Depth Video Enhancement Based on Weighted Mode Filtering

Dongbo Min, *Member, IEEE*, Jiangbo Lu, *Member, IEEE*, and Minh N. Do, *Senior Member, IEEE*

***Abstract*—This paper presents a novel approach for depth video enhancement. Given a high-resolution color video and its corresponding low-quality depth video, we improve the quality of the depth video by increasing its resolution and suppressing noise. For that, a weighted mode filtering method is proposed based on a joint histogram. When the histogram is generated, the weight based on color similarity between reference and neighboring pixels on the color image is computed and then used for counting each bin on the joint histogram of the depth map. A final solution is determined by seeking a global mode on the histogram. We show that the proposed method provides the optimal solution with respect to L_1 norm minimization. For temporally consistent estimate on depth video, we extend this method into temporally neighboring frames. Simple optical flow estimation and patch similarity measure are used for obtaining the high-quality depth video in an efficient manner. Experimental results show that the proposed method has outstanding performance and is very efficient, compared with existing methods. We also show that the temporally consistent enhancement of depth video addresses a flickering problem and improves the accuracy of depth video.**

***Index Terms*—Depth enhancement, depth sensor, multiscale color measure (MCM), temporal consistency, weighted mode filtering (WMF).**

I. Introduction

PROVIDING high-quality depth data has been one of the most important issues in the field of 3-D computer vision and can be used in many applications such as image-based rendering, 3DTV, 3-D object modeling, robot vision, and tracking. The acquisition process of accurate depth data at high resolution is nontrivial, and a variety of depth measuring methods have been developed. For example, although laser range scanner or active illumination with structured lights can provide highly accurate depth data, they are available in the limited applications such as a static environment only. Stereo matching methods can provide a depth map in real-time through a support of specialized hardware such as a graphics processing unit (GPU) [1], [2]. A number of methods have been proposed by using several cost aggregation methods [3], [4] and global optimization techniques [14], but their performance is still far from a practical solution due to lighting/occlusion problems, huge computational complexity, etc.

Recently, depth sensors such as time-of-flight (ToF) camera have been widely used in research and practice. The ToF sensor, which is based on a special complementary metal–oxide–semiconductor pixel structure, estimates the distance between the sensor and an object by extracting phase information from received light pulses [19]. Since it provides a 2-D depth map at video rate, it can be used in a dynamic environment [5]. However, the quality of depth maps obtained by TOF camera is not satisfactory due to an inherent physical limit of depth sensor. For example, depth maps obtained by a ToF sensor, i.e., "Mesa Imaging SR4000," are of low resolution (176×144) and noisy [31].

In order to overcome the physical limit of the depth sensor, Diebel and Thrun proposed a depth upsampling method based on MRF formulation by using a low-resolution depth map and its corresponding single high-resolution color image [7]. The minimization problem with MRF formulation is solved with a conjugate gradient algorithm. However, the output depth map has worse quality due to nonrobustness of a quadratic function. Lu *et al.* [8] presented an MRF-based depth upsampling method that uses a novel data term formulation which fits well to the characteristics of depth maps. This method provides high-quality depth maps, but it is computationally heavy due to a complex optimization technique. Park *et al.* [9] proposed an MRF optimization framework that combines the high-resolution color image with a nonlocal means (NLM) method [23]. They described an objective function that consists of data term, smoothness term, and NLM regularization term. In the smoothness term, a confidence weighting function is defined by using color information, segmentation, edge saliency, and bicubic interpolated depth map. The NLM regularization term is utilized to preserve thin structures by allowing pixels with similar structure to reinforce with each other [9].

Kopf *et al.* [10] presented a general framework for multimodal image enhancement and applied it to several image processing tasks such as colorization, tone mapping, and depth upsampling. In particular, the low-resolution depth map is upsampled with a guide color image in the context of a bilateral filtering scheme by leveraging the color image as a prior. Yang *et al.* [11] proposed a new method for depth upsampling with an iterative joint bilateral upsampling (JBU). In contrast to the JBU [10] that applies the filtering procedure to the depth value, they build a 3-D cost volume based on current disparity value. The joint bilateral filtering is then performed for a 2-D cost section of each depth candidate, and a final disparity value is selected by using the winner-takes-all (WTA) technique on the 3-D cost volume after a fixed number of iterations. The iterative bilateral

Manuscript received April 04, 2011; revised July 02, 2011; accepted July 12, 2011. Date of publication July 29, 2011; date of current version February 17, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Anthony Vetro.

D. Min and J. Lu are with the Advanced Digital Sciences Center, Singapore 138632 (e-mail: dbmin99@gmail.com; jiangbo.lu@adsc.com.sg).

M. N. Do is with the University of Illinois at Urbana-Champaign, Urbana, IL 61820-5711 USA (e-mail: minhdo@illinois.edu).

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/TIP.2011.2163164

filtering on the cost domain results in better edge-preserving performance, but its computational complexity is D times of that of the 2-D JBU [10], where D is the number of depth candidates. Hierarchical depth upsampling [12] was proposed for an efficient implementation, but the complexity is still high and dependent on the number of depth candidates. In this paper, we call the JBU by Kopf *et al.* [10] 2-D JBU, and the approach of Yang *et al.* [11] 3-D JBU. Another filtering-based depth upsampling method [13] was proposed on the formulation of the nonlocal means method [23]. In order to enhance the quality of the depth map while preserving fine details and depth discontinuities, an intrapatch similarity on the depth map and the corresponding color information are taken into account at the same time.

There are other existing methods for providing high-quality depth maps. Zhu *et al.* presented a method based on probabilistic fusion of ToF depth sensor and stereo camera [15]. Two cost functions, which are calculated from the ToF and stereo cameras, are adaptively combined into a data term based on a reliability value of each sensor data. A final energy function is defined on MRF formulation with a smoothness constraint and solved by loopy belief propagation [14]. This approach was extended into spatial–temporal fusion for generating depth video, which is temporally consistent over frames [16]. They used a spatial–temporal MRF formulation for taking temporal neighbors, calculated by an optical flow method, into account.

Different from these approaches on fusion of stereo and active depth sensors, our method focuses on generating high-resolution depth video with one color video (not stereo) and its corresponding low-resolution depth video. For that, we propose a *weighted mode filtering* (WMF) based on a joint histogram. The weight based on similarity measure between reference and neighboring pixels is used to construct the histogram, and a final solution is then determined by seeking a global mode on the histogram. The joint filtering means that the weight is computed with a signal different from the signal to be filtered, and it enables the histogram to be extended into a weighted filtering. We also show that the proposed filtering technique forces the filtered value to be the solution for L_1 norm minimization, which is more robust to outliers of data than L_2 norm minimization.

Weijer and Boomgaard proposed a *local* (not global) mode filtering, which is a histogram-based nonlinear filtering to preserve edges and details while suppressing noise on an image [17]. The histogram is computed from a set of data that consists of neighbors of a reference pixel, and then, it seeks a local mode on the histogram iteratively after setting an intensity of the pixel as an initial solution. Given the pixel and its corresponding histogram, the local mode filtering converges to the closest local mode on the histogram in an iterative manner [17]. In [18], it was proved that the local mode filtering is equivalent to the bilateral filtering [6], which provides an optimal solution with respect to L_2 norm minimization.

Our proposed WMF provides a solution that is optimal with respect to L_1 norm minimization, and it effectively enhances a depth video by deblurring and upsampling. Fig. 1 shows that the proposed method have the best edge-preserving performance. The synthesized view using the depth map upsampled by the proposed method is superior to those of the 2-D JBU and the 3-D JBU. We will describe this feature of the global mode filtering and its relationship with the filtering-based methods in detail later. Moreover, in order to model the characteristics of color information on an image more accurately, a multiscale color measure (MCM) is also proposed in the depth enhancement step when sampling factor $s > 1$. The color similarity between the sparse depth data on the color image grid is measured on a multigrid framework, and it leads to considering color distribution of neighboring pixels between sparse depth data completely.

