Temporally consistent depth video filter using temporal outlier reduction

Sangbeom Lee & Yo-Sung Ho

Signal, Image and Video Processing

ISSN 1863-1703

SIViP
DOI 10.1007/s11760-014-0642-7

Springer

SIViP
DOI 10.1007/s11760-014-0642-7

ORIGINAL PAPER

Temporally consistent depth video filter using temporal outlier reduction

Sangbeom Lee · Yo-Sung Ho

Received: 27 December 2012 / Revised: 20 June 2013 / Accepted: 25 November 2013

Abstract In this paper, we propose a temporal filtering algorithm to obtain the temporally consistent depth video. We apply a joint multilateral filter to depth video that is based on three Gaussian distributions: one spatial filter and two range filters for color and depth similarities. We also perform outlier reduction in the temporal domain using Gaussian-weighted least squares fitting method to deal with temporal displacement. Experimental results show that the proposed method improves depth quality by 1.22 dB on average, compared to other algorithms due to enhanced temporal consistency.

Keywords Depth video filter · Joint multilateral filter · Gaussian-weighted least squares · Temporal consistency · Three-dimensional video

1 Introduction

Advancements in three-dimensional (3D) video technologies enable us to reproduce and experience simulations of reality. In other words, the gap between the real world and virtual environments is getting closer. Owing to the rapid development of 3D displays, i.e., stereoscopic or auto-stereoscopic displays, 3D video technologies can provide us a feeling of "being there," or presence, from the simulated reality [1,2].

S. Lee (✉)
Electronics and Telecommunications Research Institute (ETRI),
1, 10-gil, Techno sunhwan-ro,Yuga-myeon,
Dalseong-gun 711-880, Daegu, Korea
e-mail: sblee230@etri.re.kr
URL: http://vclab.gist.ac.kr

Y.-S. Ho
Gwangju Institute of Science and Technology (GIST),
123 Cheomdangwagi-ro, Buk-gu, Gwangju 500-712, Korea
e-mail: hoyo@gist.ac.kr

Figure 1 shows the entire process of the 3D video system, which includes the whole processes of acquisition, processing, transmission, and rendering of 3D images including N-view color and depth videos. We produce the 3D video by utilizing various types of cameras, such as stereoscopic cameras, multi-view cameras, or depth cameras. In case of depth cameras, the depth map can be acquired directly. Otherwise, the depth information is estimated by stereo matching algorithms.

Recently, depth image-based rendering (DIBR) has received industrial attention due to its production of natural virtual images based on color and depth videos [3]. Virtual images are generated at intermediate virtual viewpoints between two real cameras. Since depth data accuracy directly affects the rendering performance of virtual views, accurate depth information is crucial.

In computer vision and image processing, researches on acquiring reliable depth information have lasted for several decades; yet, there are still many remaining problems. Among them, the temporal inconsistency problem in depth data is caused by an independent process for frame-by-frame of depth sensing methods. As a result, depth data becomes fluctuated, inducing discomfort of human eyes. Figure 2 shows a depth sequence with three consecutive frames captured by a time-of-flight depth camera [4]. At the flowerpot-plant boundary, marked by rectangles, inconsistent depth data can be observed despite the static scene.

In this paper, we propose a new algorithm for temporally consistent depth video with the aid of its corresponding color video. The main contribution of our work is to formulate the temporal inconsistency problem using *Gaussian-weighted least squares (GWLS)*. First, we employ joint multilateral filter with three Gaussian weighting functions to the depth video. Then, the outlier reduction process using GWLS is performed to deal with temporal displacement.

