Reliability-Based Multiview Depth Enhancement Considering Interview Coherence

Jinwook Choi, *Student Member, IEEE,* Dongbo Min, *Member, IEEE,*
and Kwanghoon Sohn, *Senior Member, IEEE*

***Abstract*—Color-plus-depth video format has been increasingly popular in 3-D video applications, such as auto-stereoscopic 3-D TV and freeview TV. The performance of these applications is heavily dependent on the quality of depth maps since intermediate views are synthesized using the corresponding depth maps. This paper presents a novel framework for obtaining high-quality multiview color-plus-depth video using a hybrid sensor, which consists of multiple color cameras and depth sensors. Given multiple high-resolution color images and low quality depth maps obtained from the color cameras and depth sensors, we improve the quality of the depth map corresponding to each color view by increasing its spatial resolution and enforcing interview coherence. Specifically, a new up-sampling method considering the interview coherence is proposed to enhance multiview depth maps. This approach can improve the performance of the existing up-sampling algorithms, such as joint bilateral up-sampling and weighted mode filtering, which have been developed to enhance a single-view depth map only. In addition, an adaptive approach of fusing multiple input low-resolution depth maps is proposed based on the reliability that considers camera geometry and depth validity. The proposed framework can be extended into the temporal domain for temporally consistent depth maps. Experimental results demonstrate that the proposed method provides better multiview depth quality than the conventional single-view-based methods. We also show that it provides comparable results, yet much more efficiently, to other fusion approaches that employ both depth sensors and stereo matching algorithm together. Moreover, it is shown that the proposed method significantly reduces bit rates required to compress the multiview color-plus-depth video.**

***Index Terms*—Color camera, depth sensor, depth up-sampling, interview coherence, reliability.**

I. Introduction

IN THE NEAR future, TV audiences will be able to watch multiview 3-D video without auxiliary devices as well as interact with auto-stereoscopic displays without dedicated controllers. Obtaining high-quality depth map is one of the most important techniques required to realize the next generation of 3-D broadcasting systems, since the depth map directly affects the quality of 3-D video. To represent the depth maps, various approaches have been proposed including color-plus-depth and layered depth maps [1]–[3]. Multiview color/-depth video coding methods are also required to efficiently compress and transmit large quantities of 3-D video over a network [4]–[6].

Depth acquisition methods can be generally classified into three categories: laser scanning methods, stereo matching methods, and range sensing methods. Although the laser scanning methods can provide highly accurate 3-D information, its acquisition process is very time-consuming and thus, is limited in a static scene only. Consequently, these methods are typically used to reconstruct highly accurate 3-D model [7], [8]. In contrast, the stereo matching methods estimate disparity maps using multiple images taken by two or more cameras [9], [10]. A number of algorithms have been proposed by using cost aggregation methods [11]–[13] and global optimization techniques [14]. However, their performance is still not reliable due to many factors, such as a lighting condition and an occlusion problem. Moreover, huge computational complexity is the most critical problem. Especially, its complexity increases linearly proportional to the number of color cameras used in the multiview depth estimation.

Recently, new types of sensors have been developed to overcome the limits of conventional methods [15]–[17]. The range sensing methods utilize a time-of-flight (ToF) principle to estimate a distance between a sensor and an object by extracting phase information from a received pulse of light [18], [19]. These methods are cheaper than the laser scanning-based methods, and can be used in dynamic environments since they provide depth data at video rate [20], [21]. However, the depth information obtained from the ToF sensors may be distressed with noise, and also frequently contains outliers due to difficult illumination situations and the reflectivity of an object [22]. Despite recent advances in optical technology, these sensors cannot be applied directly due to their lower resolution and higher noise in comparison to general color cameras. Therefore, effective pre- or postprocessing and fusion techniques using other types of high-resolution sensors (e.g., color cameras) are needed to yield high-quality depth data.

In this paper, we propose a novel up-sampling method that enhances multiview depth maps provided from multiple ToF

Manuscript received December 13, 2012; revised April 27, 2013, and June 26, 2013; accepted August 5, 2013. Date of publication August 15, 2013; date of current version April 2, 2014. This work was supported by the Ministry of Knowledge Economy, Korea, under the Information Technology Research Center Support Program under Grant NIPA-2013-H0301-13-1008 supervised by the National IT Industry Promotion Agency. The work of D. Min was supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology, and Research. This paper was recommended by Associate Editor J. Cai.

J. Choi and K. Sohn are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea (e-mail: cjw0512@yonsei.ac.kr; khsohn@yonsei.ac.kr).

D. Min is with the Advanced Digital Sciences Center, 138632, Singapore (e-mail: dongbo@adsc.com.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2013.2278160

sensors by incorporating multiple color cameras. The color cameras are employed to overcome the physical limits of the ToF sensors. Choi *et al.* [23], [24] enhanced the depth video in both spatial and temporal aspects by combining a ToF sensor with a single color camera, and demonstrated that it is very useful in the 3-D TV system based on single-view color-plus-depth video. These algorithms, however, were designed to obtain a single-view depth map only and thus, are insufficient in an auto-stereoscopic 3-D display, which requires the multiview color-plus-depth video. In contrast, our approach aims at enhancing the quality of multiview depth maps simultaneously. To achieve this goal, we consider both the interview coherence and the reliability of multiple warped depth maps in the hybrid 3-D video acquisition system consisting of multiple ToF sensors and color cameras. For each single color view, the initial depth maps warped from multiple ToF sensors are adaptively fused based on their reliability. The interview coherence is then explicitly enforced between up-sampled depth maps corresponding to all the color views. We will show that the proposed method provides an excellent solution for the auto-stereoscopic 3-D TV system based on multiview color-plus-depth representation.

This paper offers two major contributions. First, we propose the depth reliability metric to improve the depth accuracy in case of using multiple ToF sensors. The input low-resolution depth maps obtained from multiple ToF sensors are warped into each color view and then adaptively combined by using their reliability measures, which are defined by using a depth validity and a relative position of depth sensors. We will show that such adaptive fusion approach effectively addresses outliers and mismatches, which may exist on the multiple input depth maps. It can also suppress misalignment artifacts between color image and corresponding depth map. Second, our approach, inspired by the recent success of joint bilateral up-sampling (JBU) for depth up-sampling [23], describes how to formulate the interview coherence in the joint filtering framework. The interview consistency is enforced by adaptively considering all color views in the up-sampling process. It also reduces bit rates required to compress the multiview (up-sampled) depth maps. Note that our approach can be easily combined with other filtering-based up-sampling methods, such as the weighted mode filtering (WMF) [25], and extended into the temporal domain by using temporal filtering methods as in [26].

There are several methods that obtain accurate depth maps by combining ToF sensors with stereo matching algorithms. The stereo matching algorithms help resolve the weakness of the ToF sensor to some extent, but the performance improvement would be relatively marginal, compared to its huge computational overhead from the stereo matching method. Therefore, we focus primarily on processing multiview depth data obtained from multiple ToF sensors with no stereo matching algorithm. Experimental results show that the proposed method produces comparable results, yet much more efficiently, to these fusion approaches utilizing the stereo matching algorithms together.

The remainder of this paper is organized as follows. We introduce the background and motivation about the sensor fusion in Section II. We then describe the proposed multiple sensor setup, and present the proposed framework that generates high-quality multiview color-plus-depth video in Sections III and IV, respectively. Finally, experimental results and conclusion are shown in Sections V and VI, respectively.

II. Related Work and Motivation

Current trends in 3-D video technology have moved from stereo-based to multiview-based video. Auto-stereoscopic 3-D video applications benefit from the multiview video capturing system, but such systems consisting of only color cameras may require a large number of cameras and huge amount of bandwidth. In order to address these problems, the European 3-D consortium, called 3D4YOU [27], has introduced various 3-D video formats, such as the color-plus-depth [2], [3] and the layered depth image [1], and developed a practical multiview color and depth capturing system consisting of five color cameras and two ToF sensors for obtaining high-quality 3-D video [28]. For wider field of view (FoV), which may be important in the auto-stereoscopic 3-D system, the multiple ToF sensors were used in the capturing system. Given low-quality depth maps obtained from the depth sensors at video rate, they proposed to utilize high-quality color images for enhancing the depth maps, and to combine them with depth results from a stereo matching method.

As the number of ToF sensors increases, disagreement between depth values obtained from each sensor may occur. Depth maps obtained from ToF sensors often suffer from the presence of missing or incorrect regions due to poor reflectivity, illumination, and optical limitation. It is also important to consider the misalignment, which may occur in the registration of depth maps and color images. The proposed method addresses these issues by utilizing the reliability metric defined with the depth accuracy and the relative position of color cameras and depth sensors.

Generally, fusion approaches of depth maps and color images assume that there exists a joint occurrence between depth discontinuities and image edges. Homogeneous regions in the color images are assumed to contain smoothly varying geometry. In order to overcome the physical limit of the ToF sensor, Kopf *et al.* [29] presented a JBU based on bilateral filter by using the color information as a prior. Yang *et al.* [30] proposed an iterative JBU by building a 3-D cost volume based on the current disparity value and then applying the bilateral filtering for each slice. A final depth is selected using the winner-takes-all (WTA) method. The iterative bilateral filtering on the cost domain results in better edge-preserving performance, but is too computationally intensive. Yang *et al.* [31] also proposed the hierarchical depth up-sampling method for efficient depth enhancement; however, the complexity is still high and the performance depends on the number of discrete (quantized) depth search ranges. Chan *et al.* [32] proposed a noise aware filter that adaptively blends standard up-sampling and JBU for each pixel, depending on local characteristics of depth maps. Another method based on nonlocal mean filters was proposed by considering both intrapatch similarity and color information [33]. Min *et al.* [25] proposed a weighted

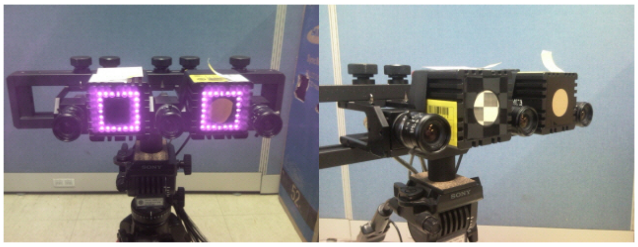


Fig. 1. Multiview color-plus-depth acquisition system: a hybrid structure with two ToF depth sensors and three color cameras.