Another contribution of this paper is to enforce temporal consistency in the procedure for depth video enhancement. We extend the WMF into a temporal domain, which generates an improved and flicker-free depth video. Temporal neighbors are determined by an optical flow method, and a patch-based reliability measure of the optical flow is used further in order to cope with errors of the estimated optical flow. This is a simplified formulation of a nonlocal video denoising [23]. Although the nonlocal video denoising shows excellent performance since it uses a set of all the possible pixels on the temporally neighboring frames, its complexity is huge. Our approach hence selects one pixel corresponding to the estimated optical flow for each frame. Since it uses the additional color information for enhancing the depth video, one optical flow vector and its patch-based reliability measure are enough to reduce the flickering problem and improve the performance, which is similar to [16].

The remainder of this paper is organized as follows: In Section II, we present the depth enhancement based on the WMF and discuss the relations to other joint filtering approaches such as the 2-D JBU [10], the 3-D JBU [11], and a histogram-based voting [22]. The MCM and temporally consistent estimate are then described in Sections III and IV. Finally, we present experimental results and conclusion in Sections V and VI, respectively.

II. Weighted Mode Filtering For Depth Enhancement

In general, the quality of depth video can be measured by the amount of noise, spatial resolution, and temporal consistency. Therefore, the depth video can be improved by suppressing the noise, increasing its spatial resolution, and handling the temporal flickering problem. In this paper, we propose a novel method based on the joint histogram for achieving these goals.

A. WMF on Histogram

Histogram of an image represents the number of pixels inside given rectangular (or any shape) regions, which corresponds to each bin, and can be referred to as a probability distribution of pixel values after a normalization step. The local mode filtering [17] constructs a relaxed histogram where each pixel is modeled by Gaussian distribution. Given 2-D function $f(p)$ of an image, relaxed histogram $H(p,d)$ at reference pixel p and dth bin can be defined as follows:

$$H(p,d) = \sum_{q \in N(p)} G_r\left(d - f(q)\right) \tag{1}$$

where G_r is a Gaussian function and $N(p)$ represents a set of neighboring pixels of pixel p. In order to compute localized his-

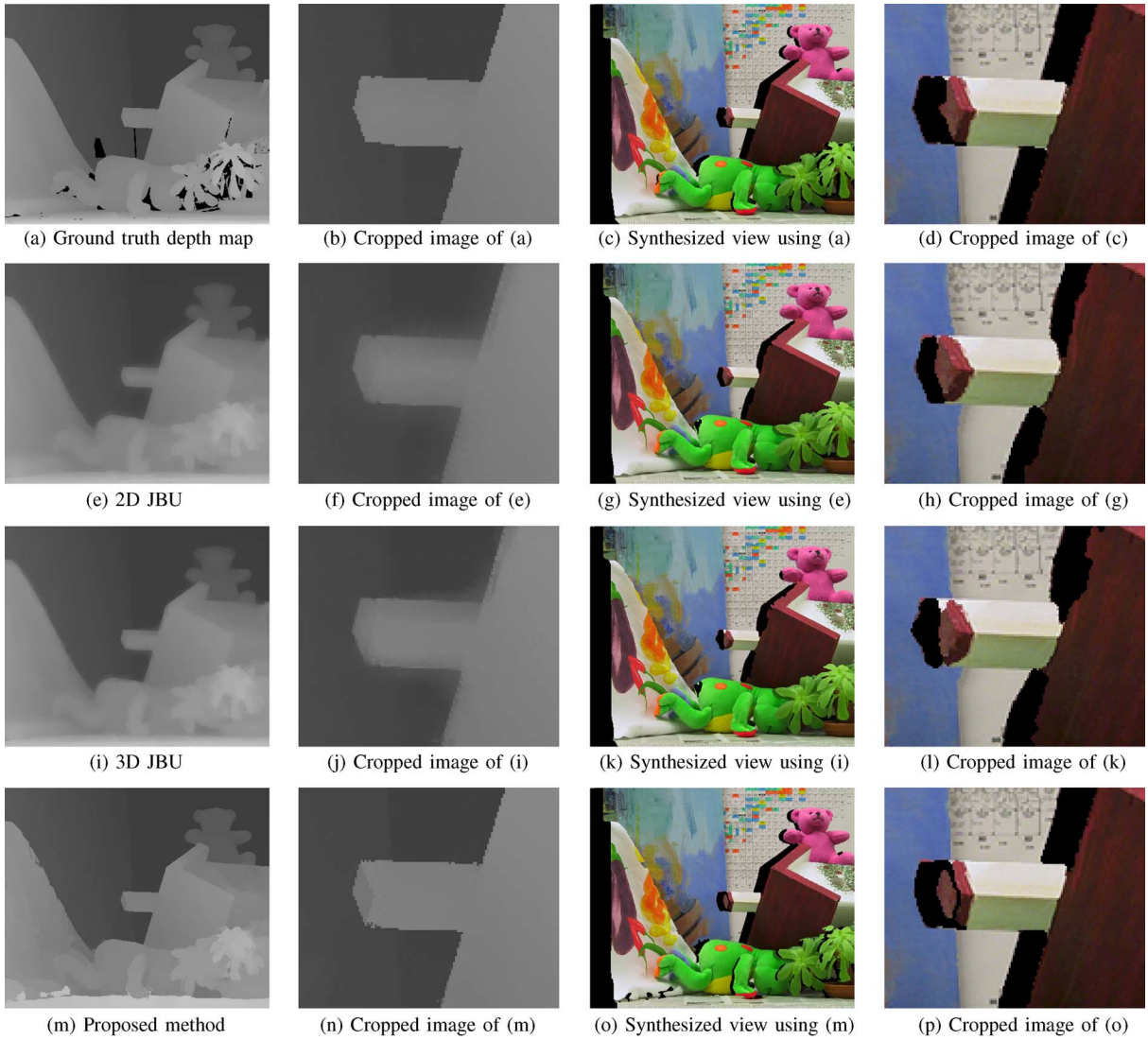


Fig. 1. Depth upsampling results for "Teddy" image: (a) Ground truth depth map, (e) 2-D JBU [10], (i) 3-D JBU [11], and (m) proposed method (WMF + MCM). The upsampling ratio is 8 in each dimension. The processing times are (e) 0.95, (i) 220, and (m) 0.55 s, respectively. The processing time of the proposed method is 0.25% of that of 3-D JBU, while it has the best edge-preserving performance. The virtual view is synthesized by using an original color image and the corresponding upsampled depth map. We can see that the enhanced depth and the virtual view of the proposed method are superior to those of 2-D JBU and 3-D JBU. (a) Ground truth depth map. (b) Cropped image of (a). (c) Synthesized view using (a). (d) Cropped image of (c). (e) Two-dimensional JBU. (f) Cropped image of (e). (g) Synthesized view using (e). (h) Cropped image of (g). (i) Three-dimensional JBU. (j) Cropped image of (i). (k) Synthesized view using (i). (l) Cropped image of (k). (m) Proposed method. (n) Cropped image of (m). (o) Synthesized view using (m). (p) Cropped image of (o).

togram $H_L(p,d)$ around pixel p, spatial Gaussian function G_s is also introduced as follows:

$$H_L(p,d) = \sum_{q \in N(p)} G_r\left(d - f(q)\right) G_S(p-q). \tag{2}$$

Pixels that are closer to reference pixel p have a larger weighting value. Local mode filtering then seeks local mode $f_L(p)$ that is the closest to intensity $f(p)$ of reference pixel p. In other words, the local minimum which is close to the initial value is chosen among a set of multiple local minima and [17] shows its effectiveness on an image denoising. Paris *et al.* [18] shows that the iterative local mode seeking step is equivalent to an iterative bilateral filtering. The $(m+1)$th solution f_L^{m+1} for the local mode filtering can be computed as follows:

$$f_L^{m+1}(p) = \frac{\sum\limits_{q \in N(p)} G_r\left(f_L^m(p) - f_L^m(q)\right) G_S(p-q) f_L^m(q)}{\sum\limits_{q \in N(p)} G_r\left(f_L^m(p) - f_L^m(q)\right) G_S(p-q)}. \tag{3}$$

