Temporal Consistency Enhancement of Background for Depth Estimation

Sang-Beom Lee, Cheon Lee, and Yo-Sung Ho

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

ABSTRACT

In this paper, we propose a new scheme to enhance temporal consistency of depth sequences for multi-view depth estimation. When we compute the matching cost for estimating appropriate depth, we add a temporal weighting function to the conventional matching function. Furthermore, since the temporal weighting function only works well for the background, we apply it only to the background. Experimental results showed that the proposed algorithm enhanced the temporal consistency for the background of the depth sequence and reduced flickering artifacts in the virtual view while maintaining visual quality.

1. INTRODUCTION

A three-dimensional television (3DTV) using multi-view video becomes attractive as one of the next-generation broadcasting services [1]. In order to acquire multi-view images, we utilize multiple cameras with parallel or convergent configuration to capture a 3-D scene with wide-viewing angle and we feel the presence from the multi-view images through 3-D displays.

Given the increasing diversity of 3-D services and displays, proper rendering of 3-D scene is necessary. For example, if the number of multi-view images is smaller than the input number of images for 3-D display, we need to reconstruct intermediate views. An intermediate view is a virtual image captured by a virtual camera which is positioned between two real cameras. By interpolating intermediate views, we not only provide us with high quality 3-D contents, but also reduce the visual discomfort of the viewer.

In order to reconstruct intermediate views of virtual viewpoints, we need depth information. Many works have been carried out for the acquisition of 3-D depth information [2]. Recently, 3-D video coding subgroup in Moving Picture Experts Group (MPEG) recognized the importance of a multi-view video and depth information and requested depth estimation tools [3]. In response to the request, Nagoya University implemented and distributed the graph cut-based depth estimation software [4].

However, there are several problems such as boundary mismatch, textureless regions, occlusion problem or wide baseline. Especially, since this 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.

Therefore, we propose a new method 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. This weighting function is available at the background since the most flickering artifacts are detected at the background. In order to separate the moving object from the background, we exploit the block-based moving object detection.

2. MULTI-VIEW DEPTH ESTIMATION

2.1. Relationship between Disparity and Depth

Figure 1 illustrates the relationship between disparity and depth. Suppose that a certain point is located at (x_r, y) in the right view and (x_l, y) in the left view. 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 obtain the depth if we estimate the disparity from multi-view images.

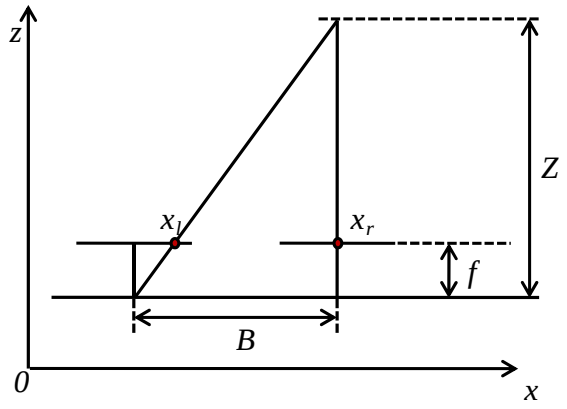


Figure 1. Relationship between disparity and depth

2.2. Depth Estimation using Multi-view Images

In response to call for depth estimation software in MPEG, Nagoya University implemented and distributed graph cut-based depth estimation software [4]. The software is categorized by three parts: disparity computation, graph cut-based error minimization, and disparity-to-depth conversion.

The first step is to compute the matching cost for each pixel of the center view. Since we have three and more views of multi-view video, we can compare center view to the left and right view simultaneously. Thus, we can solve the occlusion problem. The matching function is defined by

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

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

$$E_R(x,y,d) = | I_C(x,y) - I_R(x-d,y) | \quad (4)$$

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 furthest 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:

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

Here, Z_{far} and Z_{near} represent the farthest and nearest depth values.

3. TEMPORAL CONSISTENCY ENHANCEMENT

3.1. Temporal Weighting Function

As mentioned before, since the depth estimation software estimates 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 [5]. The temporal weighting function is defined by

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

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

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

Figure 2 illustrates an example of matching functions. The dotted line, the chain line, and the solid line represent curves of the previous matching function, the temporal weighting function, and the proposed matching function, respectively.