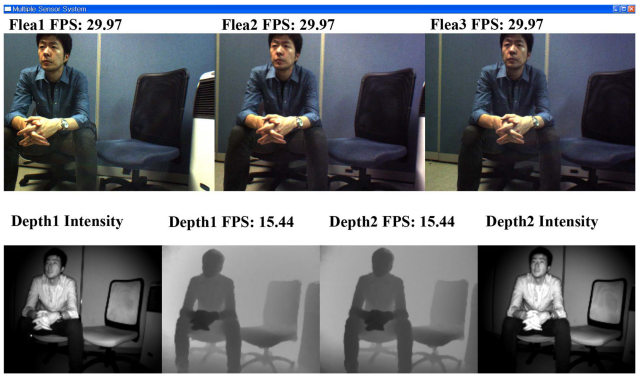


Fig. 2. Example of multiview color images and depth maps obtained from the system of Fig. 1. The color and depth images were resized for better visualization. The actual size is 1024×768 for the color image and 176×144 for the depth map, respectively. The intensity image of each depth sensor is not used in the up-sampling process.

mode filtering based on a joint histogram computed using the low resolution depth map and high resolution color image. The weight based on similarity measure between reference and neighborhood pixels is used to construct the histogram, and a final solution is then determined by seeking a global mode on the histogram.

Diebel and Thrun [34] proposed a Markov random field (MRF)-based depth up-sampling method by fusing a low resolution depth map and its high resolution color image. Park *et al.* [35] further improved the MRF-based depth up-sampling method by incorporating a nonlocal weighting term in the MRF formulation and including an additional edge weighting scheme to reinforce the preservation of fine texture.

Zhu *et al.* [36], [37] proposed a fusion algorithm based on the ToF sensor and stereo camera. Data term is derived by adaptively combining two cost functions calculated from both ToF sensor and stereo matching method. A final energy function, defined by an MRF model with a smoothness constraint, is solved by the loopy belief propagation (LBP) [14]. This method was extended into temporal fusion for generating temporally consistent high-quality depth video [38]. However, the global optimization used in these methods increases the computational cost significantly.

In contrast to these fusion approaches using both the ToF sensor and the stereo matching algorithms, our goal is to generate high-quality multiview depth maps by using only raw depth data from ToF sensors without a stereo matching algorithm. To enforce the interview consistency, we consider all color views, instead of up-sampling single-view depth map independently. Interview coherence is a crucial factor in the up-sampling process, since it seriously affects the compression ratio as well as the quality of multiview color-plus-depth video. In addition, we will show that comparable results to the fusion techniques using stereo matching methods together can be achieved with a much lower computational complexity.

III. Multisensor Setup

A. System Configuration

We built the acquisition system with three Point Grey Flead color camera [39] and two MESA Imaging SR4000 ToF sensors [15], as shown in Fig. 1. Note that this system can be directly extendable with arbitrary number of color cameras and ToF sensors. FoV can be adjusted using the baseline between the cameras. Using more color cameras and depth sensors enables a wider range of scene to be covered, which is essential to producing high-quality auto-stereoscopic 3-D video.

The resolution of the color camera is 1024×768, and the frame-rate is 30 frames/s. The resolution of the ToF sensor is 176×144, and the frame-rate is approximately 30 frames/s. In this paper, the depth frame-rate was adjusted to 15 frames/s, since a longer integration time tends to yield a more accurate depth map by accumulating multiple data to reduce a noise. All the cameras are connected with a single computer and synchronized using a trigger signal. The baseline distance between ToF depth sensor and color camera is approximately 70mm. In order to prevent an interference between multiple ToF sensors, their modulation frequency was set to 29 MHZ and 30 MHZ, respectively. Temporally synchronized images, intensity (amplitude) images, and corresponding depth maps are obtained, as shown in Fig. 2. Since the data captured from the multiple cameras have different viewpoints, the low-resolution depth maps should be warped into the color camera coordinates. Note that the intensity image of each depth sensor is used only in a camera calibration, and not in the up-sampling process.

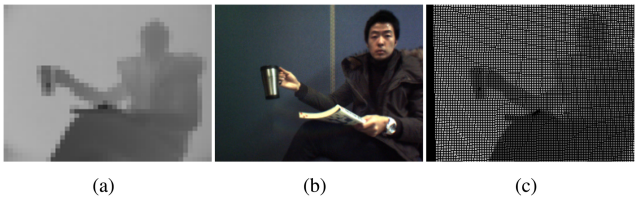
B. Depth Map Registration

To warp input depth maps into the color cameras, we first calculate projection matrices $P_c = K_c[R_c|t_c]$ and $P_d = K_d[R_d|t_d]$ for multiple depth sensors and color cameras using the Bouguet's camera calibration toolbox [40]. K, R, and t represent intrinsic, rotation, and translation matrices, respectively. Each point of the input depth map is first warped to 3-D world coordinates using the projection matrix of the depth sensor as follows:

$$P_w = R_d^{-1} {K_d}^{-1} p_d Z_d + t_d \tag{1}$$

where $P_w = (X_d, Y_d, Z_d, 1)^{\mathrm{T}}$ and $p_d = (x_d, y_d, 1)^{\mathrm{T}}$ are points in the 3-D world coordinate and the image coordinate of depth image, respectively. Then, P_w is back-projected onto the color image coordinate using the projection matrix P_c of the color camera as follows:

$$m_c = P_c \cdot \mathrm{X}_d = K_c R_c (P_w - t_c) \tag{2}$$



(a) (b) (c)

Fig. 3. 3-D warped depth map. (a) Low-resolution depth map. (b) High-resolution color image. (c) High-resolution depth map matched to the color image coordinate with holes.

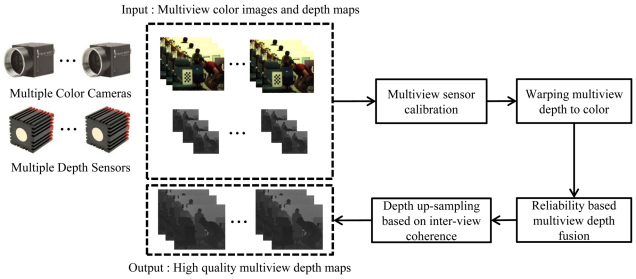


Fig. 4. Overall framework of the proposed method. Multiview sensor calibration is first performed to correct a lens distortion and warp initial multiview depth maps into multiview color images. Our contributions consist of: 1) reliability-based multiview depth fusion and 2) depth up-sampling based on interview coherence.

where $m_c = (m_1, m_2, m_3)$ is a 2-D homogeneous point in the color image coordinate. The final 2-D point p_c is computed as $(m_1/m_3, m_2/m_3)$. The depth map warped from each depth sensor is used as an initial depth map for the up-sampling process.

Fig. 3 shows an example of warping a depth map into the color image coordinate. Note that due to the resolution difference between two sensors, many holes occur in the warped depth map as shown in Fig. 3(c).

IV. Proposed Framework

Fig. 4 presents an overall framework of the proposed method. Multiview sensor calibration is applied to correct a lens distortion and warp multiview depth data into each color camera coordinate. This important warping process influences the reliability-based multiview depth fusion and the interview coherence of up-sampled multiview depth maps, which will be explained later. The main algorithms are divided into two steps: multiview depth fusion and depth up-sampling with interview coherence. Our experimental setup consists of two ToF depth sensors (D_1 and D_2) and three color cameras (C_1, C_2, and C_3). Thus, the depth maps from two ToF depth sensors are warped into each color image coordinate (three color views) and adaptively combined based on reliability as shown in Fig. 5. The initial fused depth maps for the color images are then simultaneously up-sampled by taking into account the interview coherence between the color images. For the clarity sake, we denote a method using the reliability-based fusion, the interview coherence, and the combination of the two as R-method, M-method, and RM-method, respectively.

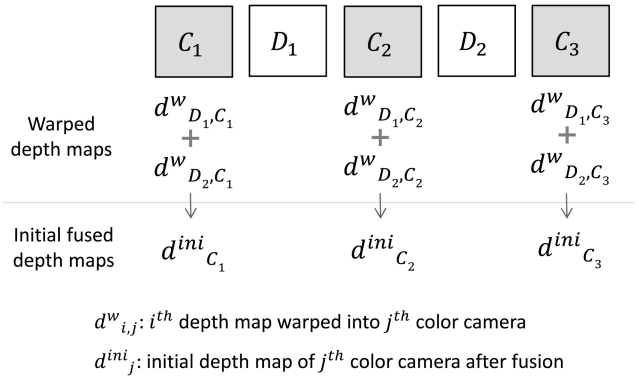
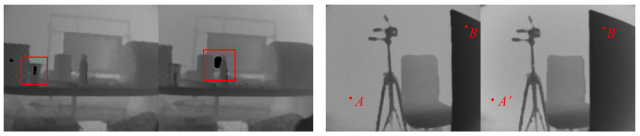


Fig. 5. Initial depth fusion process. C_i and D_j represent ith color camera and jth depth sensor, respectively. We obtain the multiple warped depth maps of the number of depth sensors (here, two) for each color camera. The warped depth maps into each color view are fused for generating the initial depth map used in the up-sampling process.



(a) (b)

Fig. 6. Problems of the depth maps obtained from multiple depth sensors. (a) Unreliable depth acquisition. (b) Different depth values at the same world point.

A. Reliability-Based Multiview Depth Fusion

Depth measurement values obtained from multiple depth sensors might be slightly different, leading to some perturbation in the warping process. Such difference between calibrated depth values may occur due to the calibration errors and the noise of input depth maps. To address these problems, we introduce two reliability metrics: geometric reliability and depth reliability. The geometric reliability metric is defined by considering baseline distance between color camera and depth sensor. The depth difference between depth sensors is utilized as a criterion of the depth reliability.

Fig. 6 shows two problems related to the depth maps obtained from multiple depth sensors. First, outliers may occur due to the characteristics of the depth sensor, which is error-prone to reflection as shown in the red box of Fig. 6(a). In addition, 2-D points in the color views projected from the same 3-D point may have different depth values, due to the perturbation of depth values. For instance, A, A' and B, B' should share the same point in world coordinate in Fig. 6(b). However, the calibrated depth values with respect to reference color view are slightly different from each other. The normalized depth values of A and B (from depth sensor D_1) with respect to color camera C_2 are 165 and 44, whereas those of A' and B' (from depth sensor D_2) are 167 and 48. These errors usually occurs in the calibration and depth acquisition stages.

Geometric reliability is defined by a baseline distance between depth sensor and color camera. As this distance gets larger, the geometric warping error on the color camera will linearly increase due to depth measurement errors. In addition,

the calibration error may lead to the misalignment in the depth registration. This is especially true when a 3-D point in world coordinate is warped away from a camera. Namely, as a depth sensor is moved away from the color camera, the translation components increase, causing more warping depth error.

