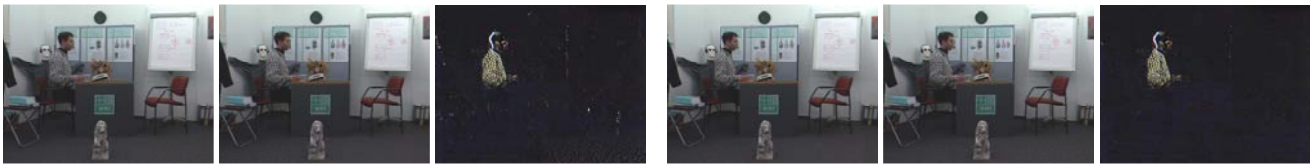
(a) Depth map without temporal consistency enhancement (b) Depth map with temporal consistency enhancement

Figure 12. Depth estimation results for 'Newspaper'



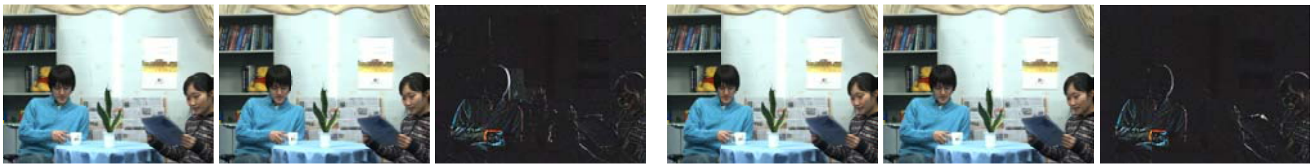
(a) Virtual view without temporal consistency enhancement (b) Virtual view with temporal consistency enhancement

Figure 13. View synthesis results for 'Alt Moabit'



(a) Virtual view without temporal consistency enhancement (b) Virtual view with temporal consistency enhancement

Figure 14. Results of virtual view for 'Book Arrival'



(a) Virtual view without temporal consistency enhancement (b) Virtual view with temporal consistency enhancement

Figure 15. View synthesis results for 'Newspaper'