Bilateral filtering is a nonlinear summation for smoothing an image while maintaining edges and details [6]. Its main idea is to combine color and spatial similarity measure between reference and neighboring pixels. A number of works have shown the relation between bilateral filtering, robust statistics, and nonlinear diffusion based on a Bayesian framework [20]. Elad [21] proved that the bilateral filtering is a solution after one iteration of the

Jacobi algorithm on Bayesian approach. The energy function is defined by using weighted least square based on L_2 norm, and a weighting function is defined by a Gaussian distribution. This method can be extended into the joint bilateral filtering [10] by using guide signal $g(p)$ different from reference signal $f(p)$ to be filtered as follows:

$$f_L^{m+1}(p) = \frac{\sum_{q\in N(p)} G_I\left(g(p)-g(q)\right) G_S(p-q) f_L^m(q)}{\sum_{q\in N(p)} G_I\left(g(p)-g(q)\right) G_S(p-q)} \quad (4)$$

where G_I is a Gaussian function for $g(p)$ whose value does not change over the iteration.

This paper proposes the weighted mode filtering that seeks *global mode* on the histogram by leveraging similarity measure between data of two pixels. When histogram $H_G(p,d)$ for the weighted mode filtering is generated, the data of each pixel inside rectangular (or any shape) regions is adaptively counted on its corresponding bin by using data similarity between reference and neighboring pixels, i.e.,

$$H_G(p,d) = \sum_{q\in N(p)} G_I\left(g(p)-g(q)\right) G_S(p-q) G_r\left(d-f(q)\right). \quad (5)$$

In this paper, $G_I(x)$, $G_S(x)$, and $G_r(x)$ are defined as Gaussian functions, where means are 0 and standard deviations are σ_I, σ_S, and σ_r, respectively. Final solution $f_G(p)$ for the WMF can be computed as follows:

$$f_G(p) = \arg\max_d H_G(p,d). \quad (6)$$

Here, $g(p)$ is a 2-D function where each pixel p has a specific data. In this paper, we focus on the WMF where G_I is a weight function of guide signal $g(p)$ different from reference signal $f(p)$ to be filtered. Since the guide signal is employed for calculating data-driven adaptive weight G_I, the proposed filtering is contextualized within the joint bilateral filtering framework [10]. However, as shown in Fig. 1, it has better edge-preserving performance than the joint bilateral filtering on the object discontinuities. The relation with the joint bilateral filtering will be described in the following section. The performance and effectiveness of the proposed method are verified by applying it to depth video enhancement, provided from ToF depth sensor. In the case of the depth enhancement task, $g(p)$ is color image $I(p)$ and $f(p)$ is depth image $d(p)$.

Fig. 2 explains the procedure that generates joint histogram H_G. Neighboring pixels m_1 and m_2 of reference pixel p are adaptively counted with a form of Gaussian function on each bin corresponding to their disparity values. The bandwidth and magnitude of Gaussian function are defined by standard deviation σ_r of G_r and the magnitude of $G_I G_S$, respectively. Standard deviation σ_r of Gaussian spreading function G_r is used for modeling errors that may exist on the input depth data. In other words, the neighboring pixels are adaptively accumulated on joint histogram H_G by using color (G_I) and spatial (G_S) similarity measures and Gaussian error model (G_r).

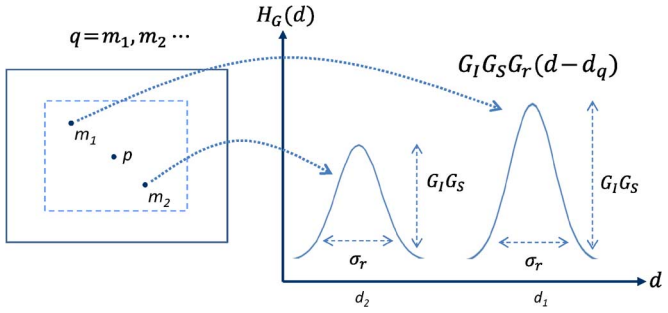


Fig. 2. Joint histogram generation: Joint histogram H_G of reference pixel p is calculated by adaptively counting neighboring pixels m_1 and m_2 with a form of Gaussian function according to their disparity values d_1 and d_2. The bandwidth and magnitude of the Gaussian function are defined by standard deviation σ_r of G_r and the magnitude of $G_I G_S$, respectively.

In general, a depth value smoothly varies inside objects and has sharp discontinuities on the object boundaries. When the depth is upsampled by the 2-D JBU, the filtered output depth value is provided by using an adaptive summation based on color information. Although an adaptive weight based on color information is used for preserving edges, it still results in unnecessary blur due to its *summation*.

B. Relations With Other Joint Filtering Approaches

As previously mentioned, there are some existing approaches for the depth enhancement (upsampling). Here, we discuss the relations with the existing approaches, particularly for the joint-filtering-based methods.

1) 2-D JBU: The 2-D JBU is an adaptive summation with a guide of corresponding color image [10]. We modify joint histogram $H_G(p,d)$ in (5) by replacing $G_r(m)$ with delta function $\delta_r(m)$, which is 1 when $m=0$, 0 otherwise, i.e.,

$$H'_G(p,d) = \sum_{q\in N(p)} G_I(I_p - I_q) G_S(p-q)\delta\left(d-d(q)\right). \quad (7)$$

The joint bilateral filtering can be then written by using (7) as follows:

$$d_{\mathrm{L}}(p) = \frac{\sum_{q\in N(p)} G_I(I_p-I_q) G_S(p-q) d(q)}{\sum_{q\in N(p)} G_I(I_p-I_q) G_S(p-q)} = \frac{\sum_d d H'_G(p,d)}{\sum_d H'_G(p,d)}. \quad (8)$$

We can find that the joint bilateral filtering computes the output depth value by adaptively summing all depth candidates according to joint histogram H'_G, whereas the WMF in (6) selects the output depth value whose histogram value is the largest among all depth candidates. Fig. 3 shows the relation between the 2-D JBU and the WMF on the joint histogram. In other words, the joint bilateral filtering provides a mean value through the adaptive summation, which is optimal with respect to L_2 norm minimization. In contrast, the WMF picks the output value that has the largest histogram value, which is optimal with respect to L_1 norm minimization. Another difference is that the joint bilateral filtering uses delta function

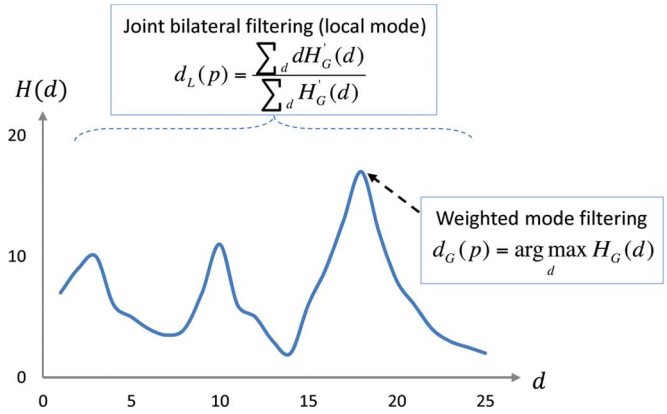


Fig. 3. Relation between 2-D JBU and WMF. The joint bilateral filtering provides the mean value through the adaptive summation (L_2 norm minimization), whereas the WMF picks the output value that has the largest histogram value (L_1 norm minimization).