Published online: 01 June 2014

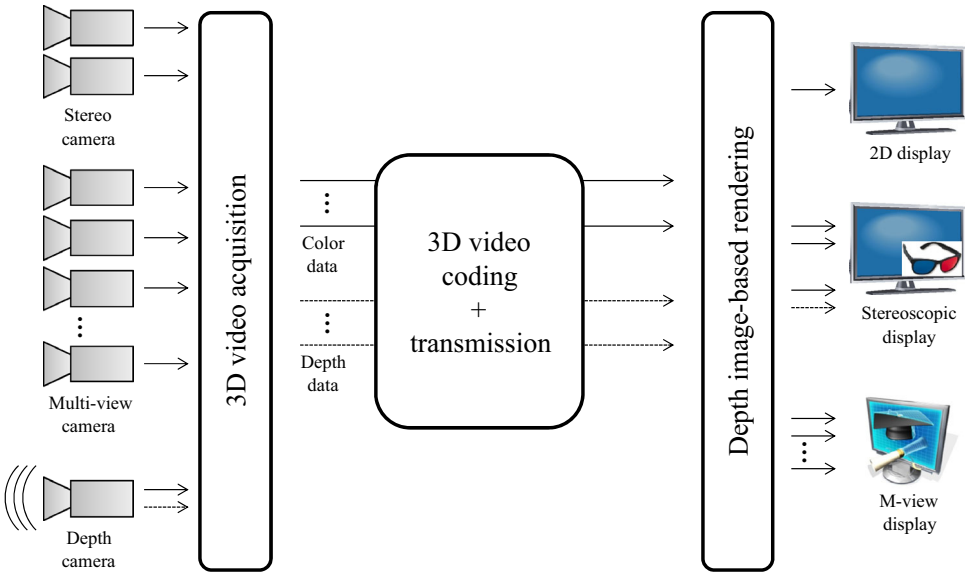

Fig. 1 Three-dimensional video system

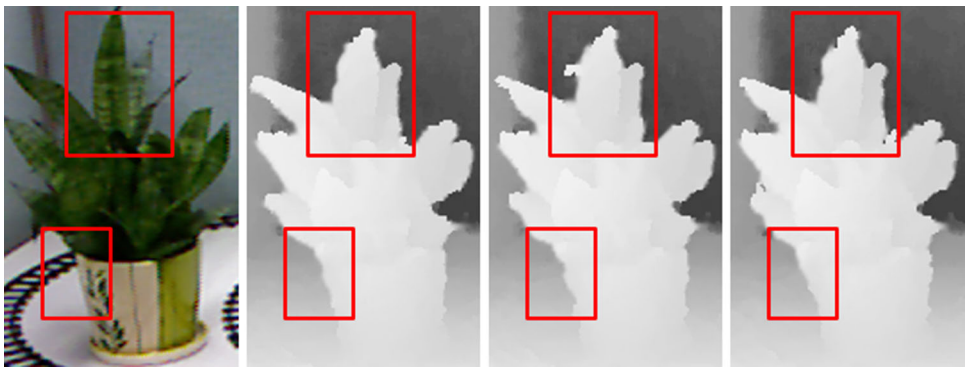

Fig. 2 Temporal inconsistency of three consecutive depth maps

The remainder of this paper is organized as follows. In Sect. 2, we introduce the conventional researches for temporal consistency of depth video. Section 3 describes the proposed method in detail. In Sect. 4, the experimental results are exhibited, and finally, conclusion is drawn in Sect. 5.

2 Related works

In the literature, two approaches exist in regard to handling temporal inconsistency of depth video: *dynamic depth estimation* and*depth video filtering*.

The former extends the energy function for depth estimation to the temporal domain. Tao et al. have presented a segment-based depth estimation algorithm under the assumption that the 3D scene is composed of many piecewise planes [5]. Depth fluctuation is reduced by the energy minimization process of the target segment, which utilizes segments of previous and successive frames. Larsen et al. have developed enhanced belief propagation for reconstructing temporally consistent depth data [6].

The latter, in which the proposed method belongs, performs data-adaptive kernel filtering of depth video. Joint bilateral filtering is prominent method in this category [7]. The joint bilateral filter (JBF) exploits spatial and range weighting functions derived from the coordinate distances and photometric similarity between a target pixel and its neighbors.

In the depth map, suppose there exist a target pixel p and one of its neighbors q. S_p and S_q are depth values at p and q. In addition, I_p and I_q are the associated color values at p and q. The new depth value $\tilde{S}_p$ via JBF is computed by

$$\tilde{S}_p = \frac{\sum_{q\in\Omega} S_q \cdot f\left(\|p-q\|\right) g\left(\left\|I_p - I_q\right\|\right)}{\sum_{q\in\Omega} f\left(\|p-q\|\right) g\left(\left\|I_p - I_q\right\|\right)} \tag{1}$$

where f and g indicate spatial and range filters, respectively. Ω is the local kernel size. If Gaussian distribution is used to model such weighting functions, they are represented by

$$f(x) = \exp\left(-\frac{x^2}{2\sigma_f^2}\right), \quad g(x) = \exp\left(-\frac{x^2}{2\sigma_g^2}\right) \tag{2}$$

where σ_f and σ_g are standard deviations of f and g, respectively.

Several algorithms using JBF have been proposed. Lai et al. have employed iterative joint multilateral filtering (JMF) which adds hard-thresholding of depth data to the conventional JBF [8]. Yang et al. have introduced a modified JBF (MJBF) [9]; this method puts an argument of the minimum depth when determining one of the 4-connected neighboring depth candidates.

Recently, a 3D JBF-based approach regarding temporal consistency has been proposed [10]. Specifically, filtering is extended to the temporal domain to reduce temporal fluctuation. Range filters for color and depth data are adaptively applied based on depth distribution inside the filter kernel. However, lack of handling temporal motion causes motion blur artifacts.

3 Proposed methods

In this section, we describe a JBF-based algorithm for temporally consistent depth video. Unlike other algorithms, the proposed method takes motion information into account, targeted for dynamic objects and moving camera environment.

