SPATIAL AND TEMPORAL UP-CONVERSION TECHNIQUE FOR DEPTH VIDEO

Jinwook Choi, Dongbo Min, Bumsub Ham and Kwanghoon Sohn

Dept. of Electrical and Electronic Eng., Yonsei University, Seoul, Korea
khsohn@yonsei.ac.kr

ABSTRACT

This paper proposes a novel framework for up-conversion of depth video resolution both in spatial and in time domain. Time-of-Flight (TOF) sensors are widely used in computer vision fields. Although TOF sensors provide depth video in real time, there are some problems in a sense that it provides a low resolution and a low frame-rate depth video. We propose a cheaper solution that enhances depth video obtained by TOF sensor by combining it with CCD camera. The proposed method provides high quality video as a cheaper solution for low resolution, and low frame-rate depth video. It is useful when depth video is used in various applications such as 3DTV, free-view TV, teleconference system. High-quality depth video can be obtained by motion compensated frame interpolation (MCFI) and extended Joint Bilateral Upsampling (JBU). Experimental results show that depth video obtained by the proposed method has satisfactory quality.

Index Terms— Depth video, CCD camera, resolution, frame-rate, up-conversion

1. INTRODUCTION

Recently, depth sensors have been widely used in computer vision research fields. The depth sensors are generally classified into three categories; laser scanning method, stereoscopic method and range sensor method using Time-of-Flight (TOF) sensor. Laser scanning method has the advantage of an accurate reconstruction of 3D object, but the acquisition process is time-consuming and the device is also very expensive. Stereoscopic method estimates the disparity map using two or more cameras. Moreover, it is not accurate especially at textureless and occlusion regions, and computational complexity is also high. Range sensor methods using TOF sensors estimate the distance between sensor and object by laser. It is cheaper than laser scanner device and can be used in real time application. However, because of physical limit of depth sensor, it just provides low resolution and low frame-rate depth video and noisy results at an object that has high reflectance. There is a depth sensor such as Z-cam that provides relatively high resolution depth map, but cost is very high [1]. Although mini version of Z-cam is recently introduced, cost and resolution is not still satisfactory to be used in general [2].

In this paper, the proposed method is meaningful in a sense that existing image up-conversion concept is expanded into video. CCD camera provides sufficiently high resolution and frame-rate video to be used in various applications such as 3DTV, FTV and teleconferencing system. Depth video can also be improved by the proposed method with CCD camera. Fig. 1 shows the overall framework of the proposed method. Depth map has characteristics that most regions are very homogeneous. That is, most energies in frequency domain are concentrated at low frequencies, which is different from the natural images captured by CCD camera. It makes motion compensated frame interpolation (MCFI) and upsampling process of depth video insensitive to error. Thus, we can implement those processes more easily than those of CCD images.

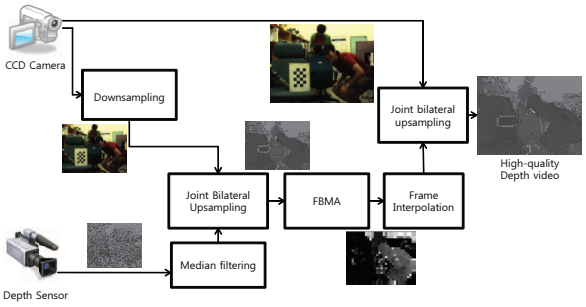

Fig. 1. Overall framework of the proposed method

The remainder of this paper is organized as follows. In Section 2, we introduce the background for the fusion of sensors and MCFI. The proposed method will be described in Section 3. Finally, we present some experimental results and conclusion in Section 4 and 5, respectively.

2. BACKGROUND

2.1. Fusion of CCD Image and Depth Map

Many methods have been proposed to combine depth information with the information of CCD images for enhancing the resolution of depth map [3], [4], [7]. Depth image is generally homogeneous except at object boundaries. Thus, when upsampling process is applied to depth map, it is important to preserve edges in the depth map well to obtain high quality depth map. In order to obtain high quality depth map, bilateral filtering of the cost volume with sub-pixel estimation [3] and Joint Bilateral Upsampling (JBU) methods [4] were proposed. They are based on bilateral filters and preserve edges of depth map very well [7]. However, since texture copying problem occurs frequently, Sebastian et al. proposed the method that uses several depth images for the reconstruction of superresolution depth image in order to avoid this problem [5], [6]. However, it has very high complexity and the difficulty of using in the textureless region.