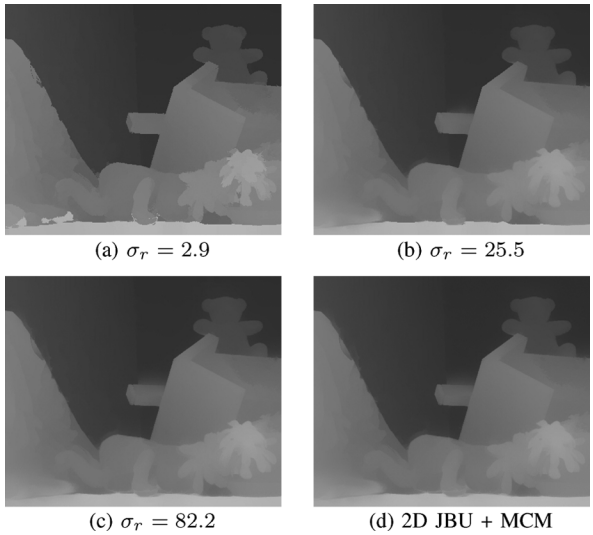


Fig. 4. WMF versus joint bilateral filtering (local mode). (a)–(c) The results of the WMF. As σ_r increases, the results of the WMF become similar to that of the joint bilateral filtering. Note that the result in (d) is different from Fig. 1(a) since the depth maps were upsampled by the MCM with original sparse depth map, which will be described in the following section. (a) $\sigma_r = 2.9$. (b) $\sigma_r = 25.5$. (c) $\sigma_r = 82.2$. (d) Two-dimensional JBU + MCM.

δ_r for modeling each depth data inside window $N(p)$, namely, weighting parameter σ_r of G_r is 0 in (5).

The joint bilateral filtering has two parameters, i.e., σ_I that defines the data-driven smoothing and σ_s that defines the spatial locality of the neighboring pixels. In contrast, the weighted mode filtering has three parameters, i.e., two parameters that are the same to those of the joint bilateral filtering and σ_r that decides the bandwidth of the Gaussian function, as shown in Fig. 2.

As σ_I increases on the joint bilateral filtering, $f_L(p)$ in (4) corresponds to a standard Gaussian filtered value, which results in a blurred output on the object discontinuities. In the WMF, as σ_r increases, the global mode of joint histogram approaches the solution of the joint bilateral filtering. Fig. 4 shows the results of the WMF and the joint bilateral filtering (which is equivalent to local mode filtering). We found that as σ_r increases, the results of the WMF become similar to those of the joint bilateral filtering. In particular, when $\sigma_r = 82.2$, the results of two methods are almost same. Note that these results were all estimated with the same initial depth map (original sparse depth map) by using a MCM, which will be described in the following section. Therefore, the result in Fig. 4(d) of the joint bilateral filtering is different from that in Fig. 1(a).

2) 3-D JBU: The 3-D JBU [11] builds a 3-D cost volume $E^m(p,d)$ by using a current depth map $d^m(p)$ and then perform the joint bilateral filtering for 2-D cost section of each depth candidate. For an initial input disparity map $d^0(p)$, the 3-D JBU is performed as follows:

$$E^0(p,d) = e(p,d) = \min\left(\left|d - d^0(p)\right|, \sigma\right) \tag{9}$$

$$E^{m+1}(p,d) = \frac{\sum_{q\in N(p)} G_I(I_p - I_q) G_S(p-q) E^m(q,d)}{\sum_{q\in N(p)} G_I(I_p - I_q) G_S(p-q)} \tag{10}$$

$$d^{m+1}(p) = \arg\min_{d\in[0,D)} E^{m+1}(p,d) \tag{11}$$

where σ is a threshold for truncation at the penalty function (9). Although it has a better performance than the 2-D JBU, the computational complexity is huge. If the complexity of the 2-D JBU is $O(PW)$, the 3-D JBU has a complexity of $O(PWDR)$, where P and W are the size of image and window N, respectively. D is the number of depth candidates, and R represents the number of iterations. In this paper, R is set to 1, namely, it is a noniterative scheme.

The 3-D JBU computes the output depth value by finding the minimum value among the depth candidates on the 3-D cost volume. In a sense that the WMF computes the solution by finding the maximum on the joint histogram, the 3-D JBU and the WMF use a similar principle. For finding a maximum value on the 3-D cost volume, we redefine cost function $e_R(p,d)$ as follows:

$$e_R(p,d) = \max\left(1 - \frac{1}{\sigma}\left|d - d^0(p)\right|, 0\right). \tag{12}$$

After applying the same joint bilateral filtering in (10), an output depth value, which is a maximum value on the 3-D cost volume, is the same to the solution in (11). If we assume that the new cost function $e_R(p,d)$ is an approximated one of Gaussian function G_r in (5), 3-D cost volume $E(p,d)$ in (10) plays a similar role to joint histogram $H_G(p,d)$ in (5), except that $E(p,d)$ is computed after the normalization step. Fig. 5 shows the relation between $G_r(d)$ and $e_R(d)$, where p is omitted. G_r and e_R model errors of the input depth map with Gaussian function and linear function, respectively.

The computational complexity of the WMF can be defined as $O(PWB)$, where B is the width of the Gaussian function G_r in Fig. 5 and determined by the number of depth candidates D and weighting parameter σ_r, which depends on the amount of noise in the input depth map. In this paper, when the number of depth candidates D is 256, B is set to 9–39. Since the 3-D JBU performs the joint bilateral filtering for all depth candidates, which consists of the adaptive summation and normalization step, the computational complexity is much higher than that of the WMF. We will show the complexity analysis in the experimental results.

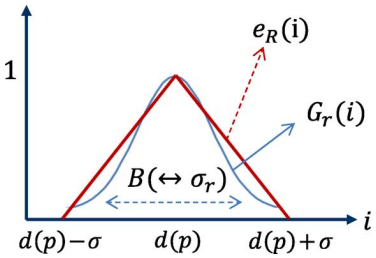


Fig. 5. Relation between $G_r(i)$ in (5) and $e_R(i)$ in (12), where p is omitted. e_R can be referred to as the approximated one of Gaussian function G_r.

3) Histogram-Based Voting: Lu *et al.* proposed a voting method based on a histogram for estimating the depth map on stereo images [22]. The voting method is used for refining the initial depth map on the histogram. It estimates an adaptive region for each pixel with an anisotropic local polynomial approximation (LPA) technique and builds the histogram with depth values inside the estimated region. The histogram is then used for refining the initial depth map. While our method is based on the joint histogram with a *soft* constraint, this can be referred to as the histogram-based approach with a *hard* constraint since the histogram is counted with a constant value (usually 1) on the adaptive region, which is determined by the LPA technique.

III. MCM

Here, we propose a method for measuring a color distance in the multiscale framework. Note that this method is for depth upsampling only. In other words, the MCM is used for preventing an aliasing effect that may happen in the depth upsampling task.

The sparse original depth values mapped into the color camera coordinate only are used for preventing the output depth value from being blurred on the depth boundaries, different from previous approaches [10], [11], which used interpolated depth values to initialize the input depth map. In order to include this notation in (5), we define binary function $R(p)$, whose value is 1 when p has an original depth value, 0 otherwise. Histogram $H_G(p,d)$ can be expressed as follows:

$$H_G(p,d) = \sum_{q \in N(p)} R(q) G_I(I_p - I_q) G_S(p-q) G_r\left(d - d(q)\right). \tag{13}$$

As shown in Fig. 6, however, if we use the sparse original depth values for the depth upsampling directly, it may cause the aliasing artifact due to different size between the original depth and color images. In (13), since color distance $G_I(I_p - I_q)$ of neighboring pixels are calculated by using sparse pixels only where they have depth values $(R(p) = 1)$, this color measure cannot represent the distribution of color information inside window $N(p)$. The existing methods [10], [11] have handled this problem by applying prefiltering methods such as bilinear or bicubic interpolation. However, this initial depth map contains contaminated values that may cause serious blur on the depth boundaries. In this paper, we handle this problem by using the MCM, instead of applying the prefiltering techniques. This method can provide an aliasing-free upsampled depth map and preserve the depth discontinuities well.