For simplicity, let us assume that a camera setup is parallel $R_c, R_d = I$ in (1) and (2), and there are no calibration error and radial distortion. Then, the 2-D homogeneous point m_c in the color image coordinate can be expressed as

$$m_c = K_c({K_d}^{-1} p_d Z_d + t_d - t_c). \tag{3}$$

For further simplification, two intrinsic parameters are assumed to be identical ($K_c = K_d$). It would be possible since 2-D depth image can be resized so that its intrinsic parameter becomes equal to that of the color camera. In the parallel camera configuration, $t_d - t_c$ can be written as a 3-D vector $t_{dc} = (B, 0, 0)^{\mathrm{T}}$, where B represents a baseline distance between the color camera and the depth sensor. The $m_c = (m_1, m_2, m_3)$ is then written as $p_d Z_d + K_c t_{dc}$. Finally, 2-D point p_c in the color image coordinate can be derived as follows:

$$p_c = (m_1/m_3, m_2/m_3) = (x_d + fB/Z_d, y_d) \tag{4}$$

where f is a focal length of the color camera. This equation represents a simple 1-D parallel stereo camera setup containing 1-D horizontal disparity fB/Z_d only. It indicates that the warping error ∇p_c increases linearly proportional to the baseline distance between the color camera and the depth sensor, when there exists the depth measurement ∇Z_d. In addition, the calibration error $\nabla t_{dc} = (\nabla B, 0, 0)$ leads to the warping error $\nabla p_c = (\nabla x_c, 0)$ in the color camera, where $\nabla x_c = f \nabla B / Z_d$.

In conclusion, the baseline distance between the depth sensor and the color camera affects the depth warping in the presence of the depth measurement error. For instance, when two depth sensors and three color cameras are used as shown in Fig. 5, warping the depth map of depth sensor D_2 to the color camera C_1 leads to worse warping results due to the errors of the depth measurement and the camera calibration.

Fig. 7 demonstrates the influence of camera calibration errors on depth registration accuracy. We analyzed the alignment results by overlaying up-sampled depth maps on each color image. Fig. 7(b) and (c), where the warping distance is larger than that of Fig. 7(d) and (e), shows much worse alignment results on the depth boundaries. Please refer to the regions pointed out by arrows in the figure. The depth registration error at the foot of bear is also worse than that of the pattered box or background. Based on these observations, we propose the reliability metric considering the baseline distance as follows:

$$R_{i,j} = \exp\left(-\frac{||c_i - c_j||}{\sigma}\right) \tag{5}$$
$$(i = D_1, ..., D_N, j = C_1, ..., C_M)$$

where N and M represent the number of depth sensors and color cameras, respectively. $R_{i,j}$ is the geometric reliability function between depth sensor i and color camera j. c_i and c_j are center points of depth sensor i and color camera j, and $||c_i - c_j||$ represents the baseline distance between depth sensor i and color camera j. In other words, the geometric reliability $R_{i,j}$ is inversely proportional to the baseline distance. σ represents the control parameter of the reliability $R_{i,j}$.

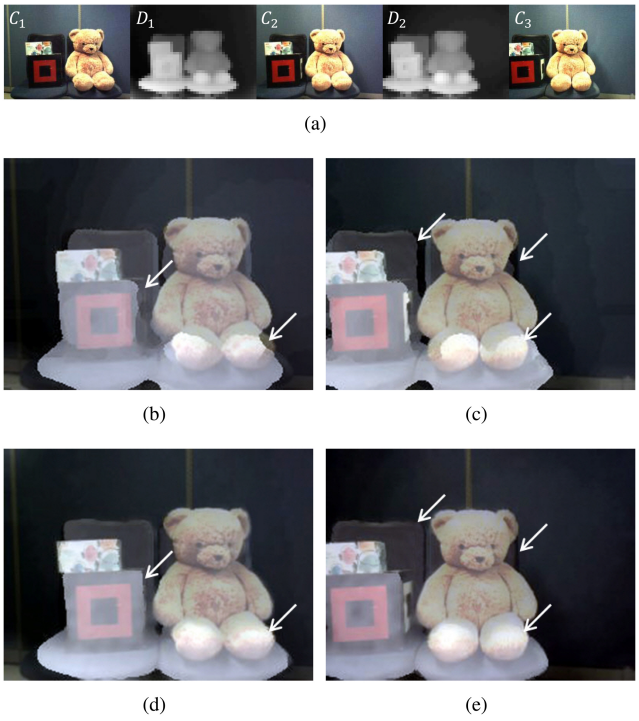
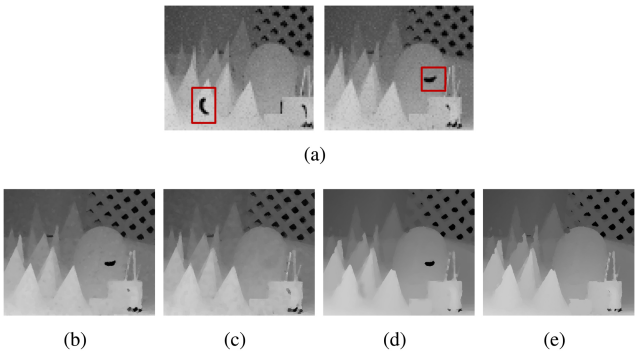


Fig. 7. Influence of camera calibration and depth measurement errors on depth registration accuracy. (a) Original color image from color camera C_1-C_3 and low-resolution depth map from depth sensor D_1-D_2. (b) Up-sampled depth maps from depth sensor D_2 to color camera C_1. (c) Up-sampled depth maps from depth sensor D_1 to color camera C_3. (d) and (e) Up-sampled depth maps to color camera C_1 and C_3 using reliability-based multiview depth fusion. We could observe that better aligned depth maps are obtained around the object boundaries. Please refer to the regions indicated by white arrows. Note that the up-sampled depth maps were obtained by using (b) and (c) the conventional JBU method and (d) and (e) the R-JBU method.

Depth values warped from multiple depth sensors are used to define the depth reliability. Finally, a depth fusion method using both the geometric and depth reliability can be defined as follows:

$$\begin{cases} d_j^{ini}(p) = \sum_{i=D_1}^{D_N} R_{i,j} d_{i,j}^w(p) / \sum_{i=D_1}^{D_N} R_{i,j}, \\ \qquad if \sum_{i=D_1}^{D_{N-1}} \sum_{k=i+1}^{D_N} |d_{i,j}^w(p) - d_{k,j}^w(p)| < \frac{N(N-1)}{2} d_{TH_1} \\ d_j^{ini}(p) = null, \\ \qquad otherwise \end{cases} \tag{6}$$

where $d_{i,j}^w(p)$ represents the depth map of the color image j warped from depth sensor i. $d_j^{ini}(p)$ is an initial depth map on the high-resolution corresponding to color camera j $(C_1, \cdots, C_M)$. This is calculated from a weighted sum of $d_{i,j}^w(p)$ using the geometric reliability $R_{i,j}$ when the condition of depth difference meets. Note that d_{TH_1} is a threshold value in the case of using only two depth sensors, i.e., N=2, and is determined empirically. As shown in the condition of (6), the threshold determining the depth difference increases according to the number of depth sensors N. Namely, when more than two depth sensors are used, the total number of cases for



(a) (b) (c) (d) (e)

Fig. 8. Up-sampled depth maps by various methods for verifying the performance of the R-method (up-sampling ratio = 4). (a) Input depth map from Middlebury Cone image. (b) Original WMF. (c) Original WMF + outlier rejection. (d) R-WMF without outlier rejection. (e) R-WMF.

TABLE I

PERFORMANCE ANALYSIS OF PROPOSED RELIABILITY METRIC ON THE DEPTH UP-SAMPLING (O.R.: OUTLIER REJECTION)

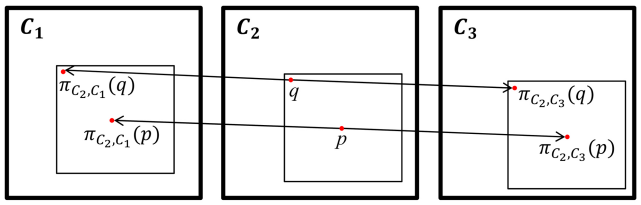
Up-samp. Ratio	Algorithm	*Teddy*		*Cone*	
		all	disc	all	disc
4×	WMF [25]	27.1	39.1	26.1	34.8
	WMF with O.R.	24.3	37.8	23.4	33.1
	R-WMF without O.R.	10.8	28.8	12.5	23.4
	R-WMF	8.75	24.4	8.65	19.2

measuring the depth difference should be considered (here, $N(N-1)/2$). When the total distortion (difference) between multiple depth measurements is larger than the threshold, they are not used in the next up-sampling process. Fig. 7 shows that the proposed reliability-based multiview depth fusion can considerably reduce the misalignment in the depth up-sampling process. Note that the up-sampled depth maps were obtained by using Fig. 7(b) and (c) the conventional JBU method, and Fig. 7(d) and (e) the R-JBU method.

In Fig. 8, the proposed reliability metric of (6) is analyzed in more details. The outlier rejection and reliability-based fusion of (6) are verified using the WMF method. The outlier rejection condition plays an important role in successfully removing dense error regions inside the red box of Fig. 8. Furthermore, we also found that the depth quality is even further improved by using the proposed reliability metric as shown in Fig. 8(e). In conclusion, the depth accuracy of the original R-WMF method is the best, meaning that the outlier rejection and the proposed reliability metric are complementary to each other. Table I shows the objective performance evaluation for depth maps up-sampled by various methods. We can find that the R-WMF method always outperforms other methods. More detailed analysis will be provided in the experimental section.

B. Depth Map Up-Sampling Based on Interview Coherence

After applying depth registration and multiple depth fusion for all color views, we up-sample each depth map considering an interview coherence. It is important to maintain the interview coherence of up-sampled depth maps, as it may directly influence the quality of 3-D video. In order to improve the resolution of the ToF sensor, a number of JBU-based methods



$\pi_{i,j}(p)$: window of i^{th} camera warped into j^{th} camera

Fig. 9. Warped window concept used in the proposed framework.

have been proposed to up-sample the depth map accurately by using an associated color image. In general, the JBU [29] is expressed as follows:

$$\tilde{d}(p) = \frac{1}{k(p)} \sum_{q \in \Omega} d(p) f(p-q) g(I(p) - I(q)). \tag{7}$$

Given a color image $I(p)$ and a depth map $d(p)$, we can obtain a solution $\tilde{d}(p)$ by using spatial f and range g kernels centered at the pixel p. Here, two kernels are defined as the Gaussian functions with standard deviation σ_s and σ_I, respectively. Ω is the spatial support of the kernel f, and $k(p)$ represents a normalization factor.

