3D JBU BASED DEPTH VIDEO FILTERING FOR TEMPORAL FLUCTUATION REDUCTION

Jinwook Choi[*], *Dongbo Min*[**], *Donghyun Kim*[*] *and Kwanghoon Sohn*[*]

[*]Dept. of Electrical and Electronic Eng., Yonsei University, Seoul, Korea
[**]Advanced Digital Sciences Center, Singapore
khsohn@yonsei.ac.kr

ABSTRACT

In this paper, we propose a three-dimensional Joint Bilateral Up-sampling (3D JBU) for the depth video which can be applied to a 3DTV system based on 2D-plus-depth video. Recently, a number of researches have been done to improve a resolution and frame-rate of the depth video. We proposed a novel method that enhances depth video obtained by Time-of-Flight (TOF) sensor by combining it with Charge-coupled Device (CCD) camera [1]. However, this method does not consider the temporal coherence of the depth video. It may cause an eye fatigue on the 3D display and increase bit rates on video coding, since it is possible to generate a temporal fluctuation problem. Therefore, in order to solve these problems, we propose a 3D JBU model which is extended conventional JBU into the temporal domain of depth video. Experimental results show that depth video obtained by the proposed method provides satisfactory quality.

Index Terms—Depth video, 3D JBU, temporal coherence, 3DTV system, 2D-plus-Depth

1. INTRODUCTION

A number of researches have been developed recently to improve a resolution and frame-rate of the depth video. High-quality depth video is needed to be applied in a 3DTV system, especially, 2D-plus-Depth based system. There are several methods that obtain the depth map or depth video. The laser scanning method provides the most accurate depth. However, it is not only time-consuming and expensive, but also is applied to only static objects. Stereoscopic methods estimate a depth using multiple images obtained by several cameras. Depth estimated by this method is not accurate at texture-less, occlusion and repeated pattern regions. Moreover, as the performance of algorithm is higher, the computation complexity is also higher. On the other hand, range sensor methods using TOF sensors are possible to provide an accurate depth relatively, and use in real-time applications. However, it provides a low resolution and low frame-rate depth video, and the noisy results of an object that has high reflectance due to the physical limits of TOF

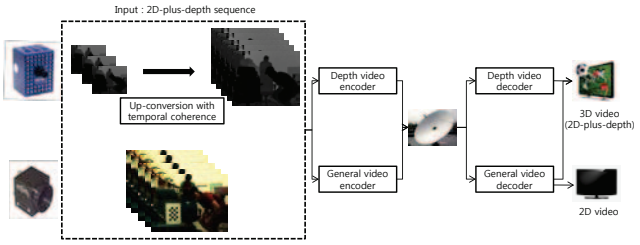


Fig. 1. An example of a 3DTV system based on 2D-plus-depth

sensors [2]. In contrast to obtain the depth using TOF sensors only, CCD cameras, used in combination with TOF sensors, provide sufficiently high resolution and frame-rate video. Therefore, CCD cameras can be used to overcome the disadvantages of TOF sensors [1].

Fig. 1 represents an example of a 3DTV system using a TOF sensor and CCD camera. 2D video and the corresponding depth video are transmitted through the communication network, and a user can then select a viewing mode in the 3DTV system based on 2D-plus-depth in the receiver part. Therefore, high-quality depth video corresponding to 2D video is required in the 3DTV system.

In the case of up-sampling in the spatial domain, Joint Bilateral Up-sampling (JBU) [3] may work well, because it preserves the edge of up-sampled depth maps accurately. The JBU is based on the bilateral filter which is an edge-preserving filter [4]. Where only one image or two images of the same size are used in the bilateral filter, JBU uses two images with different sizes. JBU has an advantage, changing a low resolution image to a high resolution image while preserving the edge and smoothing the homogeneous region. The depth map has the characteristic that most regions are homogeneous because the depth is similar throughout the same object. Fig. 2 shows examples of the Fast Fourier Transform (FFT) for both the depth map and CCD image. In case of depth map, most energy in the frequency domain is concentrated at low frequencies, which is very different from the natural images acquired by CCD cameras. Thus, if edges in the up-sampled depth map are preserved well, we can conclude that the up-sampled result is satisfactory because the edge of the depth map is the most important factor in the