2.2. Motion Compensated Frame Interpolation

Motion estimation is the most important process in MCFI. In video, a new intermediate frame is synthesized by frame interpolation. Full-search Block Matching Algorithm (FBMA) is generally used for video coding. However, the accuracy of motion vector is dependent on the block sizes since motion estimation is performed for each block. The methods based on Dynamic Programming (DP) [8], [9] and Belief Propagation (BP) [10] can estimate more accurate motion vectors than FBMA, but computational complexity is very high. Many frame interpolation algorithms which are robust to the motion error have been proposed. Most of these algorithms can be applied to FBMA which is widely used in video coding. The motion vectors are used for frame interpolation based on the assumption that all

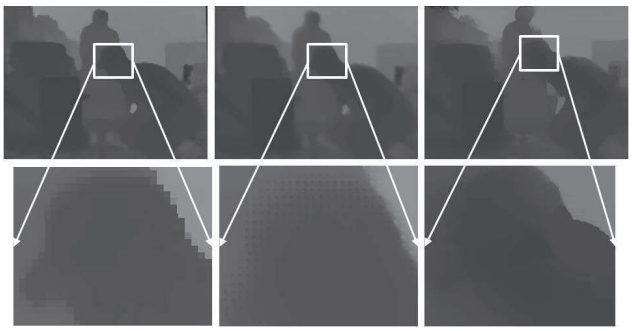

Fig. 2. (*left*) NN interpolation (*middle*) Gaussian interpolation (*right*) Our method

motion vectors are correct and can directly be used for frame interpolation. However, motion vectors may be unreliable. Thus, Huang et al. proposed the method, which uses reliable motion vectors for frame interpolation [11]. Overlapped block motion compensation algorithm by Orchard et al. [12] and adaptive method by Choi et al. [13] were also proposed.

3. PROPOSED METHOD

3.1. Up-conversion of Depth Video in Spatial Domain

In order to combine the information of depth map with that of CCD image, we should warp the depth value of the corresponding pixel in CCD image. By using the assumption that depth and CCD cameras are closely located each other, we can calculate the homography matrix between CCD camera and depth sensor. Because sensors are located very closely each other, we can ignore a little translation and consider only rotation. Depth values can be warped into CCD image by using homography matrix calculated by the affine transform. Alternative warping uses the relationship between coordinate of each sensor [15]. However, this method has higher complexity than warping using homography matrix. The camera calibration process is also needed.

Bilateral filter is an edge-preserving filter [7]. Whereas only one image or two images with the same size are used in bilateral filter, JBU uses two different size images [4]. JBU has an advantage that makes low resolution image into high resolution. It preserves edges well while homogeneous region is smoothed. The equation for JBU is as follows:

$$\overline{R_p} = \frac{1}{k_p} \sum_{q_\downarrow \in k} R_{p_\downarrow} f(||p_\downarrow - q_\downarrow||) g(||\overline{I_p} - \overline{I_q}||) \tag{1}$$

We can obtain a high resolution image $\overline{R_p}$ with low resolution image $R_{p_\downarrow}$ by using two kinds of filters. First one is the spatial filter kernel $f(||p_\downarrow - q_\downarrow||)$ representing spatial distance in low resolution image. The other one is the range filter kernel $g(||\overline{I_p} - \overline{I_q}||)$ representing the difference of pixel value in high resolution image. k_p is a normalization factor. Depth map has the characteristic that most regions are homogeneous because the depth is similar in the same object. Thus, if edges in depth map are preserved well enough, upsampling result of depth map will be better than that of the color image. It makes JBU of depth map possible. Upsampled result by our method is more accurate than those obtained by other upsampling methods such as nearest neighborhood (NN), Gaussian interpolation as shown in Fig. 2. Edge in the depth map is the most important factor in the evaluation of its quality. Thus, our method can be applied to improve the resolution of the image efficiently.