1) *Multiview-JBU:* Directly applying this up-sampling method to multiview depth maps may cause an interview inconsistency due to its view-independent processing. The local characteristics of the multiview color images (e.g., edge and color) may vary slightly according to several factors such as noise, lighting conditions, and reflectivity, which often lead to serious visual artifacts in the up-sampled depth maps. To address this problem, we propose a novel algorithm, called multiview-JBU (M-JBU), which produces multiple depth maps that are consistent between neighboring views. Specifically, the interview information as well as the spatially-neighboring color information inside each view are considered in the range kernel as follows:

$$\begin{aligned} \tilde{d}_j^{(t+1)}(p) &= \frac{1}{k(p)} \sum_{k=C_1}^{C_M} \sum_{q \in \Omega} {d_j}^{(t)}(p) f(p-q) \\ &\times w_k(p) \cdot g_k(I_k(\pi_{j,k}(p)) - I_k(\pi_{j,k}(q))) \end{aligned} \tag{8}$$

$${d_j}^{(t+1)}(p) = \begin{cases} \tilde{d}_j^{(t+1)}(p) & if\ |\tilde{d}_j^{(t+1)}(p) - {d_j}^{(t)}(p)| \le d_{TH_2} \\ {d_j}^{(t)}(p) & otherwise \end{cases}$$

where $d_j^{(0)}(p) = d_j^{ini}(p)$, and w_k $(k = C_1, ..., C_M)$ represents the weight of the range filter corresponding to each view.

The range filter kernel g is replaced with a set of g_k $(k = C_1 \cdots C_M)$ consisting of partial range filter kernels corresponding to all the color views. As explained in Fig. 9, $\pi_{j,k}(p)$ represent a pixel warped onto the color camera k from the reference color camera j. A final solution d_j represents the up-sampled depth map corresponding to color view j. Note that the proposed method iteratively reduces the interview inconsistency on the up-sampled depth maps. A texture-copying problem [32] usually occurs in the regions where neighboring pixels inside an object have different color values, leading to wrong results on the up-sampled (filtered) depth map by deforming initial depth values. This problem can

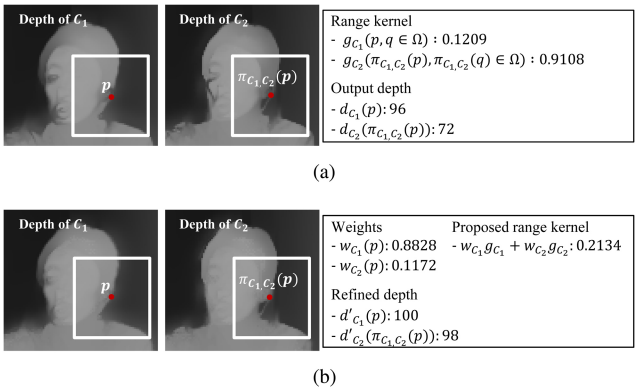


Fig. 10. Refinement using the range kernel and weights. $g_{C_1}(p, q \in \Omega)$ and $g_{C_2}(\pi_{C_1,C_2}(p), \pi_{C_1,C_2}(q) \in \Omega)$ represent real range filter kernel values calculated in each color view (C_1 and C_2). $w_{C_1}(p)$ and $w_{C_2}(p)$ represent real weights calculated according to g_{C_1} and g_{C_2} in M-JBU. Two depth values d_{C_1} and d_{C_2} independently filtered using each range kernel shows very different results, though they are from the same 3-D point. In contrast, our method effectively handles the view inconsistency on the up-sampled depth maps (d'_{C_1} and d'_{C_2}) by using the new range kernel $w_{C_1}g_{C_1} + w_{C_2}g_{C_2}$. (a) Partial results of Fig. 11(b) (conventional JBU). (b) Partial results of Fig. 11(c) (M-JBU).

also be addressed by considering the difference between two consecutive results $d^{(t+1)}$ and $d^{(t)}$ in the iteration. When the difference is larger than a threshold value, new depth value $d^{(t+1)}$ is decided as outlier distorted by the texture-copying artifact, so are not used to update the depth result. It is because the initial depth maps (d^{ini}) do not suffer from the texture-copying problem, though corrupted by several outliers. A threshold value d_{TH_2} is selected empirically.

Now, we will explain how the weighting function $w_k(p)$ is derived. The underlying assumption to define the weight function $w_k(p)$ is that the interview inconsistency can be measured by using the difference between the partial range filter kernel values on the color images. Ideally, all the corresponding pixels from different color views should contain the same color value. However, in practice, the color inconsistency between multiple views often produces the different range kernel functions in (7), resulting in inconsistent results in the single-view depth up-sampling. In order to handle this problem, we take into account all color view together to define the new range kernel function in (8). To combine a set of kernel functions adaptively, the weight $w_k(p)$ for a pixel p is defined using a ratio of partial range filter kernel values of the kth view as follows:

$$w_k(p) = \frac{\sum_{l \in \Psi \setminus k} \sum_{q \in \Omega} g_l(I_l(\pi_{k,l}(p)) - I_l(\pi_{k,l}(q)))}{\sum_{l \in \Psi} \sum_{q \in \Omega} g_l(I_l(\pi_{k,l}(p)) - I_l(\pi_{k,l}(q)))} \qquad (9)$$

$$(\Psi = \{C_i | i = 1 \cdots N\}).$$

All the color views are considered simultaneously by utilizing the proportion of weights. When all the weights are similar to each other, it becomes similar to the result of conventional JBU. Otherwise, the partial range filter kernel values of each view have a different effect on filtering of the corresponding points. For example, when the partial range filter kernel for one view is large enough to smoothen the image, whereas the

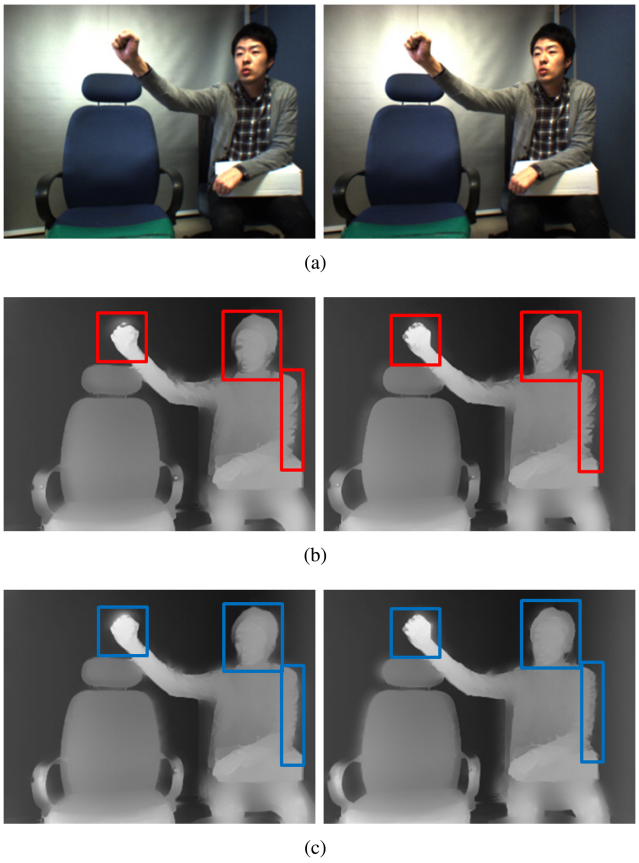
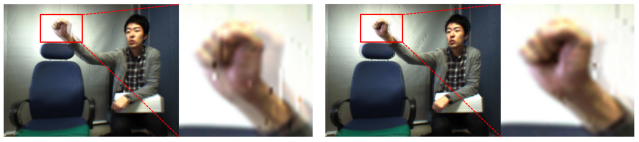


Fig. 11. Example of results with or without interview coherence. (a) Original image. (b) Depth maps up-sampled by the conventional JBU. (c) Depth maps up-sampled by the M-JBU.

corresponding range filter kernel for other view is relatively small so preserves edges, the proposed weights make all the range filter kernels preserve edges. This can help reduce the inconsistency between up-sampled multiview depth maps, particularly around the depth discontinuities.

In Fig. 10, we also analyzed the role of weights in the M-JBU. In this example, we use only one depth sensor and two color cameras, i.e., N=1 and M=2, in order to show the effect of the interview coherence more easily. In the single-view-based JBU, the range filter kernel values $g_{C_1}(p, q \in \Omega)$ and $g_{C_2}(\pi_{C_1,C_2}(p), \pi_{C_1,C_2}(q) \in \Omega)$ for each color camera C_1 and C_2 are very different, leading to incoherent filtered depth outputs $d_{C_1}(p)$ and $d_{C_2}(\pi_{C_1,C_2}(p))$ between the corresponding points. This problem is effectively handled by considering two weight functions $w_{C_1}(p)$ and $w_{C_2}(p)$ according to the partial range filter kernel values. The weight functions are decided by using the range filter kernel values calculated from each color view C_1 and C_2 as explained in (9). Therefore, in the M-JBU method, the depth values on two views become more consistent by a new range filter kernel value, $w_{C_1}g_{C_1} + w_{C_2}g_{C_2}$. In conclusion, the interview coherence is achieved by using variable weights according to the partial range filter kernel values used in the M-JBU.

Fig. 11 shows the effect of enforcing the interview coherence. We found that the result of the M-JBU is more consistent than that of the conventional JBU, especially in face, arm and



(a) (b)

Fig. 12. Results of synthesized view by Fig. 11(b) and (c). (a) Conventional JBU. (b) M-JBU.

fist. Fig. 12 demonstrates how the interview coherence affects the performance of virtual view synthesis.