3.1 Joint multilateral filter

In general, depth video accuracy is critical in many 3D video applications. Although JBF reduces depth errors near object boundaries, the frame-to-frame depth data still fluctuates. Thus, we use a different approach to improve temporal consistency, applying JMF.

Under the assumption that the color discontinuity can be used for correcting depth discontinuity, JMF exploits three Gaussian distributions: one spatial filter and two range filters which observe photometric and depth similarities. The new depth value at position p in time t $\tilde{S}_{p,t}$ is calculated by

$$\tilde{S}_{p,t} = \arg\min_{d \in N_d} \frac{\sum_{q \in \Omega} \sum_{n \in \Pi} W_{q,n} \cdot C_{q,n,d}}{\sum_{q \in \Omega} \sum_{n \in \Pi} W_{q,n}} \tag{3}$$

where t and n represent target and neighboring frames. Π is the temporal kernel size. N_d is a set of depth candidates, d, which includes 4-connected neighbors of $S_{p,t}$ and additional two neighbors, $S_{p,t-1}$ and $S_{p,t+1}$. In order to obtain clear depth boundary without blurring artifacts, the filtered depth value is determined by one of the six candidates as described in (3). $W_{q,n}$ is defined as follows:

$$W_{q,n} = f\left(\|p_t - q_n\|\right) g_I\left(\left\|I_{p,t} - I_{q,n}\right\|\right) g_S\left(\left\|S_{p,t} - S_{q,n}\right\|\right) \tag{4}$$

where g_I and g_S represent range filters for color and depth similarities. Euclidean distance is used for color similarity. We used a truncated linear model as a cost function. Since the truncated linear model allows for depth discontinuities, the cost function is robust, becoming constant as the difference becomes large. The cost function $C_{q,n,d}$ using a truncated linear model is calculated as follows:

$$C_{q,n,d} = \min\left(\lambda L, \left\|S_{q,n} - d\right\|\right) \tag{5}$$

where λ is a constant to reject outliers. L denotes the depth range and it is typically set to 256 since depth maps are represented by 8-bit gray scale. In the proposed algorithm, the

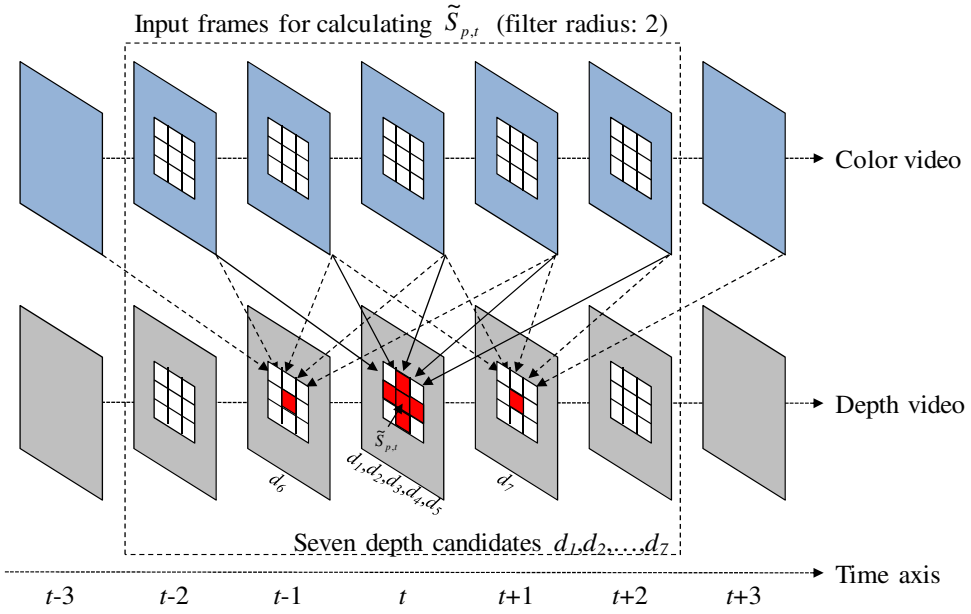

Fig. 3 Filtering process of the proposed method

cost linearly increases based on the distance between the depth candidate d and $S_{q,n}$ up to λL that controls when the cost stops increasing. Figure 3 shows the filtering process of the proposed method with a filter radius 2 in the temporal domain.