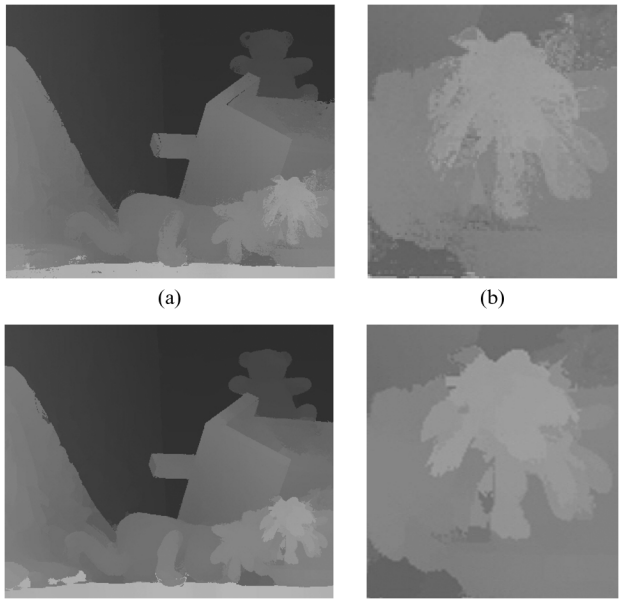


Fig. 6. Aliasing effect in the depth upsampling: Different from the previous approaches that use the bilinear or bicubic interpolations based on low-pass filtering, the sparse original depth values are used only in the proposed method. However, this may result in the aliasing effect, as shown in (a). This problem can be handled by using the MCM in (c). (a) Aliased depth map. (b) Cropped image of (a). (c) Fig. 1(c) (with MCM). (d) Cropped image of (c).

Before explaining the method in detail, we define some parameters for helping readers to understand it. Specifically, let the resolution difference between the original low-resolution depth map and high-resolution color image $S_{dc} = min(H_c/H_d, W_c/W_d)$, where H_d and H_c are the height of the depth and color images, respectively, and W_d and W_c are width of the depth and color images, respectively. The number of level $L(l : L-1 \sim 0)$ on the multiscale framework is set to $log_2 S_{dc}$. Window $N(p)$ on lth level is defined as $\{q|\ |p-q|_\infty \leq 2^l S_W\}$, where S_W is the size of the window on the original small depth domain. Namely, the actual size of window $N(p)$ is dependent on the upsampling ratio since the sparse original depth values only are used. We also define a new Gaussian filtered color image $I_G = G * I$, which is used for calculating the color distance on each level of the multiscale framework.

The sparse depth map can be upsampled by using (13) in a coarse-to-fine manner. Table I shows the pseudocode of the depth upsampling with the MCM. Gaussian lowpass-filtered color image I_G is first computed for each level, and the WMF is performed on each level by using the original and upsampled depth values only. For instance, if L is 3, the upsampling procedure starts for every $4(= 2^2)$ pixels on the coarsest level, and binary function $R(p)$ of these pixels are set to 1. In other words, the depth value of pixels upsampled on the current level can be used on the next level again. The variance of the Gaussian function for low-pass filtering of I_G is proportional to the ratio of downsampling on each level. Different from the conventional downsampling procedure where low-pass filtering is first applied and an image is then downsampled, Gaussian low-pass filtered color image I_G is first computed and color distance $G_I(I_G(p) - I_G(q))$ is then calculated on the *full*

TABLE I
PSEUDOCODE OF DEPTH UPSAMPLING WITH MCM ON HIERARCHICAL SCHEME

The number of levels on hierarchical scheme: $L = log_2 S_{dc}$

Set $l = L - 1$.
While $l \geq 0$
  For all p *which meets* $p\%2^l = 0$
    1: Compute Gaussian lowpass filtered color image $I_G = I * G$.
    2: Perform depth upsampling using Eq. (13) and (6).
    3: Get the upsampled $d(p)$ and set $R(p) = 1$.
  End
  $l = l - 1$
End

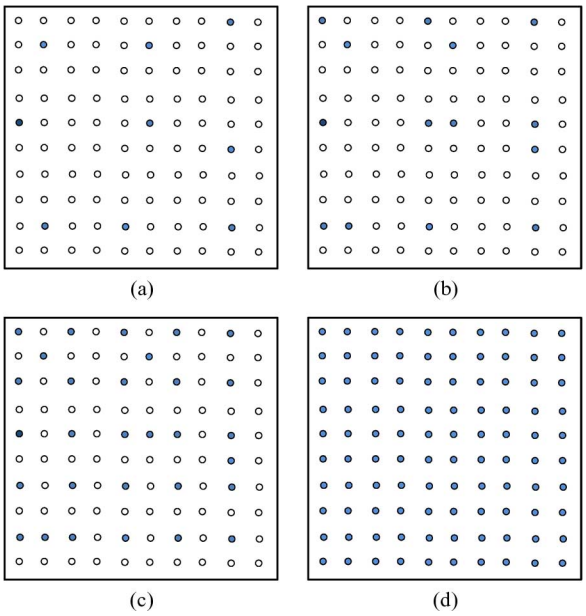


Fig. 7. Example of depth upsampling with MCM on hierarchical scheme: When scale $= 2$, blank pixels are upsampled for every $4(= 2^2)$ pixels with original and upsampled depth values on the full resolution grid. Note that the Gaussian filtered $I_G = G * I$ are used for calculating the color distance on each level. (a) Initial (nonregular) depth. (b) Scale $= 2$ (for every 4 pixels). (c) Scale $= 1$ (for every 2 pixels). (d) Scale $= 0$ (for all pixels).

resolution grid (not coarse resolution grid). In other words, filtered color image I_G is not downsampled.

Fig. 7 shows an example of the depth upsampling with the MCM. Note that the depth and color images are always processed on the full resolution. Given the initial sparse (irregularly sampled) depth map, the depth values on the lth level are upsampled (refined) for every 2^l pixels by using neighboring pixels on the full resolution inside $N(p)$. Instead of changing the size of the color and depth images, the processing unit for upsampling adjusts on each level. This multiscale scheme has two advantages. First, since the size of the color and depth images is the same to all levels, an additional memory used on the coarse level is not needed. Second, the initial sparse depth map, warped into the color camera coordinate, is generally irregularly sampled. Therefore, downsampling of the depth map may lose some information of the irregularly sampled depth data on the coarse domain. The proposed method can consider the irregularly sampled depth data completely on the full resolution for all levels.

IV. ENFORCING TEMPORAL CONSISTENCY

In order to apply the proposed method to the depth video, temporal consistency should be considered by using the information of temporally neighboring frames. Temporally consistent estimate of the correspondences from the low-quality depth video provides a flicker-free depth video and improves the accuracy of an output depth video.

There are some requirements when enforcing temporal consistency. First, an additional complexity for the temporally consistent estimate should be small compared to that of the WMF for a single depth image. The temporal consistency is enforced for handling the flickering problem and improving the quality from the temporal aspect; thus, it is not suitable and meaningful to use a computationally heavy algorithm whose complexity is higher than that of the WMF for a single depth map. Second, the information of the temporal neighbors should be incorporated in a way that is robust to errors on the depth discontinuities. Considering the temporal aspect may cause some problems due to errors that might exist on the information of the temporal neighbors on the depth discontinuities, where a motion vector is generally difficult to estimate. In this paper, the temporal neighbors are determined by Lucas–Kanade (LK) optical flow method [25], [26], and a patch-based reliability measure is used for handling the error of the optical flow. Recently, a number of optical flow estimation methods have been proposed [27]. However, their performance is still far from a practical solution, and the complexity is still too huge to be used in this application.