2) *Extension to M-WMF:* The proposed up-sampling framework can also be applied to the WMF [25], which proposes to seek a global mode on the histogram by leveraging the similarity measure between the data of two pixels. When the histogram $H_G(p, d)$ is generated for each pixel p and its disparity variable d, the data (depth) of pixels inside a window centered at p is adaptively counted on its corresponding bin d by using the data similarity between reference and neighboring pixels as

$$H_G(p, d) = \sum_{q \in \Omega} f(p - q) g(I(p) - I(q)) G_r(d - d(q)) \quad (10)$$

where $G_r(x)$ is defined as the Gaussian function with σ_r. $I(p)$ and $d(p)$ represent the color image and the input low-resolution depth map. The final solution $d_G(p)$ for the WMF can be computed as follows:

$$d_G(p) = \arg\max_d H_G(p, d). \quad (11)$$

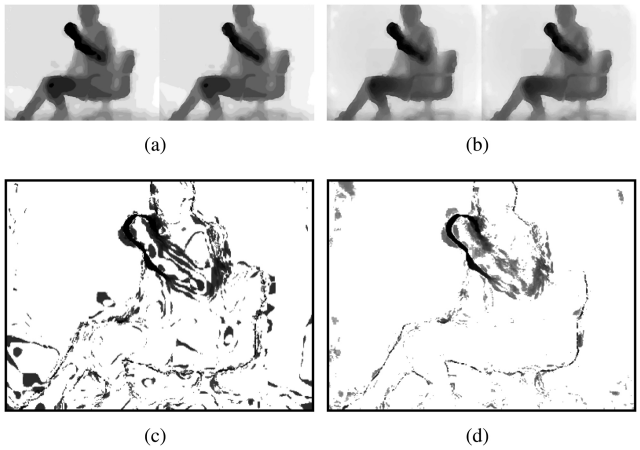
Similarly, the WMF is extended into the multiview depth up-sampling method by taking into account the interview coherence as follows:

$$\begin{aligned} H_{G,j}(p, d) = &\sum_k \sum_{q \in \Omega} f(p - q) G_r(d - d(q)) \\ &\times w_k(p) g_k(I_k(\pi_{j,k}(p)) - I_k(\pi_{j,k}(q))). \end{aligned} \quad (12)$$

The weight function $w_k(p)$ is computed in the manner similar to the M-JBU. Note that the edge-preserving performance of the WMF was shown to be superior to that of the JBU, because the JBU provides a mean value through an adaptive summation, while the WMF selects an output value that has the largest histogram value. Please refer to [25] for more information. In the experiments, we will show that the WMF, which is the state-of-the art method in the depth up-sampling method, can also be improved in case of up-sampling the multiview depth maps.

3) *Extension to Temporal Domain:* In multiview depth up-sampling procedure, the temporal coherence of a depth video is also an important factor in producing high-quality multiview 3-D video. Both RM-JBU and RM-WMF methods can be easily extended into the temporal aspect using the 3-D filtering concept in [26] as follows:

$$\begin{aligned} \tilde{d}_j(p_t) = &\frac{1}{k(p_t)} \sum_{k=C_1}^{C_M} \sum_{m \in N(t)} \sum_{q_m \in \Omega(p_m)} d_j(p_t) f(||p_t - q_m||) \\ &\times w_k(p_t) \cdot g_k(||I_k(\pi_{j,k}(p_t)) - I_k(\pi_{j,k}(q_m))||) \end{aligned} \quad (13)$$



(a) (b) (c) (d)

Fig. 13. Example of results with or without temporal coherence. (a) Consecutive depth maps up-sampled by RM-WMF without temporal coherence. (b) With temporal coherence. (c) Difference image of (a). (d) Difference image of (b). For better visualization, the contrast of the results was enhanced. The temporally consistent results were obtained in (b). Note that there are moving objects (e.g., arm) in the depth video.

$$\begin{aligned} H_{G,j}(p_t, d) = &\sum_k \sum_{m \in N(t)} \sum_{q_m \in \Omega(p_m)} f(p_t - q_m) G_r(d - d(q_m)) \\ &\times w_k(p_t) g_k(I_k(\pi_{j,k}(p_t))) - I_k(\pi_{j,k}(q_m))) \end{aligned} \quad (14)$$

where t and $N(t)$ represent the reference frame and the number of neighborhood frames, respectively. p_m represents a corresponding pixel of p_t in the mth frame, and q_m is a neighboring pixel of p_m. This correspondence can be obtained using various motion estimation methods such as the full-search block matching algorithm (FBMA) and the optical flow method. Here, the FBMA method was used for experiments. Equations (13) and (14) additionally use the neighbors of the temporally neighboring frames in the joint filtering algorithm. Using such temporal neighbors together in the filtering can reduce the temporal fluctuation effectively. Fig. 13 shows the effectiveness of the 3-D filtering. The difference images between consecutive frames were used to show the improvement on temporal aspect. We found that temporally consistent depth values were obtained in most parts of Fig. 13(d), except for moving objects such as arm, compared to Fig. 13(c). Please refer to an electric version for better visibility. In addition, in the context of video coding, the reduction of the temporal fluctuation can help save the bit rate. Please refer to [26] for more information.

V. Experimental Results

The proposed method was implemented using the Visual Studio 2010, with the exception of the depth image acquisition which used MATLAB, and was tested on an Intel Core i7 2.8 GHz processor with 4 GB RAM. Input images are multiple color images and depth maps obtained from our acquisition system, which consists of two depth sensors and three color cameras. In our experiments, the operation range of depth sensor set from 0.5m to 5.0m. Out of range data (depth) means invalid information and is excluded from depth fusion and up-sampling process. The proposed method is tested with the

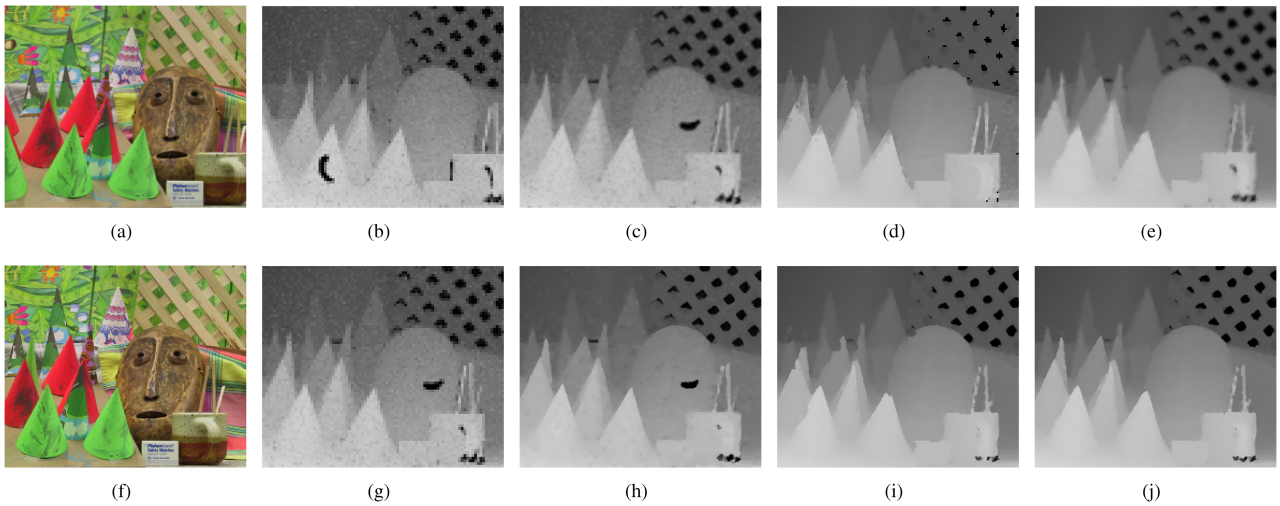


Fig. 14. Depth up-sampling results of cone images on color view C_6 with an up-sampling factor of four. (a), (b), (f), and (g) Input color images and noisy low-resolution depth maps (C_2, C_6 color cameras and D_2, D_6 depth sensors). (c) and (h) Up-sampled depth maps by conventional JBU and WMF. (d) and (i) Up-sampled depth maps considering the reliability-based fusion (R-JBU and R-WMF). (e) and (j) Up-sampled depth maps considering the reliability-based fusion and interview coherence (RM-JBU and RM-WMF).

TABLE II

OBJECTIVE PERFORMANCE EVALUATION FOR DEPTH MAPS UP-SAMPLED BY VARIOUS METHODS

Up-samp. Ratio	Algorithm	*Teddy*		*Cone*		*Venus*	
		all	disc	all	disc	all	disc
2×	JBU	31.7	39.7	31.9	38.9	9.77	14.4
	R-JBU	7.22	14.8	8.29	19.1	1.18	11.1
	RM-JBU	5.43	15.1	9.22	11.0	0.54	5.27
	WMF	8.92	16.3	8.91	16.5	1.06	7.65
	R-WMF	5.35	14.3	5.72	13.5	0.62	5.84
	RM-WMF	4.53	11.9	3.91	8.13	0.55	2.03
3×	JBU	30.2	41.2	32.4	41.9	7.43	13.9
	R-JBU	8.18	22.4	9.03	19.7	1.26	14.1
	RM-JBU	7.09	15.8	8.66	16.2	0.63	5.60
	WMF	17.1	33.5	18.6	29.0	1.22	8.30
	R-WMF	7.69	21.5	6.94	15.6	0.41	4.93
	RM-WMF	7.04	18.9	6.44	14.2	0.72	2.71
4×	JBU	33.6	44.3	34.2	45.1	7.68	14.3
	R-JBU	9.78	26.4	10.1	20.2	1.38	13.6
	RM-JBU	10.1	18.8	11.5	19.3	0.75	5.83
	WMF	27.1	39.1	26.1	34.8	1.81	8.56
	R-WMF	8.75	24.4	8.65	19.2	0.61	7.23
	RM-WMF	8.34	22.7	8.45	18.5	0.73	5.69

same parameters for all images. The control parameter σ in (5) is set to 2.8 by considering an actual distance between sensors (e.g., 6.0cm in our setup). Following the original WMF paper [25], the weighting parameters σ_I, σ_s and σ_r in (8) and (12) are set to 5, 7, and 12, respectively. The threshold values d_{TH_1} and d_{TH_2} in (6) and (8) are usually determined by the measurement range of the active depth sensor used in the experiments. In our experiments, it ranges from 0.5m to 5.0m, and the two threshold values are set to 0.1m and 0.3m. These values are converted into five and 15, considering that the range data is normalized from 0 to 255. We also empirically found that (8) converges within three iterations for all the experiments. The size of window Ω varies according to the image size. In our experiments, it is set to 7×7 and 13×13 for the Middlebury data sets and the real images, respectively.

A. *Objective Evaluation With Middlebury Datasets*

For quantitative evaluation, we first performed experiments using the Middlebury data sets. Two views (C_2 and C_6) of cone image with image size of 450 × 375 were used in this experiment since only two depth maps (D_2 and D_6) among all the color views ($C_1 \sim C_9$) were provided, i.e., $N, M = 2$. To verify the effectiveness of the reliability fusion, these input depth maps were obtained by down-sampling ground truth depth maps with factors of two, three, and four for each dimension, and then by adding several types of noise such as the additive white Gaussian noise (AWGN) where a standard deviation is 20, impulsive noise where a density is 0.03, and reflective noise generated by considering the depth sensor characteristics. Fig. 14(a), (b), (f), and (g) shows examples of the input color images and low-quality depth maps.

In Table II, we first performed an objective evaluation for the depth maps up-sampled by various methods. Note that the Tsukuba image was excluded from this evaluation, since only single-view depth map is provided. The depth accuracy was measured using the percent (%) of bad matching pixels (where the absolute disparity error is greater than 1 pixel) for all (all pixels in the image) and disc (the visible pixels near the occluded regions) regions [41]. We could find that the R-methods and RM-methods based on the JBU and WMF approaches always outperform the conventional methods. Interestingly, even the original WMF method [25], proven as the state-of-the-art methods in the depth up-sampling, showed only relatively marginal improvement in the presence of various types of noise and color distortion. In addition, as already explained in Fig. 10, the RM-methods outperform the R-methods around depth discontinuities (disc).