3.2 Outlier reduction in temporal domain

The simple extension of JMF to the temporal domain is not sufficient for dynamic scenes due to motion blur and temporal outliers, as shown in Fig. 4. Figure 4b is obtained by Depth Estimation Reference Software (DERS) which uses the graph

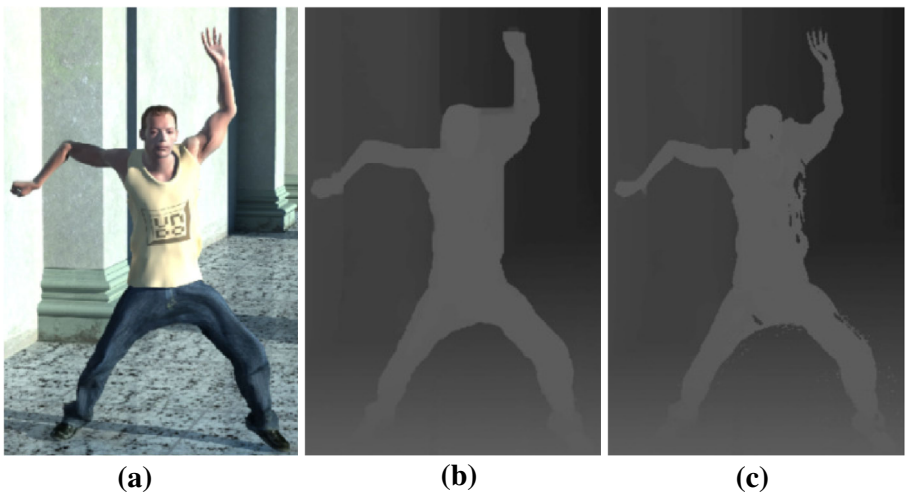

Fig. 4 Filtering errors of the simple JMF. **a** Original *color image*. **b** Result for DERS. **c** Result with outlier (color figure online)

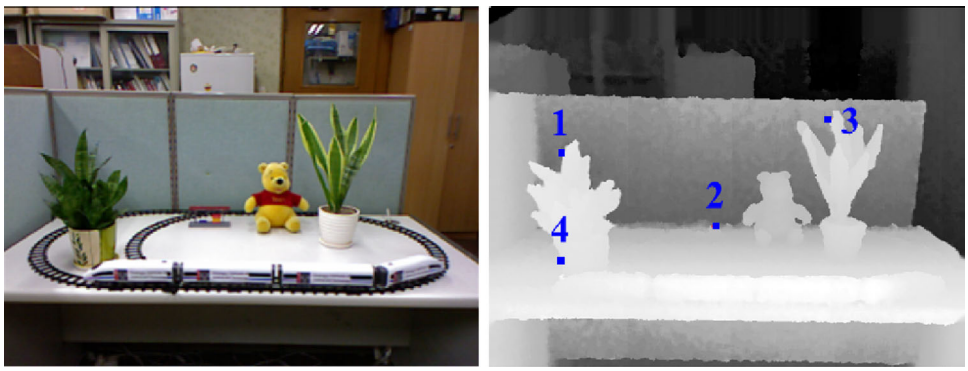

Fig. 5 Images captured by a depth camera. **a** Color image. **b** Depth map (color figure online)

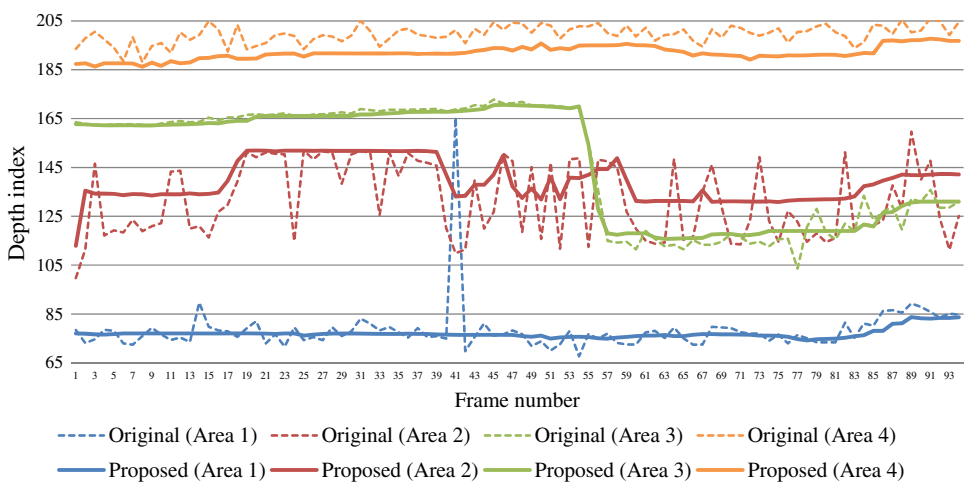

Fig. 6 Average depth data of four blocks

cut optimization [11]. In Fig. 4c, depth data inside the object are deformed due to temporal outliers.

In order to reduce temporal outliers, initially we add a range filter for depth similarity; ample depth differences can be removed including temporal outliers. It is defined as follows:

$$g_S\left(\left\|S_{p,t}-S_{q,n}\right\|\right)=\exp\left(-\frac{\left\|S_{p,t}-S_{q,n}\right\|^2}{2\sigma_S^2}\right) \tag{6}$$

where σ_S represents the standard deviation of depth.