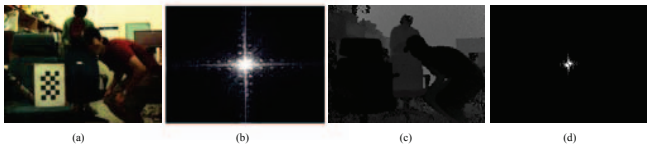


Fig. 2. CCD and depth image analysis in the frequency domain: (a) CCD image, (b) FFT result of (a), (c) original depth map, (d) FFT result of (c).

evaluation of its quality. It makes JBU of the depth map possible. The equation for JBU is as follows:

$$\tilde{R}_p = \frac{1}{k_p} \sum_{q_\downarrow \in \Omega} R_{p_\downarrow} f(\| p_\downarrow - q_\downarrow \|) g(\| \tilde{I}_p - \tilde{I}_q \|) \quad (1)$$

Given a high resolution image, $\tilde{I}$, and a low resolution, R, we can obtain an up-sampled solution $\tilde{R}$ by using two kinds of filters. The first one is f representing the spatial filter kernel such as a Gaussian centered over $p_\downarrow$ in a low-resolution image. The other one is g representing the range filter kernel, centered at the pixel value at p in a high-resolution image. p, q in Eq. (1) represent the locations of pixels in $\tilde{I}$, and $p_\downarrow$, $q_\downarrow$ represents the corresponding locations of pixels in a low resolution image, R. Ω is the spatial support of the kernel f, and k_p represents a normalization factor, the sum of the $f \cdot g$ filter weights.

However, in the case of extending the image up-sampling to the video, conventional JBU may cause temporal fluctuation problems, because it is performed without considering any temporal information. To address this problem, we extended conventional JBU into a 3-dimension volume including the temporal domain. The contribution of this paper is that a 3D JBU model is proposed to reduce the temporal fluctuation of depth video. It can help reduce the bit rates in depth video coding and eye fatigue on the 3D display.

Temporal fluctuation which is also known as stationary area fluctuation [5] or flickering artifact [6] is important issue to be addressed. It is usually considered in video coding. Fan *et al* proposed a modified encoder with improved intra prediction [6] and Yang *et al* proposed a robust filtering [7] in order to reduce temporal fluctuation problems. In addition, rate control method for minimizing temporal fluctuation was proposed [8]. However, these methods are related to video coding. The proposed method differs from them, since it is considered as video processing applied to the up-conversion of the depth video. In this paper, we focus on not aspect of video coding but up-conversion of the depth video by 3D JBU in order to reduce temporal fluctuation problems fundamentally.

The remainder of this paper is organized as follows. In Section 2, we describe the depth video filtering technique using the proposed method. Finally, we present experimental results and conclusions in Section 3 and 4, respectively.

2. DEPTH VIDEO FILTERING WITH 3D JBU

In [1], we improved the resolution of depth video in the spatial and temporal domains in order to make high quality 2D-plus-depth contents. However, temporal fluctuation problems may occur because [1] does not consider any other information of neighboring frames in temporal domain. There is also a very important problem in viewing 2D-plus-depth contents on the 3D display. In contents which have serious temporal fluctuation problems, humans may feel fatigue easily [9]. Therefore, we solve the temporal fluctuation problem using 3D JBU, which is extended into a 3-dimensional volume, considering the temporal domain. JBU is generally a 2-dimensional process because both range and spatial filter kernels are 2-dimensional structures. JBU is extended into a 3-dimensional volume by accumulating neighboring frames.

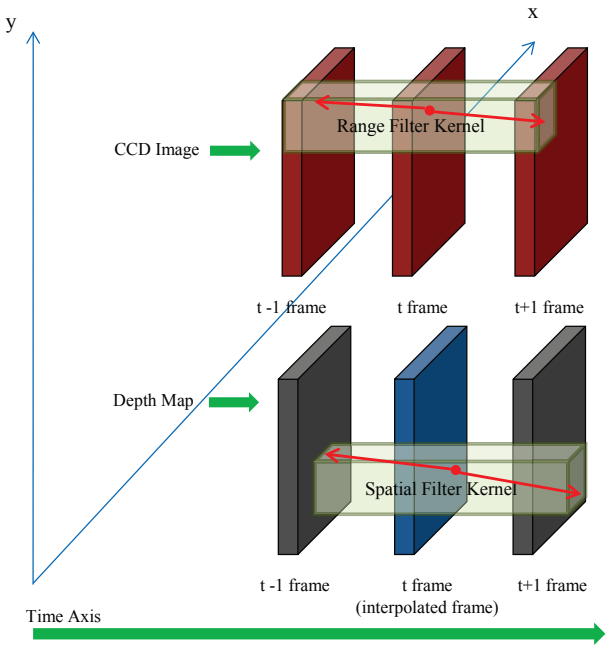


Fig. 3. Structure of 3D JBU (N=3).