Fig. 14 shows the depth maps up-sampled by various methods. The results with the up-sampling factor of four only were shown due to the lack of space. Note that the warping-based registration was not used in this experiment, since two color views (C_2 and C_6) and corresponding depth

(a) (b) (c) (d) (e) (f)

Fig. 15. Synthesized view results of cone image using depth maps of Fig. 14. (a) Conventional JBU. (b) R-JBU. (c) RM-JBU. (d) Conventional WMF. (e) R-WMF. (f) RM-WMF. Please refer to an electronic version for better readability.

TABLE III
PERFORMANCE ON PSNR AND SSIM FOR SYNTHESIZED VIEW RESULTS BY VARIOUS METHODS

Algorithms	*Teddy*		*Cone*		*Venus*	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
JBU	24.94	0.7681	25.77	0.8506	26.55	0.8262
R-JBU	25.35	0.7970	26.28	0.8735	28.38	0.8779
RM-JBU	25.59	0.8126	26.37	0.8907	28.85	0.8997
WMF	25.08	0.7767	26.13	0.8625	28.64	0.8957
R-WMF	25.52	0.8027	26.56	0.8880	28.94	0.8995
RM-WMF	25.68	0.8142	26.81	0.8877	28.94	0.8998

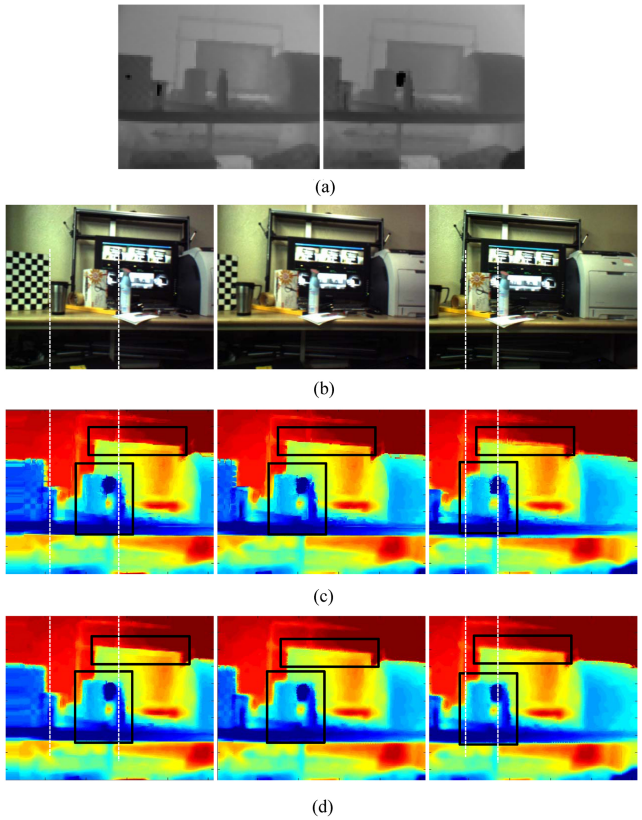
maps (D_2 and D_6) were already co-aligned. The depth maps of the R-JBU and R-WMF methods are superior to those of the conventional methods. Especially, our methods effectively suppress the various noises through the proposed fusion using the reliability metric, while the JBU and WMF methods still suffer from serious noise, as shown in Fig. 14(c) and (h).

Furthermore, the results obtained from the RM-JBU and RM-WMF methods showed that by using the interview coherence, one can better preserve edges and maintain the interview coherence for all color views (C_2 and C_6), even compared to the R-JBU and R-WMF methods.

Fig. 15 shows the synthesized view results by the depth maps of Fig. 14. We could find that R-methods and RM-methods outperform the conventional methods on the object boundaries. WMF-based results are also slightly better than JBU-based results, since the WMF can better sharply preserve the depth boundaries than the JBU as shown in Fig. 14. In Table III, we calculated the PSNR and SSIM for evaluating the objective performance. Synthesized view results were compared with corresponding original image (C_4).

B. Up-Sampled Depth Results on Real Data Sets

Next, we performed the depth up-sampling using the input data sets obtained from our acquisition system in Fig. 1, i.e., $N = 3$ and $M = 2$. The spatial resolution of the input color and depth image used in the experiments is 1024×768 and 176×144, respectively. Figs. 16 and 17 show the depth results up-sampled using the reliability-based depth fusion and the interview coherence. For better visualization, their contrast was adjusted with a color mapping. Especially, by comparing the results of Figs. 16 and 17, i.e., (M-method versus RM-method), we could find that the proposed reliability measure is very effective in fusing erroneous input depth maps. It was also shown in Fig. 16(d) that when only the depth sensor D_2 is used, the depth result up-sampled on the color camera C_3 is more aligned than that up-sampled on the color camera C_1,



(a) (b) (c) (d)

Fig. 16. Effect of interview coherence on the up-sampled depth maps. (a) Original depth maps obtained from the depth sensor D_1 and D_2. (b) Original color images obtained from the color camera C_1, C_2, and C_3. (c) Conventional JBU results using the depth sensor D_2 for each color view. (d) M-JBU results using the depth sensor D_2 for each color view: M-method maintains boundaries of objects more consistently than the conventional method, especially in the black boxes. However, up-sampled depth maps are still misaligned, especially with the color camera C_1 due to the baseline distance.

since the warping distance of C_3 is shorter (refer to Fig. 5). As shown in Fig. 16, the M-JBU method can maintain the shape of an object consistently in all views compared to conventional JBU. Both methods, however, still suffer from outliers. Such errors were effectively resolved in Fig. 17 using the reliability-based fusion. The depth results were aligned very well with the corresponding color image for all views, and reduced outliers as shown in Fig. 17(b) and (c). As expected, we could also confirm that the RM-WMF outperforms the RM-JBU in terms of an edge-preserving capability.

Fig. 18 shows 3-D visualization of those results in Fig. 16. The original low-resolution depth maps have many outliers

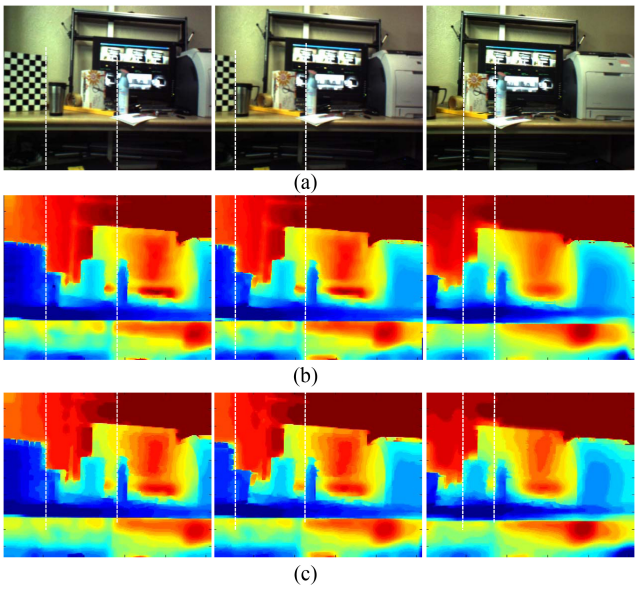


Fig. 17. Effect of the reliability-based fusion on the up-sampled depth maps. (a) Original color images obtained from the color camera C_1, C_2, and C_3. (b) RM-JBU results using the depth sensor D_1 and D_2. (c) RM-WMF results using the depth sensor D_1 and D_2: RM-methods are aligned with corresponding color images better than M-methods. In addition, outliers in the black boxes were removed and objects were represented more accurately than Fig. 16.

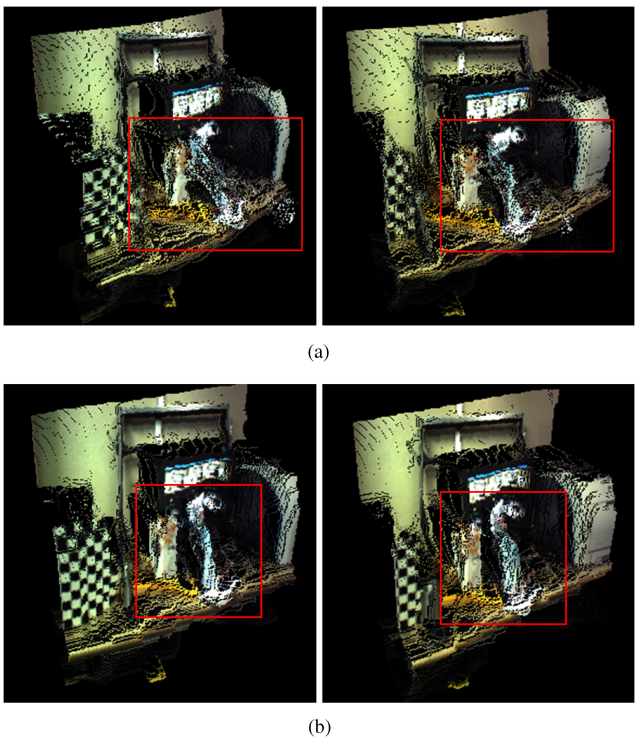


Fig. 18. 3-D visualization results of Figs. 16 and 17. Up-sampled depth maps corresponding to color camera C_1 and C_2 using (a) conventional method (3-D visualization results of the conventional JBU) and (b) RM-JBU (3-D visualization results of the RM-JBU). Our results outperform those of the conventional method. Please refer to the visualization results in the red boxes, and an electronic version for better readability.

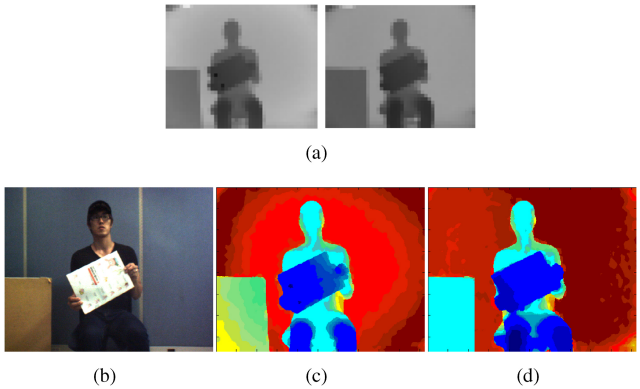


Fig. 19. Comparison of the RM-method with the conventional method. (a) Original depth maps from depth sensor D_1 and D_2. (b) Original color image from color camera C_3. (c) Up-sampled depth map by conventional WMF using only D_1. (d) Up-sampled depth map by RM-WMF using D_1 and D_2.