However, if Eq. (1) is applied to low resolution depth map, texture-copying problem may occur as shown in Fig. 3 since only

Fig. 3. Texture copying problem

range filter kernel of CCD image is considered in this equation. Texture of CCD image is copied to upsampled depth image in regions that have similar depth values but different color values. In order to avoid this problem, noise aware filter was proposed [16]. We use the modified noise aware filter for lowering the complexity as follows:

$$\begin{aligned} \overline{R_p} = \tfrac{1}{k_p} \sum_{q_\downarrow \in k} R_{p_\downarrow} f(||p_\downarrow - q_\downarrow||) \times \\ (\alpha g(||\overline{I_p} - \overline{I_q}||) + (1-\alpha) h(||\overline{I_{p_\downarrow}} - \overline{I_{q_\downarrow}}||)) \end{aligned}$$

$$\begin{cases} \alpha = 1, (\overline{I_{q_\downarrow,\max}} - \overline{I_{q_\downarrow,\min}}) \geq threshold \\ \alpha = 0, (\overline{I_{q_\downarrow,\max}} - \overline{I_{q_\downarrow,\min}}) < threshold \end{cases} \tag{2}$$

Range filter kernel is applied not only to CCD image but also to depth map. When the difference between minimum and maximum values in the window of depth map is lower than a threshold value, α is set to 0 and this region can be considered as homogeneous region. Thus, JBU is performed by using a range filter kernel of depth map. α is the weighting factor and threshold value is empirically decided. If range filter kernel of CCD image is also considered in this case, texture of CCD image is propagated into the upsampled version of depth map. Otherwise, α is set to 1, and this region can be considered as edge region. Thus, only range filter kernel of CCD image is used. If range filter kernel of depth map is also considered in this case, the quality is decreased in the upsampled version of depth map. In [16], α is defined as blending function. However, we simply modified the equation because the results are almost same as those in [16] although α is binary value, and the complexity can be reduced. Before upsampling is performed, we apply median filter into the original depth map to suppress the salt and pepper noise frequently appeared due to the physical limit of depth sensor. CCD image is downsampled to almost same size as the original depth map. It is computationally efficient to perform MCFI with the small size CCD image. It makes the intermediate depth map better becuase estimated motion could be more accurate. Thus, the original depth map is upsampled by downsampled CCD image. Using these results, an intermediate depth map can be interpolated, and then an intermediate depth map is upsampled once again by full size CCD image using JBU as shown Fig. 1. This algorithm eventually provides the full size depth map without degradation of quality. In this process, fast implementation of bilateral filter which consists of IIR filter and subsampling concept of image can be used for reducing the computation time [14], [17].

3.2. Up-conversion of Depth Video in Temporal Domain

In general, the quality of the interpolated image is dependent on motion estimation, because accurate motion vector may improve the quality of the interpolated image. However, depth map is slightly different from CCD image, because depth map is less sensitive to the error of motion vector. Fig. 4 shows the results with BP and FBMA. Although BP is more accurate algorithm, complexity of BP is very high and there is not much difference between BP and FBMA in the resulting depth map. Thus, accurate motion estimation algorithm is not necessary. We set on the basis of CCD image corresponding to

Fig. 4. Interpolated depth map. (*left*)FBMA (*right*)BP

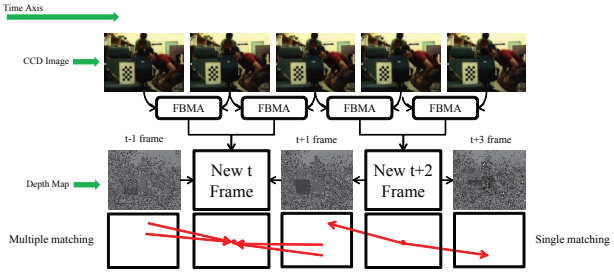

Fig. 5. Matching problem according to the direction of motion estimation.

the interpolated depth map. It is possible because frame-rate of CCD camera is usually higher than that of depth sensor. It can prevent one pixel from matching to multiple pixels. As shown in Fig. 5, when motions are estimated from the neighboring images ($(t-1)$, $(t+1)$ frame) to the CCD image corresponding to the interpolated depth map (t frame), there is a possibility of generating multiple matching for one pixel. However, if estimated direction is inverse, it prevents multiple matching for one pixel. Thus, single matching method can be chosen in order to solve the matching problem.