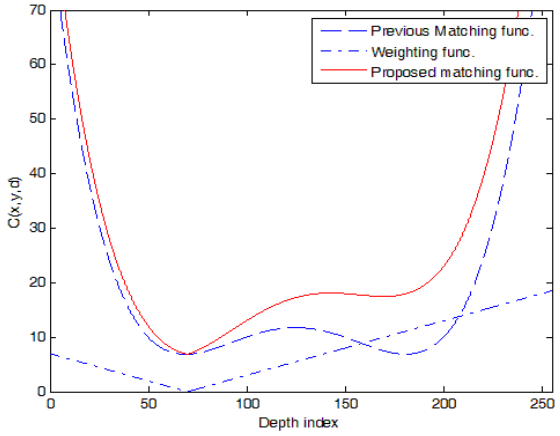


Figure 2. Curves of matching functions

3.2. Proposed Temporal Consistency Enhancement

We noticed that the depth sequence had ghost artifacts near moving objects when applying the temporal weighting function to the whole scene [6]. Figure 3 depicts the depth map and the virtual view with ghosting artifacts for 'Lovebird1'.

(a) conventional depth map and virtual view

(a) depth map and virtual view with ghosting artifact

Figure 3. Ghosting artifacts for 'Lovebird1'

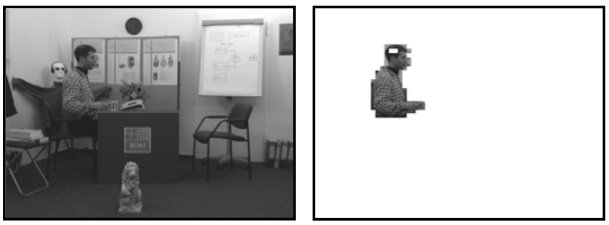
This problem is caused by the propagation of depth values at the previous frame. It degraded the visual quality of the synthesized views. Since the

temporal weighting function refers to the previous depth value at exactly the same position, it only works well at the static background.

Therefore, we detect the moving object and apply the temporal weighting function only to the background. Since viewers mostly feel the flickering artifacts at the background, it is satisfactory that we apply the weighting function only 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. Figure 4 shows the result of the moving object detection for 'Book Arrival'.



(a) Original image (b) Detected moving object

Figure 4. Moving object detection for 'Book Arrival'

4. EXPERIMENTAL RESULTS

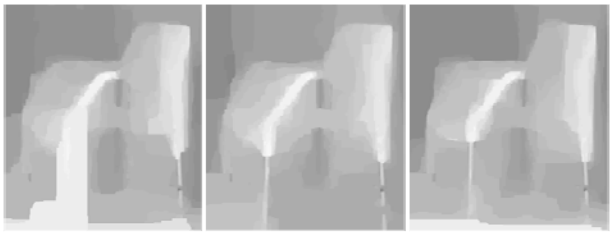
In order to evaluate our algorithm, we used 10 test sequences: 'Book Arrival', 'Door Flower', 'Leaving Laptop', and 'Alt Moabit' provided by Heinrich-Hertz-Institut (HHI), 'Champagne Tower', 'Pantomime', and 'Dog' provided by Nagoya University, 'Newspaper' provided by Gwangju Institute of Science and Technology (GIST), 'Lovebird1' and 'Lovebird2' provided by MPEG-Korea Forum and Electronics and Telecommunications Research Institute (ETRI).

Table 1. Threshold for moving object detection

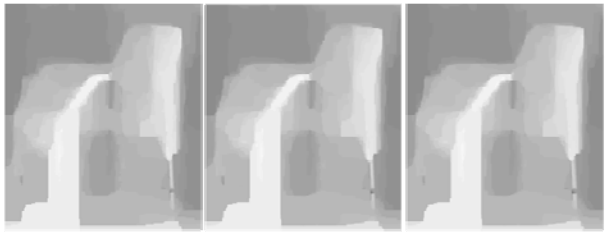
Sequence	Th	Sequence	Th
Book Arrival	2.50	Pantomime	1.00
Door Flower	2.50	Dog	2.00
Leaving Laptop	2.50	Newspaper	1.50
Alt Moabit	2.50	Lovebird1	1.50
Champagne Tower	1.00	Lovebird2	1.50

These sequences are distributed for the purpose of 3-D video coding of MPEG. We used Depth Estimation Reference Software (DERS) provided by Nagoya University for depth estimation. In addition, we used the parameters for moving object detection as represented in Table 1.