As an additional process of temporal outlier reduction, we exploit Gaussian-weighted least squares (GWLS) fitting method. We assume that in static areas, color (or depth) data at the same position linearly change in the temporal domain. Hence, the estimate of linear model coefficients can be calculated. Another assumption is that estimation of linear model coefficients depends on color (or depth) difference and temporal distance with respect to the data at time t. Therefore, we adopt GWLS fitting method where the Gaussian weighting factor is included in the fitting process.

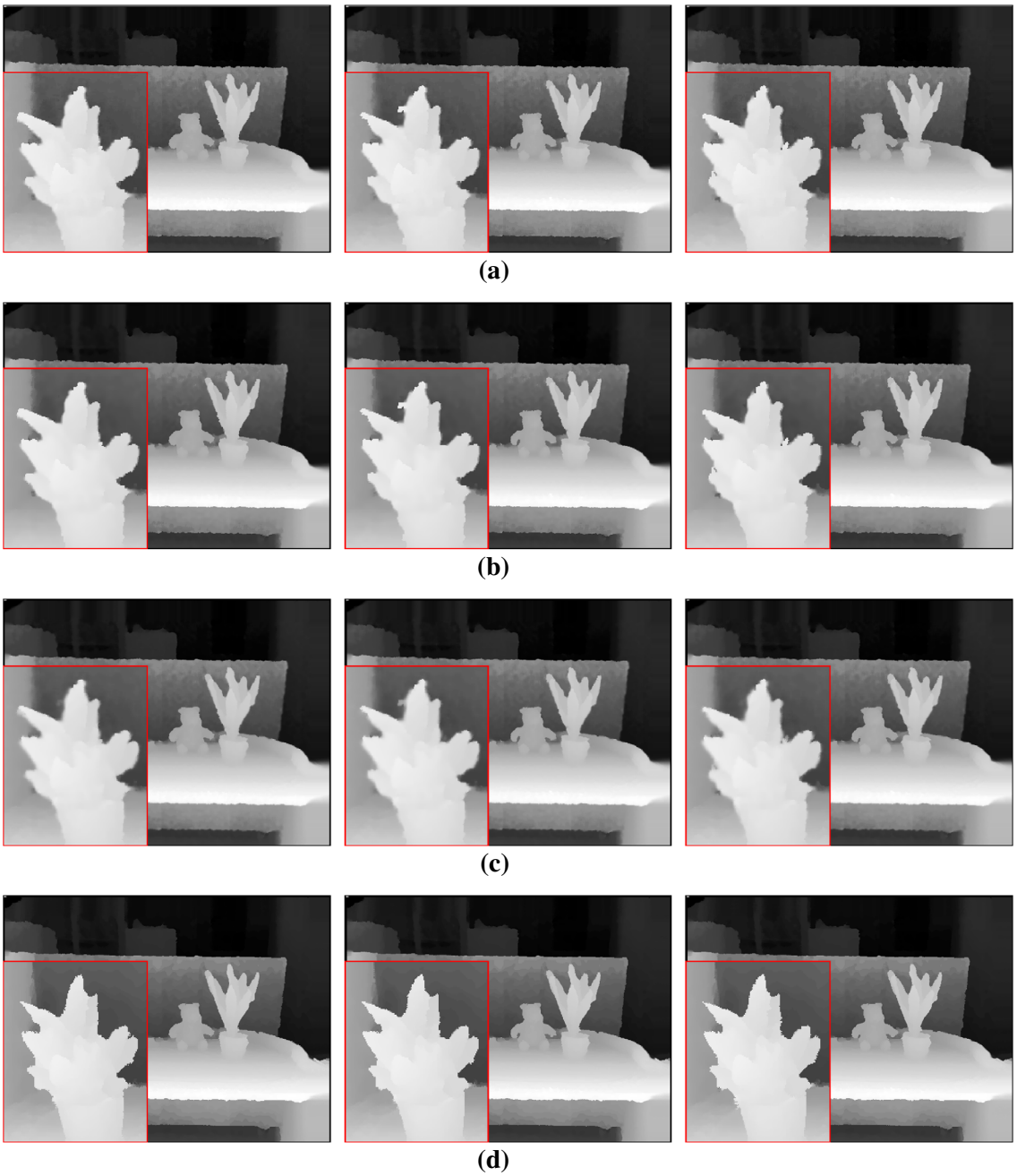

Fig. 7 Effect of the depth video filter on depth sequence captured by the time-of-flight depth camera for 170th, 171st, and 172nd frame. **a** Preprocessed depth sequence. **b** Result for JMF. **c** Result for 3D JBF. **d** Result for the proposed method