The proposed method makes unreliable pixels to empty space. In case of depth map unlike CCD image, most regions are homogeneous. Thus, depth value at empty pixel can be interpolated by the neighboring reliable motions. In our method, intermediate depth map is accurately interpolated through bidirectional MCFI, since more motion vectors can be used compared to the conventional MCFI. These concepts are used to up-conversion of depth video in temporal domain. When synthesizing an intermediate frame of depth video by using forward and backward motions, we should determine first whether forward and backward motions in a certain pixel are reliable or not. Our bidirectional adaptive MCFI is computed as follows:

$$f_t^{mid}(p,q) = \tfrac{1}{2} f_{t-1}(p - mv_{1x}, q - mv_{1y}) + \tfrac{1}{2} f_{t+1}(p + mv_{2x}, q + mv_{2y})$$

$$f_t(p,q) = \begin{cases} f_t^{mid}(p,q) & ||mv_1 - mv_2|| \le \varepsilon_1 \\ & and \ ||I_{p-mv1} - I_{p+mv2}|| \le \varepsilon_2 \\ null & otherwise \end{cases} \tag{3}$$

We assume that the depth video has linear motion. Thus, if the difference between forward and backward motion is above the threshold value, it is considered as unreliable pixel. We also consider the difference between values of pixels moved by motions in the depth map as shown in Eq. (3).

Interpolated depth map can be obtained by using forward and backward motion. mv_1 between t and $(t-1)$ frame is backward motion, and mv_2 between t and $(t+1)$ frame is forward motion. $f_t^{mid}(p,q)$ is an interpolated depth map. We used two threshold values in order to determine reliable motions. If the difference between mv_1 and mv_2 is higher than threshold ε_1 or the difference between

Fig. 6. Comparison of interpolated results. (*left*) Method using all pixels (*right*) Our method

Fig. 7. Up-conversion in Spatial Domain. (*left*) Spatial resolution enhanced depth map (*right*) Original depth map.

intensity values at pixels moved by motions mv_1, mv_2 is higher than threshold ε_2, it remains an empty pixel. Empty pixels of $f_t(p,q)$ are filled with the algorithm used in Sec. 3.1. Threshold values ε_1 and ε_2 are determined empirically. Thus, final equation is as follows:

$$\overline{R_p} = \tfrac{1}{k_p} \sum_{q \in k} D(p,q) R_p f(||p-q||) \times (\alpha g(||\overline{I_p} - \overline{I_q}||) + (1-\alpha) h(||\overline{I_p} - \overline{I_q}||)) \tag{4}$$

$$\begin{cases} D(p,q) = 0, & f_t(p,q) = null \\ D(p,q) = 1, & f_t(p,q) = f_t^{mid}(p,q) \end{cases}$$

$D(p,q)$ is a decision factor that decides whether a pixel (p,q) is reliable or not. $\overline{R_p}$ is the value in unreliable pixel that will be filled with reliable pixels. In other words, Eq. (4) is applied only for empty pixels in order to fill the null space with neighboring reliable value. As shown in Fig 6, using only reliable pixels makes the quality of intermediate depth map better compared to the case of using also unreliable pixels. Although FBMA is less accurate in estimating the motion than other methods such as BP or DP, we can use it to apply for video coding system because it is possible to implement efficiently.