The LK tracker generally allows estimating a small displacement only over two frames and provides erroneous results on the depth boundaries. We hence use the patch-based reliability measure $w_n(p, p_n)$ between reference pixel p on the tth frame and corresponding pixel p_n on the nth frame with the estimated optical flow together, when combining the joint histograms of the temporally neighboring frames. Fig. 8 shows the WMF for the temporally neighboring depth frames. The temporally consistent joint histogram $H^t_{GT}(p, d)$ of the tth frame can be computed through an adaptive summation of the joint histograms of the temporal neighbors, i.e.,

$$H^t_{GT}(p, d) = \sum_{n \in T(t) \bigcup t} w_n(p, p_n) H^n_G(p_n, d). \quad (14)$$

$T(t)$ represents the neighboring frames of the tth frame. Note that H^t_{GT} is an output joint histogram after applying a temporally adaptive summation on the tth frame. Patch reliability measure $w_n(p, p_n)$ can be computed as follows:

$$w_n(p, p_n) = \frac{1}{Z} \exp\left(- \sum_m \left| I_t(p + m) - I_n(p_n + m) \right| / \sigma_p \right). \quad (15)$$

In this paper, the size of neighboring pixels m and σ_p are set to 5×5 and 40, respectively. Weighting function w_n is computed by the patch-based reliability measure between p on the original tth frame and p_n on the nth frame and then normalized with Z, namely, $\sum_{n \in T(t) \bigcup t} w_n = 1$.

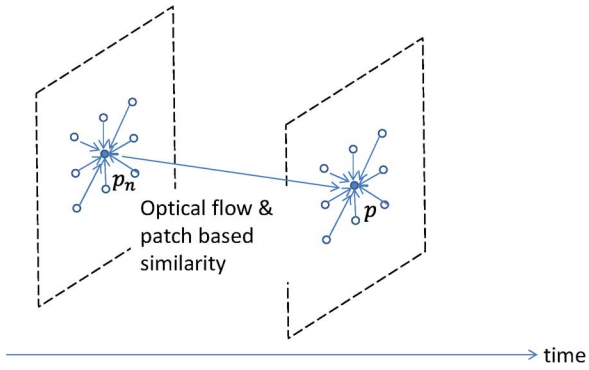


Fig. 8. Temporal neighboring frames for weighted mode filtering.

In order to reuse the temporally consistent estimates on the previous frames, we divide $T(t)$ into previous neighbors $T_{\text{prev}}(t)$ and next neighbors $T_{\text{next}}(t)$, respectively, i.e.,

$$H^t_{GT}(p,d) = \sum_{n\in T_{\text{prev}}(t)} w_n(p,p_n)H^n_{GT}(p_n,d) + \sum_{n\in T_{\text{next}}(t)\bigcup t} w_n(p,p_n)H^n_G(p_n,d). \quad (16)$$

In other words, joint histogram H^t_{GT}, which is computed on the tth frame, is reused as an input joint histogram on the $t+1^{th}$ frame.

This formulation is similar to a nonlocal video denoising [23], which uses the patch-based similarity measure on the neighboring frames. Although it provides the excellent performance and is easy to implement, the complexity is huge since all possible pixels on the temporally neighboring frames are used. In this paper, we hence use a simplified approach that uses both the simple optical flow and the patch-based reliability, which is motivated by the nonlocal means approach. Fig. 9 shows a relation between the nonlocal means approach and the proposed method. Single pixel on each frame is used only so that the complexity is small. The errors of the estimated optical flow are suppressed by patch-based reliability measure w_n. Since the additional color information is also used for enhancing the depth video, one optical flow vector and its patch-based reliability measure are enough to reduce the flickering problem and improve the accuracy.

V. Experimental Results

We have verified the performance of the proposed method through various experiments in terms of edge-preserving performance, noise suppression, flickering reduction, and complexity. The performance was compared with the 2-D JBU [10] and the 3-D JBU [11]. All the experiments were performed on a Dell laptop, which consists of Intel i5 2.4-GHz central processing unit and 6-GB random access memory. We need to classify the depth enhancement task into two categories according to the size of the input depth map: 1) depth upsampling and 2) depth refinement. In the depth upsampling task, the input depth maps, which are noisy and of low resolution, are upsampled with the MCM. In contrast, the depth refinement task improves the accuracy of the input depth maps, estimated by existing stereo matching methods [2]. Thus, the MCM is not used since the color image and its corresponding input depth map are of the same resolution.

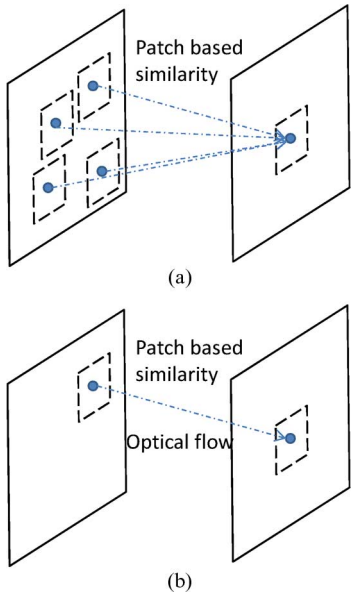


Fig. 9. Nonlocal means and the proposed scheme. In the proposed method, single pixel, which is estimated by the optical flow, is used only, whereas the nonlocal means filtering uses all possible pixels. The errors of the estimated optical flow is suppressed by patch-based reliability measure w_n. (a) Nonlocal means for video. (b) Proposed scheme.

Fig. 10. Experimental setup for the proposed method. The capturing system consists of a "Point Grey Flea" color camera and a "Mesa Imaging SR4000" depth sensor.

A. Geometric Calibration for Fusing Depth and Color Sensors

Given a noisy depth map with low resolution, our goal is to enhance its quality and/or resolution efficiently and accurately, particularly for the depth discontinuities, by using its corresponding color image. Note that the color image should be aligned with the depth map spatially and temporally as they should be acquired at the same viewpoint and time. As shown in Fig. 10, our system consists of a "Mesa Imaging SR4000" depth sensor [31] and a "Point Grey Flea" color camera [32] for acquiring the depth and color videos. Since frame rates of the color and depth sensors are different, time stamps are used to synchronize two sensors. For the spatial alignment, we performed camera calibration and mapped depth data into the color camera coordinate. After calculating a calibration matrix that consists of intrinsic and extrinsic parameters, the depth data provided by the ToF depth sensor is mapped into an image coordinate of

the color sensor. Specifically, let the projection matrices of the depth and color sensors be $M_d = K_d[R_d; -R_dT_d]$ and $M_c = K_c[R_c; -R_cT_c]$, respectively, and the homogeneous image coordinate on the depth and color images $p_d = (x_d, y_d, 1)^T$ and $p_c = (x_c, y_c, 1)^T$, respectively. Given depth value Z_d of pixel p_d on the depth image, 3-D point P_w at the world coordinate corresponding to pixel p_d is computed as follows:

$$P_w = R_d^{-1}\left(K_d^{-1}p_dZ_d\right) + T_d. \tag{17}$$

The 2-D point p_c and its corresponding 3-D point $P_c = (X_c, Y_c, Z_c)^T$ on the color sensor are then computed with the intrinsic and extrinsic parameters K_c, R_c, and T_c as follows:

$$p_c = K_cP_c = K_cR_c(P_w - T_c). \tag{18}$$

Finally, the depth map aligned on the color image is assigned with $D_c(p_c) = Z_c$. Note that this depth value is the distance between object and color cameras, different from displacement vector (pixel) of stereo matching algorithms. When the depth image is smaller than the color image, namely, the sampling factor $s > 1$, the warped depth image has sparse depth values only. Additionally, due to occlusion, some background points, which should be removed as occluded points, might be mixed with foreground points whose depth values are smaller than that of background points. In order to remove these occluded depth values, we used a smoothness constraint of 3-D surface [24]. If the depth value of the pixel in the warped depth image is larger than those of neighboring pixels, the pixel is considered to be occluded and removed from the warped depth image.

B. Depth Upsampling

We have implemented the proposed method and evaluated the performance by using the depth video, provided by ToF depth sensor "Mesa Imaging SR4000." For the objective evaluation, we first performed experiments with ground truth depth maps provided by the Middlebury test bed [2], i.e., "Tsukuba," "Venus," "Teddy," and "Cone." They provide stereo image pairs and corresponding depth maps. Low-quality depth map, which is generated by downsampling the ground truth depth map, is upsampled by the proposed method.