To illustrate the GWLS fitting process, suppose we have color (or depth) data that can be modeled by a first-degree polynomial as follows:

$$y = b_1 x + b_2 \tag{7}$$

where x and y represent time coordinate and color (or depth) data within the temporal kernel size Π, respectively. In order to solve this equation for the unknown coefficients b_1 and b_2, suppose that the system *SSE* (sum of square error) is defined as the summed square of the residuals as follows:

$$SSE = \sum_{n\in\Pi} w_n \{y_n - (b_1 x_n + b_2)\}^2 \tag{8}$$

Since such a process minimizes S, the coefficients are determined by differentiating *SSE* with respect to each parameter, i.e., b_1 and b_2, with setting each result to zero.

$$\frac{\partial SSE}{\partial b_1} = -2\sum_{n\in\Pi} x_n w_n \{y_n - (b_1 x_n + b_2)\} = 0$$
$$\frac{\partial SSE}{\partial b_2} = -2\sum_{n\in\Pi} w_n \{y_n - (b_1 x_n + b_2)\} = 0 \tag{9}$$

From the simultaneous equation, we obtain b_1 as follows:

$$b_1 = \frac{\sum_{n\in\Pi} w_n \sum_{n\in\Pi} w_n x_n y_n - \sum_{n\in\Pi} w_n x_n \sum_{n\in\Pi} w_n y_n}{\sum_{n\in\Pi} w_n \sum_{n\in\Pi} w_n x_n^2 - \left\{\sum_{n\in\Pi} w_n x_n\right\}^2} \tag{10}$$

Solving for b_2 using the b_1:

$$b_2 = \frac{\sum_{n\in\Pi} w_n y_n - b_1 \sum_{n\in\Pi} w_n x_n}{\sum_{n\in\Pi} w_n} \tag{11}$$

Here, Gaussian weights are calculated as follows:

$$w_n = \exp\left\{\frac{\|x_n - x_t\|^2}{2\sigma_x^2}\right\} \exp\left\{\frac{\|y_n - y_t\|^2}{2\sigma_y^2}\right\} \tag{12}$$

where x_t and y_t are time coordinate and color (or depth) data at time t.

We obtain the estimated linear model from the above process. In order to distinguish temporal outliers, we calculate the distance between each point and the fitted line. If the distance is larger than a certain threshold, the point is regarded as an outlier. The threshold is defined as T_{GWLS} which represents the distance between the data coordinates at time t and the fitted line as follows:

$$T_{GWLS} = \frac{|b_1 x_t - y_t + b_2|}{\sqrt{b_1^2 + 1}} \tag{13}$$

The GWLS fitting process is performed to both color and depth data. Frames containing temporal outliers for color and depth data are excluded and the remaining frames are used for the filtering process.

4 Experimental results and analysis

In order to evaluate the proposed method, we acquire depth videos from both active and passive sensor-based method. We applied MJBF, JMF, 3D JBF, and the proposed method for comparison. The standard deviations of the proposed method were set to 10 and 255 for color and depth, respectively. Through several experiments, we have confirmed that we obtained the best depth quality when the constant parameter λ for outlier rejection is set to be 0.1. In other words, we regard the depth discontinuity as 10 % of the depth range.

4.1 Experiments for active sensing method

We conducted experiments using the Kinect depth camera with a built-in structured light sensor, which captures both color and depth images in 640×480 resolution. Since the capturing process of the Kinect depth camera has difficulties in some areas, such as occlusion areas caused by different viewpoints of the sensor transmitter and receiver, black-colored areas, slanted, or shiny surface, etc., the depth map presents a lot of holes with black intensity. Therefore, these holes must be filled before we apply the proposed algorithm. In these experiments, we used an in-painting algorithm proposed by

Table 1 Variances of depth data for 100 frames

Area	Original	Proposed method
1	101.3260	3.8489
2	232.0264	69.4549
3	571.6297	510.8932
4	13.3535	7.4868

Table 2 Average PSNR of filtering results with depth video using DERS

Method	Filter radius	Sequence	
		Undo_Dancer	Mobile
DERS	N/A	38.6973	29.3727
MJBF	1	39.0275	29.4972
	2	40.7222	29.6723
	3	40.7386	29.7597
JMF	1	38.6618	29.4971
	2	38.6652	29.5862
	3	38.5531	29.6783
3D JBF	1	39.6497	29.4717
	2	39.9297	29.3825
	3	40.0300	29.7063
Proposed method	1	40.1845	29.7258
	2	41.5713	29.8434
	3	41.7083	29.8589

Telea as a preprocessing step since this is very powerful for recovery of lost or corrupted parts of the image data [12]. Figure 5 shows input color and preprocessed depth images captured by the depth camera.

In order to quantitatively verify the improvement of temporal consistency, we have measured the amount of depth fluctuation by calculating average depths in the 4×4 block as denoted in Fig. 5b.