As shown in Fig. 3, we use not only the CCD image and depth map for the interpolated frame, but also neighborhood frames to maintain temporal constancy in the up-converted depth video. 3D memories, corresponding to the spatial filter kernel of depth images and the range filter kernel of CCD images, can be made by integrating neighborhood frames in the temporal domain as shown in Fig. 3. If the number of frames used in 3D JBU is 3, in order to up-sample the resolution of the interpolated frame (t frame), we use images in its temporal neighborhood, such as frames *(t - 1)* and *(t + 1)*. The up-sampled results are improved significantly from those of the conventional JBU, because both temporal and spatial information are considered. By using a 3-dimensional

Gaussian filter as the spatial filter kernel and a 3D window as the range filter kernel, temporal fluctuation can be reduced, while the edge is preserved in the video. The 3D JBU equation is as follows:

$$\tilde{R}_{p,t} = \frac{1}{k_{p,t}} \sum_{N} \sum_{q_{\downarrow} \in \Omega} R_{p_{\downarrow},t} f(\| p_{\downarrow,t} - q_{\downarrow,t_N} \|) g(\| \tilde{I}_{p,t} - \tilde{I}_{q,t_N} \|) \quad (2)$$

In Eq. (2), t and t_N represent the reference and neighborhood frames used in 3D JBU. $p_{\downarrow,t}$, $q_{\downarrow,t_N}$ in Eq. (2) represent the locations of pixels in the frame t and t_N of a low resolution image R (depth video). The basic structure of 3D JBU is similar to that of the conventional JBU, except that filter kernels contain temporal neighborhood information. f indicates the 3-dimensional filter such as a Gaussian centered over $p_{\downarrow,t}$ in the frame t of a low-resolution image. g indicates the range filter kernel, centered at the pixel value at p in the frame t of a high-resolution image. In other words, it means the range filter kernel based on the difference in intensity values between corresponding pixels and neighborhood pixels in the 3D window of neighborhood frame, N. N represents the number of neighborhood frames used in 3D JBU. In this paper, we used 3-neighborhood frames (N=3) because of computational complexity. When depth video is up-sampled by 3D JBU, the previous neighborhood images are needed, so that N-neighborhood frames per up-sampling process are needed in a memory.

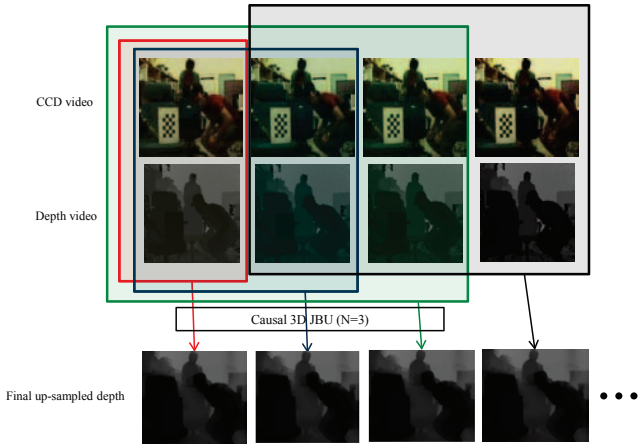
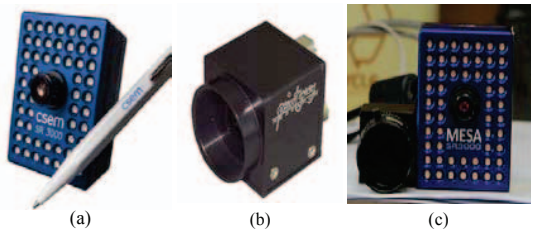


Fig. 4. Example of causal 3D JBU (N=3).