The proposed method is tested with the same parameters for all images. Weighting parameters σ_I and σ_S in (4) are set to 6 and 7, respectively. σ_r is determined by bandwidth B of Gaussian function G_r. We compute weighting parameter σ_r, which meets condition $G_r(B/2) = 0.3$. Then, σ_r can be expressed as $\sqrt{-0.5(B/2)^2/ln0.3} = B/3.1$. When the number of depth candidates is 256 and the input depth map is noise-free, bandwidth B is set to 9. In the case of a noisy depth map, bandwidth B increases for taking the noise into account in the input depth map. In this paper, we set a maximum value of bandwidth B to 39. Note that bandwidth B plays a similar role as threshold σ of the 3-D JBU in (12).

The size of window S_W on the original small depth domain is 2. Although the size of window $N(p)$ changes on each level when the MCM is used, the number of the neighboring depth values, which are used for computing joint histogram $H(p, d)$, is approximately similar on all levels. In Fig. 4, when scale l is 2, the depth value is computed for every $4(= 2^2)$ pixels so that actual window $N(p)$ is $-2^2S_W \sim 2^2S_W$ on x and y axes.

Fig. 11 shows the results of the proposed depth upsampling method for the test bed images, when the downsampling ratio is 8 in each dimension. The results of the 2-D JBU [10] and the 3-D JBU [11] were included for a visual evaluation. Note that these methods used the bilinear interpolation technique for computing the initial input (dense) depth maps. The size of window $N(p)$ is set to 11×11 since the MCM is not used. In order to fairly compare the performance of the filtering-based methods by using the same input depth maps, the results that were upsampled with the MCM were included as well, i.e., 2-D JBU + MCM and 3-D JBU + MCM. Note that the 2-D JBU (3-D JBU) uses the interpolated (dense) depth map as an input value, whereas the 2-D JBU (3-D JBU) + MCM uses the original (sparse) depth map.

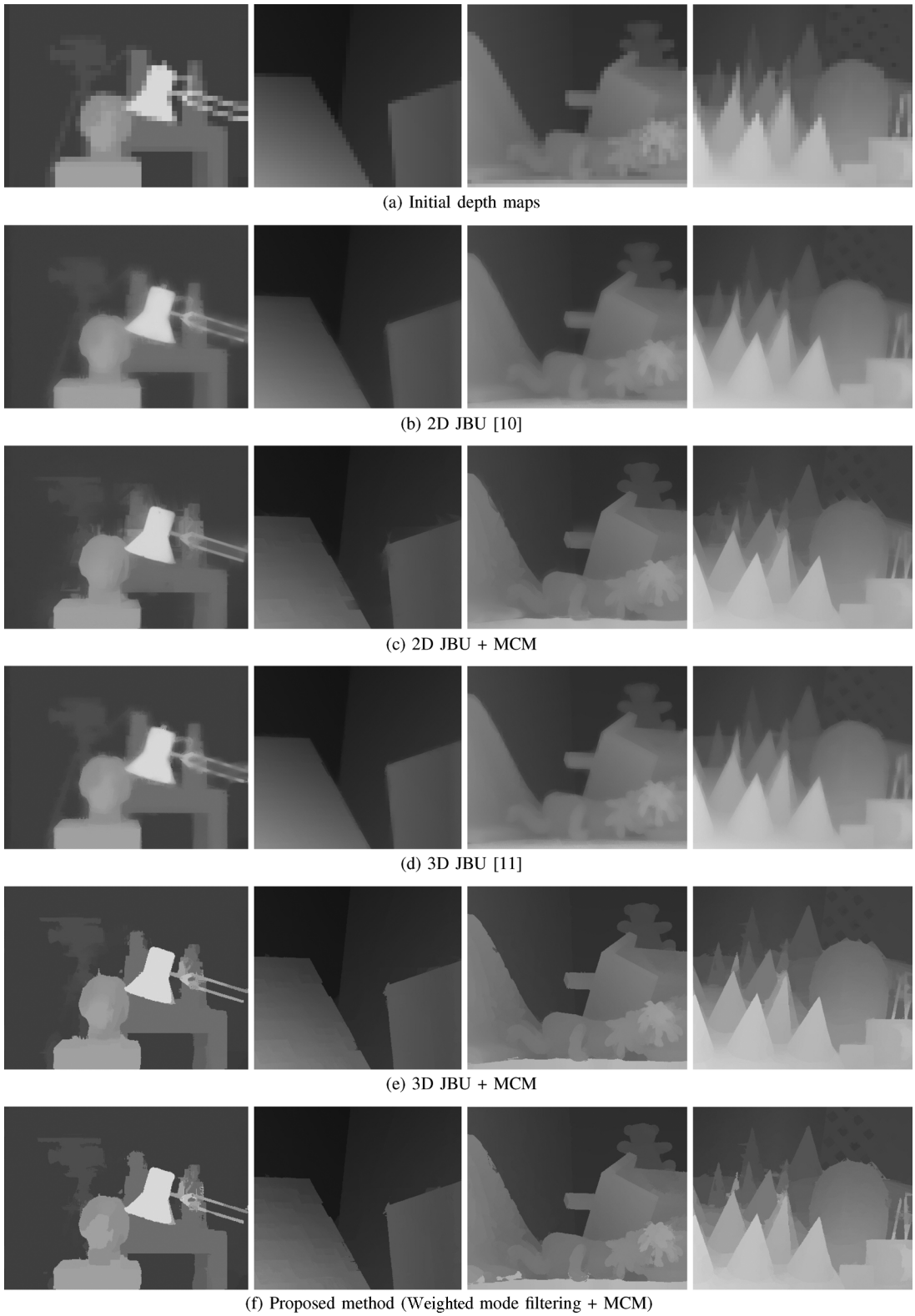
The proposed method yields the superior results over the existing methods, particularly on the depth discontinuities. The performance of the 3-D JBU + MCM is very similar to that of the WMF since the MCM prevented the upsampled depth map from being blurred on the depth discontinuities. The 2-D JBU + MCM does not improve the edge-preserving performance, even compared with the 2-D JBU that used the blurred depth maps as the initial value. We can hence conclude that the 2-D JBU or the 2-D JBU + MCM results in the blurred depth map due to its summation over all depth candidates, even though they use the adaptive weight based on color information and the MCM on the hierarchical scheme. The objective evaluation of these methods is shown in Table II. The accuracy is evaluated by measuring the percent (%) of bad matching pixels (where the absolute disparity error is greater than 1 pixel). The measurement is computed for two subsets of a depth map, i.e., all (all pixels in the image) and disk (the visible pixels near the occluded regions). As expected, the proposed method (WMF + MCM) is superior to the existing methods (2-D JBU and 3-D JBU) and comparable to the 3-D JBU + MCM.

The processing time of the methods is presented in Table III. The processing time of the 2-D JBU + MCM is the smallest among all methods, but its quality is the worst. Although the 3-D JBU + MCM has a similar accuracy to the proposed method, the computational complexity of the proposed method is nearly 0.8% of that of the 3-D JBU + MCM. Since the 3-D JBU performs the joint bilateral filtering for all depth candidates repeatedly, it results in huge computational complexity.

The results for noisy depth maps are also shown in Fig. 12. Additive white Gaussian noise (AWGN) was added with a mean of 0 and a standard deviation of 20 to the low-resolution depth maps, whose downsampling ratio is 8 in each dimension. Bandwidth B of G_r in (4) is set to 39. We found that the proposed method provides the accurate high-resolution depth map, even when the input depth map is very noisy.