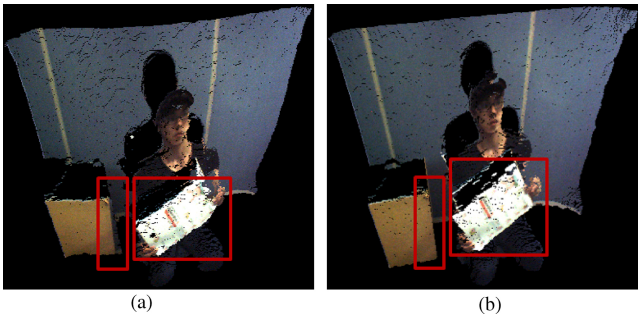


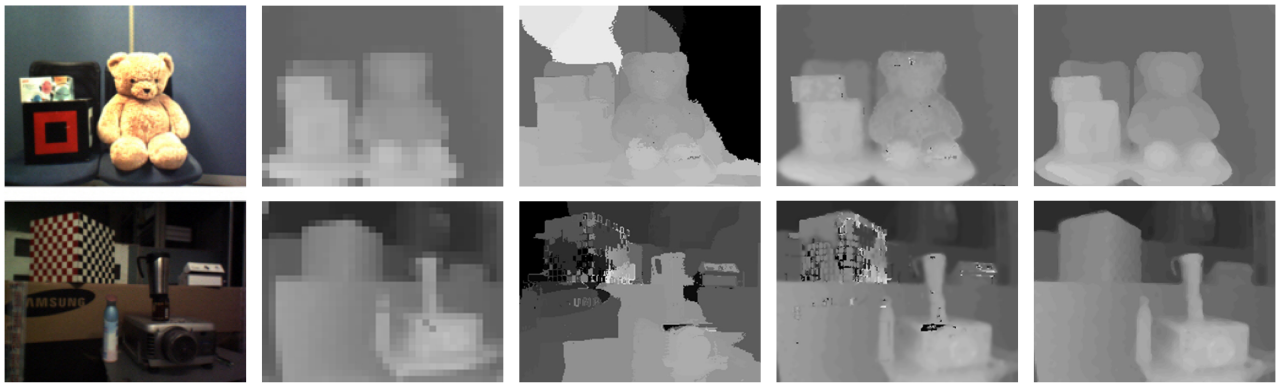
Fig. 20. 3-D visualization results of Fig. 19(c) and (d). Up-sampled depth maps corresponding to color camera C_3 using (a) conventional WMF (3-D visualization results of the conventional WMF) and (b) RM-WMF (3-D visualization results of the RM-WMF). Our result outperforms that of the conventional method. Please refer to the visualization results in the red boxes, and an electronic version for better readability.

at many regions including the monitor, check patterned box, and tumbler. We can verify that Fig. 18(b) is more interview consistent than Fig. 18(a) in the aspect of reconstructed shape and depth value for each view.

Figs. 19 and 20 also show the excellent performance of the proposed method. Fig. 19(a) represents original color image (C_3) and depth map (D_1 and D_2). The result obtained from the RM-WMF using D_1 and D_2 outperforms the conventional WMF using only D_1 as shown in Fig. 19(c) and (d). Fig. 20 shows 3-D visualization of the results in Fig. 19. We can find that the proposed method is aligned better than the conventional method, especially inside the red box. In addition, the depth measurement error in the book of D_1 is removed by the reliability metric. Additional video results were also provided in [42] to demonstrate the effectiveness of our approach more clearly.

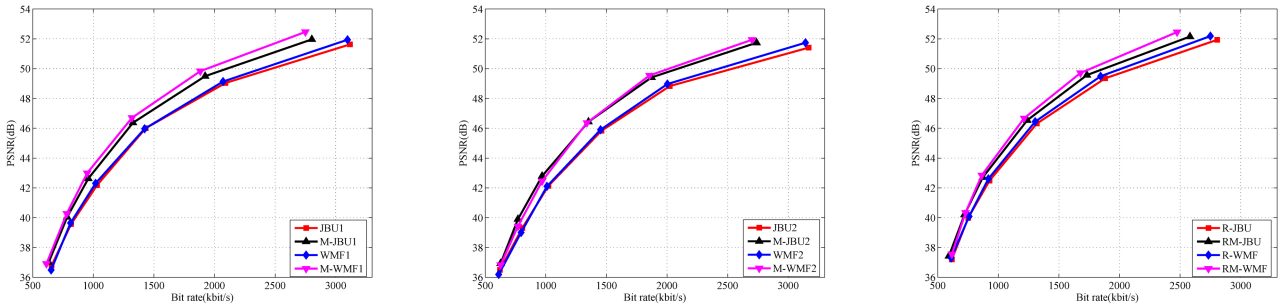
C. Comparison With Other Hybrid Methods

Fig. 21 compares the results with those of hybrid system [31], which consists of three views and one depth sensor. This system is similar to our approach, but utilizes the stereo matching algorithm together and adaptively combines two results (from stereo matching and depth up-sampling methods) for improving the depth quality. Note that the depth maps were



(a) (b) (c) (d) (e)

Fig. 21. Comparison of the proposed method with the hybrid system. (a) Original color image. (b) Initial depth map. (c) Stereo matching result based on [43] and [44]. (d) Depth map obtained by adaptive combination of the stereo matching and depth up-sampling results as in [31]. (e) Depth map up-sampled by the RM-WMF. Note that in this figure, the depth maps were converted to disparity maps using stereo camera parameters (e.g., baseline and focal length) since the stereo matching results are combined with the depth maps obtained from the depth sensor.



(a) (b) (c)

Fig. 22. Effect of the interview coherence on RD performance. (a) Total bit rates for three depth videos, corresponding to color views (C_1, C_2, and C_3), up-sampled using only depth sensor D_1 (method versus M-method). (b) Total bit rates of three depth videos up-sampled using only depth sensor D_2 (method versus M-method). (c) Total bit rates of three depth videos up-sampled using the reliability-based fusion (D_1 and D_2) (R-method versus RM-method). Enforcing the interview coherence on the up-sampling process leads to a significant bit rate saving in the depth video compression.

converted to disparity maps using stereo camera parameters (e.g., baseline and focal length), since the stereo matching results are combined with the depth maps obtained from the depth sensor.

We could find that our results are comparable to those of [31], even though the stereo matching is not utilized. In the experiments, we substituted the stereo matching algorithm of [31] with the method using the census transform [43] and the belief propagation [44] for better stereo results. Especially, the results of [31] preserve edges well, but severe depth errors may occur frequently at homogeneous and repetitive patterned regions as shown in Fig. 21(d). Such errors mainly come from the outliers of the stereo matching, which leads to serious artifacts in the final results despite the adaptive combination based on the reliability [31]. Moreover, it is difficult to suitably control the weights used to combine the results from the depth sensor and the stereo matching method in a practical environment.

The proposed method also has a computational advantage over the stereo matching-based hybrid system. For instance, our (unoptimzed) implementation takes about 12.94 s (for three views) for up-sampling the low-resolution depth map (176×144) to high-resolution (1024×768) on a single core CPU, while the stereo matching-based hybrid method takes about 97.29 s (for one view) due to the huge computational complexity of the stereo matching on high resolution stereo images with a large search range (e.g., 230 pixels). Most of the runtime (89.64 s) is from the disparity estimation stage. Note that our C implementation was not fully optimized, but such complexity analysis would be a good indicator of verifying the computational efficiency of our method.

Let us analyze the timing results from a computational perspective. For instance, the complexity of a local stereo matching can be defined as $O(SML)$, where S, M, and L represent an image size, a matching window size, and a search range, respectively, while that of the proposed depth up-sampling (filtering) method is $O(SM)$. In the local stereo matching methods, the nonlinear filtering should be applied repeatedly for all the disparity hypotheses, leading to a huge computational cost depending on the search range L. Although many methods have been developed to reduce the computational complexity in terms of M and L, these techniques solving a discrete labeling problem (defined on a discrete label space with a size of L) are still much slower than the relatively simple depth filtering approach such as our method. It should also be noted that the global stereo approach (e.g., using belief

TABLE IV
PERFORMANCE ON BIT RATES FOR DEPTH VIDEOS UP-SAMPLED WITH OR WITHOUT INTERVIEW COHERENCE

Algorithms	ΔBit rate(%)					
	QP=24	QP=29	QP=34	QP=39	QP=42	QP=45
JBU1 v.s. M-JBU1	-10.12	-8.15	-7.31	-6.94	-4.08	-3.65
JBU2 v.s. M-JBU2	-13.36	-7.48	-7.36	-5.11	-4.51	-0.99
WMF1 v.s. M-WMF1	-11.21	-9.08	-7.13	-6.47	-5.94	-3.85
WMF2 v.s. M-WMF2	-14.02	-7.35	-7.76	-3.98	-3.32	-1.62
R-JBU v.s. RM-JBU	-8.99	-7.13	-6.85	-5.49	-4.28	-3.22
R-WMF v.s. RM-WMF	-10.22	-8.98	-7.12	-4.99	-2.49	-0.97

TABLE V
PERFORMANCE ON PSNR FOR DEPTH VIDEOS UP-SAMPLED WITH OR WITHOUT INTERVIEW COHERENCE

Algorithms	ΔPSNR(dB)					
	QP=24	QP=29	QP=34	QP=39	QP=42	QP=45
JBU1 v.s. M-JBU1	0.34	0.45	0.37	0.44	0.46	0.23
JBU2 v.s. M-JBU2	0.33	0.58	0.60	0.65	0.58	0.43
WMF1 v.s. M-WMF1	0.52	0.68	0.73	0.69	0.63	0.41
WMF2 v.s. M-WMF2	0.19	0.57	0.46	0.35	0.41	0.63
R-JBU v.s. RM-JBU	0.22	0.35	0.24	0.21	0.18	0.20
R-WMF v.s. RM-WMF	0.31	0.24	0.32	0.23	0.26	0.28

propagation), which was used in our experiment, is generally much slower than the local stereo approach.

D. Improvement in Depth Video Coding

In many 3-D video applications (e.g., 3-D TV and freeview TV), which often require transmitting the depth video over network, the depth video should be compressed in an efficient manner. In this section, we will show that the proposed method also improves compression performance of the depth video by maintaining the interview coherence. Experimental results were obtained with various quantization parameters (QP) ranging from 24 to 45, by using the reference software for the MVC, the joint multiview video coding [45].