4. EXPERIMENTAL RESULTS

The proposed method has been implemented in Visual C++ 6.0, and tested on computer with Intel Core2 Quad 2.5GHz processor and 2GHz RAM. In the experiment, Flea camera made in Point Grey Research Inc. and SR3000 depth sensor made in SwissRanger Inc. were used. The resolution of Flea camera is 1024X768 and frame-rate is 30Hz whereas the resolution of depth sensor is 176X144 and frame-rate is 15Hz. Up-conversion results in the spatial domain is shown in Fig. 7. Depth image of 176X144 is enlarged into 1024X768 which is the same size as CCD image while maintaining

Table 1. Processing time in each step

Step	1st JBU	FBMA	MCFI	2nd JBU	Total
Time(sec)	0.892	0.046	0.106	1.902	2.946

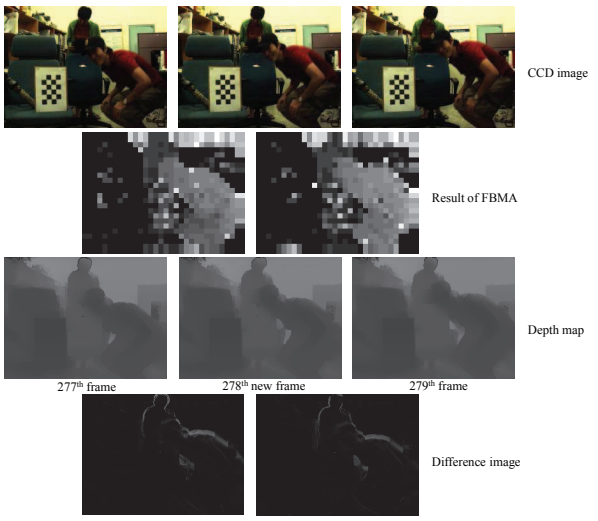

Fig. 8. Up-conversion in Temporal Domain

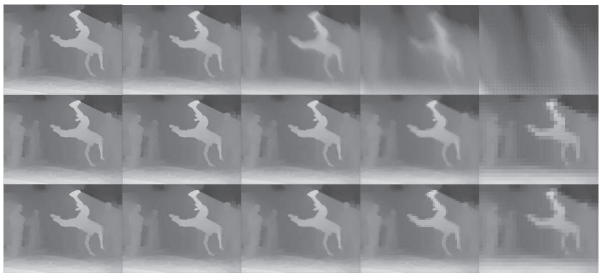

Fig. 9. Up-conversion result. (*top*)Gaussian interpolation (*middle*) Nearest Neighborhood interpolation (*bottom*) Our method

the quality. We can confirm that texture copying problem is removed. Up-conversion result in temporal domain is shown in Fig. 8. Motion information estimated from the CCD video using FBMA is used for interpolating an intermediate depth map (the 278th frame). The frame-rate of depth sensor increases to 30 Hz which is the same level as CCD camera.

"Breakdancer" sequences provided by Microsoft Research are also used in the experiments. The 87th frame of "Breakdancer" is upsampled according to the interpolation method and up-conversion rate as shown in Fig. 9. As the original image size is smaller, that is, up-conversion rate is higher, the quality of upsampled image is decreased as shown in Fig. 10. When up-conversion rate is 32, our method produces poor results because the estimated motion is inaccurate in the downsampled image of very small size. However, as shown in Fig. 10, up to up-conversion rate is 8, interpolated results maintain almost same quality compared to other methods. Thus, our experiment is performed successfully because size of CCD image is quadruple of that of depth map, that is, up-conversion rate is 4. We can also confirm that our result is almost same as the ground truth map of the 96th frame as shown in Fig.11 when the up-conversion

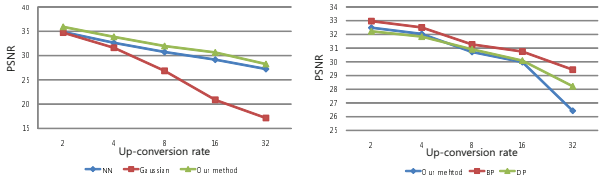

Fig. 10. Comparison of PSNR according to (*left*) interpolation methods (*right*)motion estimation methods used in MCFI

Fig. 11. Comparison the result with ground truth. (*left*) Result in up-conversion rate 4 (*middle*) Ground truth map (*right*) The difference image

rate is 4. We use fast implementation of bilateral filter to measure the processing time, which is shown in Table 1. If it is tested on GPU, it is possibly implemented in real time.