Figure 6 demonstrates average depth values of four blocks specified in Fig. 5b. The dotted line and the solid line represent the preprocessed depth data and the results for the proposed method, respectively. Figure 7 illustrates the filtering results of three consecutive depth sequences. The variance of depths for 100 frames is given in Table 1. From such results, we verified that the proposed method reduces depth fluctuation, accomplishing temporal consistency.

4.2 Experiments for passive sensing method

Regarding the passive sensing method, we initially performed depth estimation by DERS, then applied filtering algorithms. The original and filtered depth videos were compared in terms of peak signal-to-noise ratio (PSNR). We have evaluated our algorithm using "Undo_Dancer" and "Mobile" sequences provided by Nokia and Philips, respectively. They are synthetic including ground truth depth data

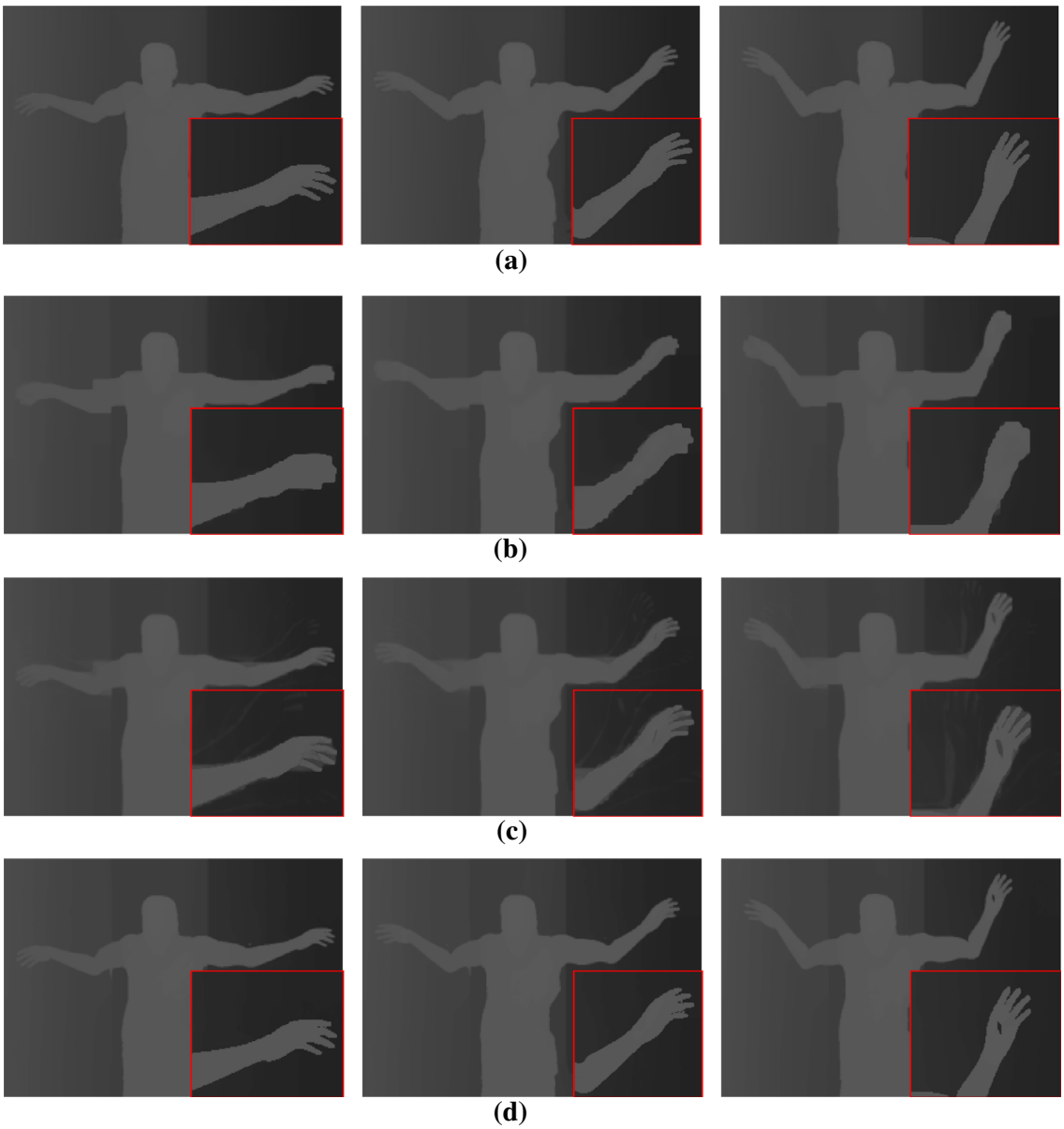

Fig. 8 Effect of the depth video filter on "Undo_Dancer" for 24, 25, and 26th frame. **a** Original depth sequence. **b** Result for JMF. **c** Result for 3D JBF. **d** Result for the proposed method