Fig. 22 shows that enforcing the interview coherence leads to a significant bit rate saving in compressing the up-sampled depth video. JBU1 (or 2) and WMF1 (or 2) represent total amount of bit rates required to compress the depth maps of three color views up-sampled by the JBU and the WMF with the initial low-resolution depth maps from the depth sensor D_1 (or D_2). Similarly, M-JBU1 (or 2) and M-WMF1 (or 2) represent total amount of bit rates of depth maps up-sampled considering the interview coherence with the initial low-resolution depth maps from the depth sensor D_1 (or D_2). In Fig. 22(c), the effect of interview coherence on the R-methods (using depth sensor D_1 and D_2) was also analyzed. Table IV and V show that the results with the interview coherence (the M-methods and the RM-methods) reduce the bit rates of 6.22% and increase the PSNR of 0.41dB averagely, compared to other results with no interview coherence. In this experiment, we compared the M-methods with the conventional methods and the RM-methods with R-methods in order to fairly verify the effect of the interview coherence on the multiview video coding.

VI. CONCLUSION

In this paper, a novel approach for producing high-quality multiview depth maps was proposed. The proposed method increased the depth quality by improving the interview consistency on the up-sampled multiview depth maps as well as by considering the reliability of initial multiple depth maps obtained from multiple depth sensors. It was verified through various experiments that the multiview depth maps up-sampled by the proposed method were aligned with the corresponding color images very well. The experimental results show that the proposed method outperforms the existing up-sampling methods, which have been usually developed to enhance single depth map only. We also demonstrated that our method also decreases the total amount of bit rates in compressing multiview depth videos.

In comparison to the stereo matching-based hybrid approaches, our framework can be easily applied into relatively low-cost up-sampling algorithms, such as the JBU and the WMF. In the proposed framework, the convenient adjustment of the sensors allows high-quality freeview video to be obtained in a relatively compact manner, in that a computationally-heavy stereo algorithm is not employed. In terms of the computational complexity which is very crucial to producing a high-resolution 3-D video, we showed that the proposed method is more efficient than other hybrid methods using stereo matching algorithms together, but with comparable depth maps.

The proposed method will be implemented on graphics processing units for obtaining a real-time performance in future works. Moreover, the active depth sensor used in the proposed method would be substituted with different type of sensor such as Kinect [17], making the proposed system more widely applicable. Further research will also include accurate 3-D reconstruction based on the proposed approach.

REFERENCES

[1] J. Duan and J. Li, "Compression of the layered depth image," *IEEE Trans. Image Process.*, vol. 12, no. 3, pp. 365–372, Mar. 2003.
[2] C. T. E. R. Hewage, S. T. Worrall, S. Dogan, S. Villette, and A. M. Kondoz, "Quality evaluation of color plus depth map-based stereoscopic video," *J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 304–318, 2009.
[3] B. Bartczak and R. Koch, "Dense depth maps from low resolution time-of-flight depth and high resolution color views," in *Proc. 5th Int. Symp. Adv. Visual Comput.*, 2009, pp. 228–239.
[4] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proc. Int. Conf. Image Process.*, 2007, pp. 201–204.
[5] S.-U. Yoon and Y.-S. Ho, "Multiple color and depth video coding using a hierarchical representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1450–1460, Nov. 2007.
[6] J. Y. Lee, H. Wey, and D.-S. Park, "A fast and efficient multi-view depth image coding method based on temporal and interview correlations of texture images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 12, pp. 1859–1868, Dec. 2011.
[7] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2006, pp. 519–528.
[8] M. Levoy, K. Pulli, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton, S. E. Anderson, J. Davis, J. Ginsberg, J. Shade, and D. Fulk, "The digital Michelangelo project: 3-D scanning of large statues," in *Proc. ACM SIGGRAPH*, 2000, pp. 131–144.

[9] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, nos. 1–3, pp. 7–42, 2002.
[10] B. J. Tippetts, D.-J. Lee, J. K. Archibald, and K. D. Lillywhite, "Dense disparity real-time stereo vision algorithm for resource-limited systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 10, pp. 1547–1555, Oct. 2011.
[11] K.-J. Yoon and I.-S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.
[12] D. B. Min and K. Sohn, "Cost aggregation and occlusion handling with WLS in stereo matching," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1431–1442, Aug. 2008.
[13] C. Pham and J. Jeon, "Domain transformation-based efficient cost aggregation for local stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 7, pp. 1119–1130, Jul. 2013.
[14] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. F. Tappen, and C. Rother, "A comparative study of energy minimization methods for Markov random fields with smoothness-based priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068–1080, Jun. 2008.
[15] MESA Imaging [Online]. Available: http://www.mesa-imaging.ch
[16] PMD Technologies [Online]. Available: http://www.pmdtec.com
[17] Microsoft, Kinect [Online]. Available: http://www.xbox.com/en-US/kinect
[18] S. Foix, G. Alenya, and C. Torras, "Lock-in time-of-flight (ToF) cameras: A survey," *IEEE Sensors J.*, vol. 11, no. 9, pp. 1917–1926, Sep. 2011.
[19] S. Gokturk, H. Yalcin, and C. Bamji, "A time-of-flight depth sensor: System description, issues and solutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn. Workshops*, Jun. 2004, p. 35.
[20] R. Larsen, E. Barth, and A. Kolb, "Special issue on time-of-flight camera based computer vision," *Comput. Vision Image Understand.*, vol. 114, no. 12, p. 1317, 2010.
[21] D. Alexiadis, D. Zarpalas, and P. Daras, "Real-time, full 3-D reconstruction of moving foreground objects from multiple consumer depth cameras," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 339–358, Feb. 2013.
[22] S. Guomundsson, H. Aanaes, and R. Larsen, "Environmental effects on measurement uncertainties of time-of-flight cameras," in *Proc. Int. Symp. Signals, Circuits Syst.*, 2007, pp. 1–4.
[23] J. Choi, D. B. Min, B. Ham, and K. Sohn, "Spatial and temporal up-conversion technique for depth video," in *Proc. Int. Conf. Image Process.*, 2009, pp. 3525–3528.
[24] J. Choi, D. B. Min, D. Kim, and K. Sohn, "3-D JBU based depth video filtering for temporal fluctuation reduction," in *Proc. Int. Conf. Image Process.*, 2010, pp. 2777–2780.
[25] D. Min, J. Lu, and M. N. Do, "Depth video enhancement based on weighted mode filtering," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1176–1190, Mar. 2012.
[26] J. Choi, D. Min, and K. Sohn, "2-D-plus-depth based resolution and frame-rate up-conversion technique for depth video," *IEEE Trans. Consumer Electron.*, vol. 56, no. 4, pp. 2489–2497, Nov. 2010.
[27] *3D4YOU-Content Generation and Delivery for 3-D Television* [Online]. Available: http://www.3d4you.eu
[28] A. Frick, B. Bartczack, and R. Koch, "3-D-TV LDV content generation with a hybrid ToF-multicamera rig," in *Proc. 3-DTV-Conf. True Vision Capture, Transmission Display 3-D Video*, 2010, pp. 1–4.
[29] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, 2007.
[30] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2007, pp. 1–8.
[31] Q. Yang, K.-H. Tan, W. B. Culbertson, and J. G. Apostolopoulos, "Fusion of active and passive sensors for fast 3-D capture," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Oct. 2010, pp. 69–74.
[32] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A noise-aware filter for real-time depth upsampling," in *Proc. Workshop Multi-camera Multi-modal Sensor Fusion Algorithms Applicat.*, 2008, pp. 1–12.
[33] B. Huhle, T. Schairer, P. Jenke, and W. Straßer, "Fusion of range and color images for denoising and resolution enhancement with a non-local filter," *Comput. Vision Image Understand.*, vol. 114, no. 12, pp. 1336–1345, 2010.
[34] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proc. Adv. Neural Inform. Process. Syst.*, 2005, pp. 291–298.
[35] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I.-S. Kweon, "High quality depth map upsampling for 3D-TOF cameras," in *Proc. IEEE Int. Conf. Comput. Vision*, Nov. 2011, pp. 1623–1630.
[36] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2008, pp. 1–8.
[37] J. Zhu, L. Wang, R. Yang, J. E. Davis, and Z. Pan, "Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1400–1414, Jul. 2011.
[38] J. Zhu, L. Wang, J. Gao, and R. Yang, "Spatial-temporal fusion for high accuracy depth maps using dynamic MRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 899–909, May 2010.
[39] *Point Grey Research* [Online]. Available: http://www.ptgrey.com
[40] *Camera Calibration Toolbox for Matlab Provided by Caltech* [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc
[41] [Online]. Available: http://vision.middlebury.edu/stereo
[42] [Online]. Available: http://diml.yonsei.ac.kr/cjw0512/reliability
[43] M. Humenberger, T. Engelke, and W. Kubinger, "A census-based stereo vision algorithm using modified semi-global matching and plane fitting to improve matching quality," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn. Workshops*, Jun. 2010, pp. 77–84.
[44] J. Sun, N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 787–800, Jul. 2003.
[45] *WD 4 Reference Software for MVC*, ISO/IEC JTC1/SC29/WG11/JVT-AD207, Geneva, Switzerland, 2009.

Jinwook Choi (S'09) received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2008, where he is currently pursuing the joint M.S. and Ph.D. degree in electrical and electronic engineering.

His current research interests include 3-D image and video processing, computer vision, 3-D modeling, hybrid sensor systems, super-resolution, HDR imaging, and intelligent vehicle systems.

Dongbo Min (M'09) received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2003, 2005, and 2009, respectively.

He was a Post-Doctoral Researcher with the Mitsubishi Electric Research Laboratories, Cambridge, MA, USA, from June 2009 to June 2010. He is currently with the Advanced Digital Sciences Center, which was jointly founded by the University of Illinois at Urbana-Champaign, Urbana, IL, USA, and the Agency for Science, Technology, and Research, a Singapore government agency. His current research interests include 3-D computer vision, video processing, 3-D modeling, and hybrid sensor systems.

Kwanghoon Sohn (M'92–SM'12) received the B.E. degree in electronic engineering from Yonsei University, Seoul, Korea, in 1983, the M.S.E.E. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 1985, and the Ph.D. degree in electrical and computer engineering from North Carolina State University, Raleigh, NC, USA, in 1992.

He was a Senior Member of the Research Staff with the Satellite Communication Division, Electronics and Telecommunications Research Institute, Daejeon, Korea, from 1992 to 1993, and a Post-Doctoral Fellow with the MRI Center, Medical School, Georgetown University, Washington, DC, USA, in 1994. He was a Visiting Professor with Nanyang Technological University, Singapore, from 2002 to 2003. He is currently a Professor with the School of Electrical and Electronic Engineering, Yonsei University. His current research interests include 3-D image processing, computer vision, and image communication.

Dr. Sohn is a member of the SPIE.