Figure 5 and Figure 6 show the enlarged figures of depth sequences for 'Book Arrival' and 'Alt Moabit'. As shown in Figure 5(a) and Figure 6(a), we noticed that the depth sequences have inconsistent depth near the chair's legs and near the tree's branches, whereas the depth sequences in Figure 5(b) Figure 6(b) have consistent depth.

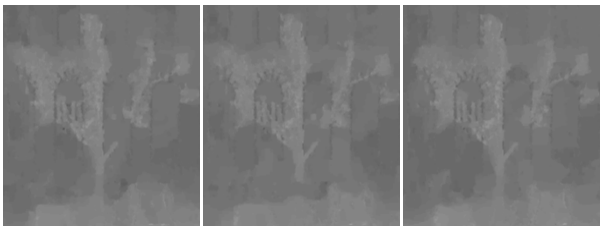


(a) depth map without temporal enhancement

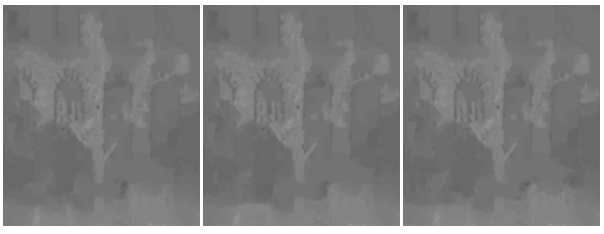


(b) depth map with temporal enhancement

Figure 5. Results of depth map for 'Book Arrival'



(a) depth map without temporal enhancement



(b) depth map with temporal enhancement

Figure 6. Results of depth map for 'Alt Moabit'

We used View Synthesis Reference Software (VSRS) provided by Nagoya University to synthesize the virtual view [4]. Figure 7 and Figure 8 show the enlarged figures of virtual views for 'Book Arrival' and 'Alt Moabit'. We also noticed that the virtual views have inconsistent shapes as shown in

Figure 7(a) and Figure 8(a), whereas the virtual views in Figure 7(b) and Figure 8(b) have consistent shapes even if they include boundary errors at the chair's legs and the tree's branches.

(a) virtual view without temporal enhancement

(b) virtual view with temporal enhancement

Figure 7. Results of virtual view for 'Book Arrival'

(a) virtual view without temporal enhancement

(b) virtual view with temporal enhancement

Figure 8. Results of virtual view for 'Alt Moabit'

Table 2 shows the results of PSNR comparison between the original view and the virtual view. As shown in the results, PSNR values of the proposed scheme was almost the same as that of the previous work. In other words, the flickering artifacts of synthesized views were reduced without any degradation of the objective quality.

5. CONCLUSIONS

We have proposed the temporal consistency enhancement scheme for 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. We reduced the flickering artifacts of virtual views without any degradation of visual quality from the experimental results.

Table 2. Average PSNR

Sequence	View	Temporal enhancement		ΔdB
		Off (dB)	On (dB)	
Book Arrival	8	34.40	34.48	+0.08
Door Flower	8	36.16	36.20	+0.04
Leaving Laptop	8	36.07	35.98	-0.09
Alt Moabit	8	35.15	35.34	+0.19
Champagne Tower	39	28.77	28.72	-0.05
Pantomime	39	35.87	35.83	-0.04
Dog	39	31.13	31.13	-
Newspaper	4	24.37	24.37	-
Lovebird1	6	30.99	31.00	+0.01
Lovebird2	6	34.18	34.20	+0.02

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. A. Smolic, K. Müller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, "3D Video and Free Viewpoint Video – Technologies, Applications and MPEG Standards," IEEE International Conference on Multimedia and Expo, pp. 2161-2164, July 2006.
2. D. Sharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms," Proc. of IEEE Workshop on Stereo and Multi-Baseline Vision, pp. 131-140, Dec. 2001.
3. ISO/IEC JTC1/SC29/WG11 "Call for Contributions on 3D Video Test Material," N9595, Jan. 2008.
4. ISO/IEC JTC1/SC29/WG11 "Reference Software of Depth Estimation and View Synthesis for FTV/3DV," M15836, Oct. 2008.
5. S. Lee and Y. Ho, "Multi-view Depth Map Estimation Enhancing Temporal Consistency," Proc. of International Technical Conference on Circuits/Systems, Computers and Communications, pp. 29-32, July 2008.
6. ISO/IEC JTC1/SC29/WG11 "Experiment on Temporal Enhancement for Depth Estimation," M15852, Oct. 2008.