Table 3 Average PSNR of filtering results with deformed depth video

Method	Filter radius	Sequence	
		Undo_Dancer	Mobile
Noise	N/A	27.0573	27.0802
MJBF	1	36.1632	34.2594
	2	41.6712	37.2594
	3	40.9638	37.8596
JMF	1	27.5580	27.6017
	2	27.7825	27.8387
	3	27.8856	27.9474
3D JBF	1	30.8803	31.1158
	2	31.3984	31.5470
	3	31.5015	31.6247
Proposed method	1	40.0543	40.8923
	2	43.5310	39.8319
	3	43.2088	38.6740

captured in moving camera environment [13,14]. The resolutions are 1,920 × 1,080 for "Undo_Dancer" and 720 × 540 for "Mobile."

The average PSNR for 200 frames is shown in Table 2. From the results, we verified that the proposed algorithm achieves better depth quality compared to conventional algorithms. With a filter radius 3, gains of average PSNR were 1.75, 0.53, 1.67, and 0.92 dB compared to DERS, MJBF, JMF, and 3D JBF, respectively.

Figure 8 shows the effects of the depth video filtering by comparing the depth sequence with three consecutive frames. As shown in Fig. 8d, we proved that the proposed method clearly reconstructs the object boundary and improves temporal consistency compared to other methods.

In order to further evaluate the performance of the proposed method, we conducted experiments using the deformed depth video by additive Gaussian noise with mean 0 and standard deviation 16. Then, we measured PSNR between the ground truth depth video and the filtering results. Table 3 shows average PSNR of filtering results for 200 frames with deformed depth video. As shown in Table 3, the proposed method outperforms the conventional methods. This means that the noise-reducing effect of the proposed method was better than other methods.

5 Conclusions

In this paper, we proposed a temporal filtering algorithm for the depth video using temporal outlier reduction. We applied Gaussian-weighted least squares fitting method to deal with temporal displacement. From our experimental results, we have shown that the proposed method improves the depth quality by 0.91 dB for the depth video obtained by DERS and 8.65 dB for the deformed depth video on average, compared to the conventional algorithms. In terms of temporal consistency, the proposed method reduced depth data fluctuation successfully.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2013-067321)).

References

1. Barfield, W., Weghorst, S.: The sense of presence within virtual environments: a conceptual framework. In: Salvendy, G., Smith, M. (eds.) Human Computer Interaction: Software and Hardware Interfaces, pp. 699–704. Elsevier, Amsterdam (1993)
2. Freeman, J., Avons, S.E.: Focus group exploration of presence through advanced broadcast services. SPIE Hum. Vis. Electron. Imaging **3959**, 530–539 (2000)
3. Fehn, C.: Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3-D TV. SPIE Conf. Stereosc. Disp. Virtual Real. Syst. **5291**, 93–104 (2004)
4. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1297–1304 (2011)
5. Tao, H., Sawhney, H.S., Kumar, R.: Dynamic depth recovery from multiple synchronized video streams. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 118–124 (2001)
6. Larsen, E.S., Mordohai, P., Pollefeys, M., Fuchs, H.: Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In: IEEE International Conference on Computer Vision, pp. 1–8 (2007)
7. Kopf, J., Cohen, M.F., Lischinski, D., Uyttendaele, M.: Joint bilateral upsampling. In: SIGGRAPH, pp. 96–100 (2007)
8. Lai, P., Tian, D., Lopez, P.: Depth map processing with iterative joint multilateral filtering. In: Picture Coding Symposium, pp. 9–12 (2010)
9. Yang, Q., Wang, L., Ahuja, N.: A constant-space belief propagation algorithm for stereo matching. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1458–1465 (2010)
10. Choi, J., Min, D., Sohn, K.: 2D-plus-depth based resolution and frame-rate up-conversion technique for depth video. In: IEEE Transactions on Consumer Electronics, pp. 2489–2497 (2010)
11. Tanimoto, M., Fujii, T., Suzuki, K.: Reference software of depth estimation and view synthesis for FTV/3DV. In: ISO/IEC JTC1/SC29/WG11, M15836 (2008)
12. Telea, A.: An image inpainting technique based on the fast marching method. J. Gr. Tools **9**(1), 25–36 (2004)
13. Hannuksela, M., Rusanovskyy, D.: Extension of existing 3DV test set toward synthetic 3D video content. In: ISO/IEC JTC1/SC29/WG11, M19221 (2011)
14. Bruls, F., Gunnewiek, R.K., Walle, P.: Philips response to new call for 3DV test material: arrive book & mobile. In: ISO/IEC JTC1/SC29/WG11, M16419 (2011)