However, a system, such as that shown in Fig. 3, can be non-causal, since frame *(t + 1)* is used in the up-conversion process. It may cause the delay at the up-conversion process. Therefore, we can solve this problem by using only previous frames, such as *(t – 2)* and *(t – 1)*. Fig. 4 shows the example of causal 3D JBU when the number of neighborhood frame N is 3. The first frame in the depth video can be up-sampled by using only corresponding frames, since there is no previous frame. The subsequent frames can then be up-sampled by using a number of frames added by 1 until the N^{th} frame is input. By extending 2D spatial information into 3D temporal information, we can obtain temporally consistent quality depth video. From the view-point of video coding, the reduction of temporal fluctuation can help save the bit rate. In the case of a 2D still image, the compression rate of an image in which most regions are homogeneous may be better than that of an image in which noise is randomly distributed. Similarly, temporal fluctuation problems may degrade the performance of video coding. Therefore, temporal fluctuation problems in the depth video should be removed in order to reduce both the bit rate in the coding and eye fatigue on the 3D display.

3. EXPERIMENTAL RESULTS

In the experiment, the Flea camera made by Point Grey Research, Inc. [10] and the SR3000 depth sensor, made by MESA Imaging [11], are used to acquire CCD images and depth images, as shown in Fig. 5 (a) and (b). The resolution of the Flea camera is 1024×768, and the frame-rate is 30 fps (frames per second). The resolution of the depth sensor is 176×144, and the frame-rate is about 15 fps. The two sensors are synchronized with each other in spite of the difference of frame-rate. As shown in Fig. 5 (c), in this experiment, two sensors have been placed near each other in parallel. The video used in the experiment is "Two Men" (obtained by our fusion system, which consists of the SR3000 and Flea camera).



(a) (b) (c)

Fig. 5. Experimental device and setup: (a) depth sensor SR3000, (b) CCD camera Flea, (c) experimental setup.

Temporal fluctuation problem in the frame interpolated depth video using 2D JBU may be caused, since only one pair of images which consist of CCD image and corresponding depth map are used without considering temporal coherence. Fig. 6 shows the results (277^{th}, 278^{th}, 279^{th} frames) up-converted by 2D JBU and difference images between the frames (277^{th}-278^{th}, 278^{th}-279^{th}). The background regions which contain severe time fluctuation in the red box of Fig. 6 may cause eye fatigue in viewing 2D-plus-depth based 3D contents. It does not consider temporal coherence, so that varying depth values at the background pixels were assigned although the scene is static. In Fig. 7, 3D JBU (N=3) was used in the up-sampling process in order to remove temporal fluctuation. We can find that temporally

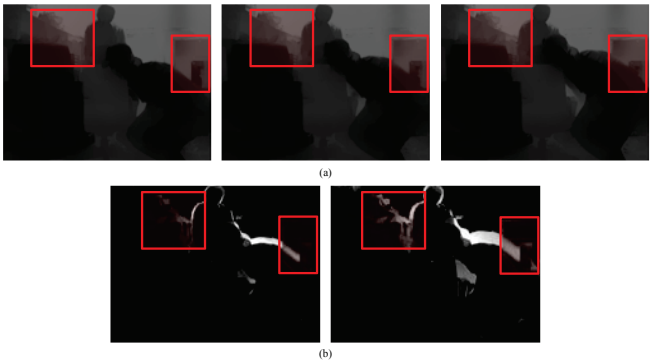


Fig. 6. (a) Up-sampled depth video of "Two men" obtained by 2D JBU and (b) difference images.

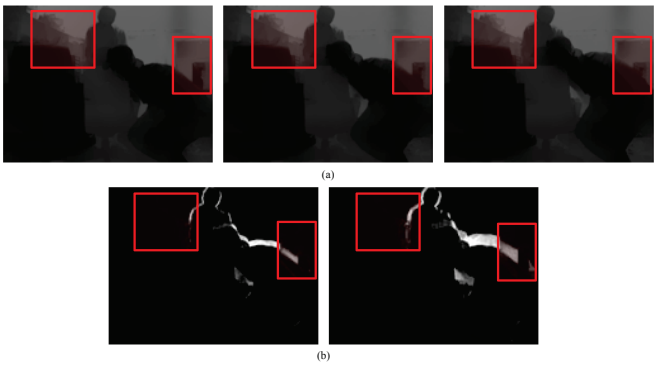


Fig. 7. (a) Up-sampled depth video of "Two men" obtained by 3D JBU and (b) difference images.

consistent depth values were obtained in most part except for moving objects using 3D JBU compared to Fig. 6. That is, difference of values between frames in the red box is almost removed.