5. CONCLUSION

In this paper, we propose the novel method that overcomes the physical limits of TOF sensor. By using the proposed framework, we can convert low resolution, low frame-rate depth video into high quality depth video, and show the possibility applicable to computer vision system such as 3DTV broadcasting in real time. However, it is difficult to assess the quality of the result obtained by the proposed method since there is no ground truth map. In order to obtain the ground truth map, accurate depth acquisition device is needed such as laser scanner or Z-cam. For further research, high-quality depth video obtained by the proposed method will be used for subjective assessment of 1-view and 1-depth videos. In addition, when the fast algorithm of JBU is implemented, texture copying problem can not be reduced. Thus, this problem will be solved as further works.

6. REFERENCES

[1] G. J. Iddan, and G. Yahav, "3D imaging in the studio (and elsewhere)," *Proc. SPIE*, pp. 48-55, Jan. 2001.

[2] 3dv systems, z-cam. http://www.3dvsystems.com.

[3] Q. Yang, R. Yang, J. Davis, and D.Nister. "Spatial-depth super resolution for range images," *IEEE Proc. CVPR*, 2007.

[4] J. Kopf, MF. Cohen, D. Lischinski and M. Uyttendaele, "Joint bilateral upsampling," *ACM SIGGRAPH*, 2007.

[5] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. "High-quality scanning using time-offlight depth superresolution," *IEEE CVPR Workshop on Time-of-Flight Computer Vision*, 2008.

[6] S. Borman and R. L. Stevenson. "Super-resolution from image sequences - a review," *Proc. Midwest Symp. Circuits and Systems*, 1998.

[7] C. Tomasi and R. Manduchi. "Bilateral filtering for gray and color images," *IEEE Proc. ICCV*, pp. 839-846, 1998.

[8] M. Gong and Y. Yang, "Fast Stereo Matching Using Reliability-Based Dynamic Programming and Consistency Constraints," *IEEE International Conference on Computer Vision*, pp.610, vol. 1, 2003.

[9] L. Wang, M. Liao, M. Gong, R. Yang and D. Nister, "High-Quality Real-Time Stereo Using Adaptive Cost Aggregation and Dynamic Programming," *3DPVT Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission*, pp. 798-805, 2006

[10] J Sun, NN Zheng and HY Shum, "Stereo Matching Using Belief Propagation," *IEEE Trans. PAMI*, vol. 25, no. 7, pp. 787-800, 2003.

[11] Huang A and Nguyen TQ, "A multistage motion vector processing method for motion-compensated frame interpolation," *IEEE Trans. Image Process*, pp. 694-708, 2008.

[12] M. T. Orchard and G. J. Sullivan, "Overlapped block motion compensation: An estimation-theoretic approach," *IEEE Trans. Image Process*, vol. 3, no. 5, pp. 693-699, Sep. 1994.

[13] Choi, B.-D, Han, J.-W, Kim, C.-S and Ko, S.-J., "Motion-Compensated Frame Interpolation Using Bilateral Motion Estimation and Adaptive Overlapped Block Motion Compensation," *IEEE Trans. CSVT*, vol. 17, pp. 407-416, April. 2007.

[14] S. Paris and F. Durand. "A Fast Approximation of the Bilateral Using a Signal Processing Approach," *Proc. ECCV*, 2006.

[15] J. Zhu, L. Wang, R. Yang, J. Davis, "Fusion of Time-of-Flight Depth and Stereo for High Accuracy Depth Maps," *IEEE Proc. CVPR*, 2008.

[16] D. Chan, H. Buisman, C. Theobalt, S. Thrun, "A Noise-Aware Filter for Real-Time Depth Upsampling", *M2SFA2 2008: Workshop on Multi-camera and Multi-modal Sensor Fusion*, 2008.

[17] Ian T. Young and Lucas J,van Vliet, "Recursive implementation of the Gaussian filter", *Signal Processing*, vol. 44, Issue. 2, pp. 139-151, 1995.