In addition, this process improves the compression rate of the depth video. By reducing the temporal fluctuation, we can obtain gain in video coding. As shown in Fig. 8, compression rate of depth video using 3D JBU is higher than that using 2D JBU while PSNR is preserved to satisfactory quality that people cannot distinguish the original with the compressed videos. Bit rate decreases in overall QP values as shown in Fig. 8. Experimental results were compressed by H.264 AVC JM 12.4 according to various QP values. Reduction of bit rate means that amount of depth data transmitted to the receiver is reduced in 2D-plus-depth based 3DTV system. Therefore, reduction of temporal fluctuation is certainly needed to 3DTV system.

Table 1 shows the processing time of 3D JBU proposed in this paper and 2D JBU implemented in [1].

Table 1. Processing time comparison of 3D JBU with conventional JBU.

Step	2D JBU	3D JBU
Time(sec)	1.902	9.214

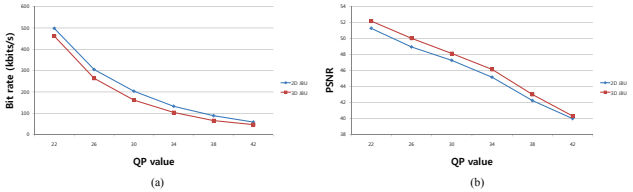


Fig. 8. (a) Bit rate comparison, (b) PSNR comparison of 3D JBU with 2D JBU in "Two men" video.

4. CONCLUSION

In this paper, we propose the 3D JBU model for high-quality depth video. By using the proposed model, we can remove the temporal fluctuation problem of up-converted depth video in the temporal domain. It can reduce the bit-rates of the depth in video coding and eye fatigue on the display. However, 3D JBU process is time-consuming compared to conventional JBU. In the future research, fast algorithm of 3D JBU will be studied. In addition, 2D-plus-depth contents made by 3D JBU will be used for subjective quality assessment on a 3D display as well as analysis by difference image between frames in order to analyze the relationship between temporal fluctuation and eye fatigue accurately.

5. REFERENCES

[1] J. Choi, D. Min, B. Ham and K. Sohn, "Spatial and temporal up-conversion technique for depth video," *in Proc. IEEE Conf. Image Processing*, pp. 3525-3528, 2009.

[2] S. A. Gudmundsson, H. Aanaes and R. Larsen, "Environmental Effects on Measurement Uncertainties of Time-of-Flight Cameras," *In IEEE Symp. on Signals Circuits and Systems* (ISSCS), vol.1, pp. 1-4, Jul. 2007.

[3] J. Kopf, M. F. Cohen, D. Lischinski and M. Uyttendaele, "Joint bilateral upsampling," ACM Trans. Graphics, vol. 26, no. 3, article. 96 (Proc. SIGGRAPH 2007), Jul, 2007.

[4] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *in Proc. Int. Conf. Computer Vision*, pp. 839-846, 1998.

[5] M. Yuen, "Coding Artifacts and Visual Distortions," in H. R. Wu and K. R. Rao, Eds., *Digital Video Image Quality and Perceptual Coding*, pp. 87-122, CRC Press, 2006.

[6] X. Fan, W. Gao, Y. Lu and D. Zhao, "Flicking reduction in all intra frame coding," *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, JVT 5th Meeting*, Oct. 2002.

[7] J. Yang and H. Wu, "Robust Filtering Technique for Reduction of Temporal Fluctuation in H.264 Video Sequences," *IEEE Trans. Circuits Syst. Video Technol.*, 2010 (forthcoming article)

[8] A. Matsumura, S. Naito, R. Kawada and A. Koike, "Effective rate control method for minimizing temporal fluctuations in picture quality applicable for MPEG-4 AVC/H.264 encoding," *in Proc. IEEE Conf. Image Processing*, pp. 569-572, 2005.

[9] Y. Nojiri, H. Yamanoue, A. Hanazato, M. Emoto and F. Okano, "Visual comfort/discomfort and visual fatigue caused by stereoscopic HDTV viewing," *in Proc. SPIE stereoscopic Displays and Virtual Reality Systems*, vol. 5291, pp. 303-313, 2004.

[10] [Online] Point Grey Research, Inc.: http://www.ptgrey.com.

[11] [Online] MESA Imaging: http://www.mesa-imaging.ch.