C. Depth Refinement

Next, the depth enhancement method is evaluated by applying it to refining depth maps, which were estimated by several existing stereo matching algorithms [2]. As previously mentioned, the MCM is not used. Namely, the depth map is refined with a nonhierarchical scheme. All the parameters are the same to



(a) Initial depth maps

(b) 2D JBU [10]

(c) 2D JBU + MCM

(d) 3D JBU [11]

(e) 3D JBU + MCM

(f) Proposed method (Weighted mode filtering + MCM)

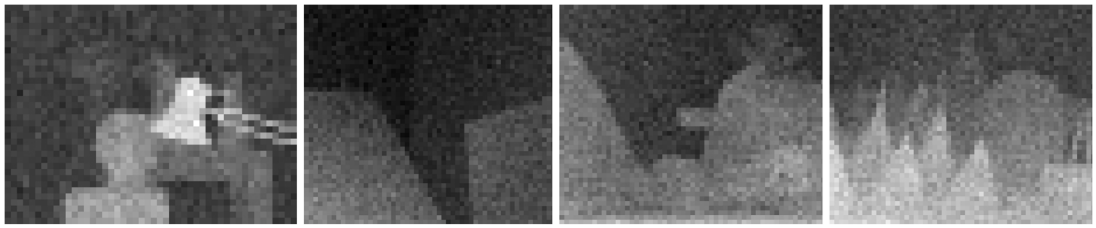
Fig. 11. Depth upsampling results for test bed images: (a) Initial low-resolution depth maps whose downsamping ratio is 8 in each dimension, (b) 2-D JBU results [10], (c) 2-D JBU + MCM, (d) 3-D JBU [11], (e) 3-D JBU + MCM, and (f) proposed method (WMF + MCM).

those of the depth upsampling, except that window size $N(p)$ is set to 7×7.

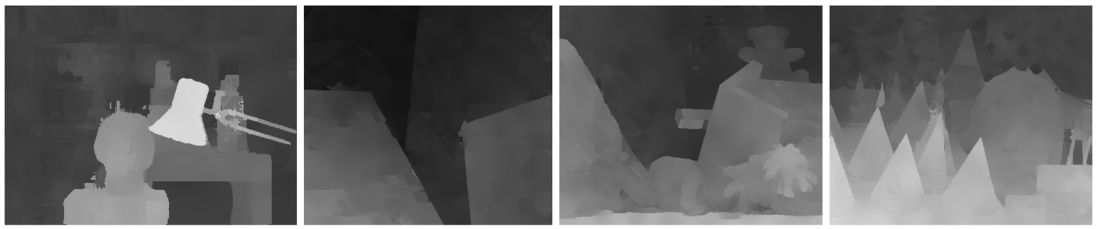
Fig. 13 shows the original depth maps and the enhanced ones, which were calculated by the stereo matching method [2]. The results for "GC+occ" algorithm are shown here. We could find that the proposed filtering improves the accuracy of the depth maps for the discontinuities and occluded regions. Table IV shows an objective evaluation for the depth refinement by mea-

TABLE II
OBJECTIVE EVALUATION (THE PERCENT (%) OF BAD MATCHING PIXELS) FOR DEPTH UPSAMPLING ON ALL (ALL PIXELS IN THE IMAGE) AND DISK (THE VISIBLE PIXELS NEAR THE OCCLUDED REGIONS) REGIONS WITH THE MIDDLEBURY TEST BED

Algorithm	*Tsukuba*		*Venus*		*Teddy*		*Cone*	
	all	disc	all	disc	all	disc	all	disc
Input depth (bilinear)	10.4	46.2	3.26	36.6	11.9	35.5	14.7	36.4
2D JBU [10]	9.04	40.4	2.04	22.1	14.0	37.6	14.7	34.8
2D JBU + MCM	9.71	38.0	2.01	16.3	16.1	38.3	16.7	33.9
3D JBU [11]	7.89	35.0	1.67	17.8	10.7	30.6	12.1	30.0
3D JBU + MCM	4.46	20.4	0.66	5.83	8.88	23.7	10.6	22.6
Proposed method	4.35	20.2	0.61	5.73	9.51	23.7	9.43	19.2



(a)



(b)

Fig. 12. Depth upsampling results in noisy environment: The downsampling ratio is 8 in each dimension, and AWGN was added with a mean of 0 and a standard deviation of 20. (a) Input noisy depth maps. (b) Upsampled depth maps.

TABLE III
PROCESSING TIMES OF DEPTH UPSAMPLING FOR MIDDLBURY TEST BED

Algorithms	Tsukuba	Venus	Teddy	Cone
2D JBU [10]	0.61s	0.94s	0.95s	0.95s
2D JBU + MCM	0.23s	0.33s	0.34s	0.34s
3D JBU [11]	143.0s	215.5s	220.6s	219.2s
3D JBU + MCM	43.3s	65.2s	65.8s	65.6s
Proposed method	0.36s	0.54s	0.55s	0.55s

TABLE IV
OBJECTIVE EVALUATION (AVERAGE RANK OF DEPTH MAPS) OF SEVERAL EXISTING STEREO MATCHING ALGORITHMS "BEFORE" AND "AFTER" APPLYING OUR PROPOSED DEPTH REFINEMENT TECHNIQUE

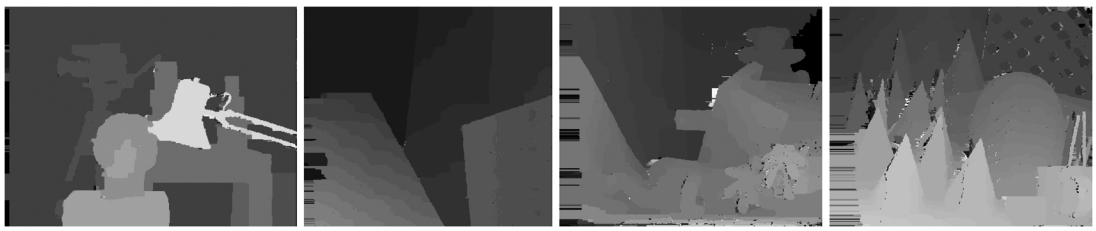
Algorithm	Average Rank	
	Before	After
AdaptingBP	5.2	4.8
CoopRegion	5.2	5.2
DoubleBP	7.3	5.8
OutlierConf	8.2	6.2
SubPixDoubleBP	10.8	8.2
SurfaceStereo	11.8	9.8
WarpMat	12.8	11.3
Undr+OvrSeg	17.3	13.6
GC+SegmBorder	18.1	18.6
AdaptOvrSegBP	19.6	16.5
SymBP+occ	21.7	17.4
MultiResGC	24.7	18.9
RealTimeLAW	36.3	26.8
RealTimeABW	42.1	35.6
RealtimeBP	45.6	36.6
GC+occ	51.2	42.5

suring an average rank of the depth maps. The proposed method improves the accuracy of the depth maps for almost all the algorithms, even for the top-ranking algorithms such as "AdaptingBP" or "DoubleBP" [2]. In "GC+SegmBorder," the output result is a little worse $(18.1 \rightarrow 18.6)$. The proposed method is likely to provide a piecewise constant value so that the output depth map was slightly degenerated at the piecewise linear regions of "Teddy" image. The processing times of the depth enhancement are 0.42 s for "Tsukuba," 0.64 s for "Venus," 0.66 s for "Teddy," and 0.66 s for "Cone."

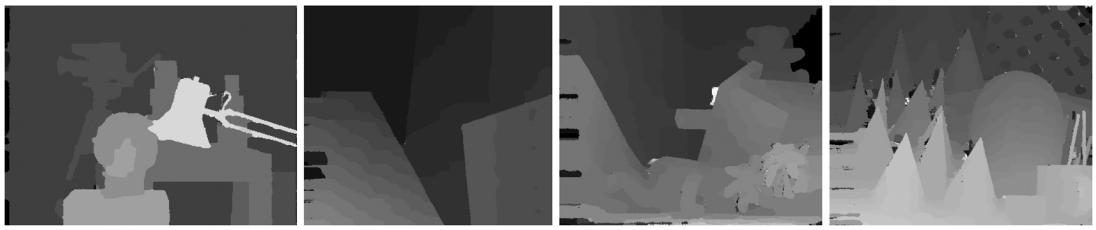
D. Temporal Consistency

In order to evaluate the temporally consistent estimate performance of the proposed method, we performed experiments with color and ground truth depth videos "Tanks" (400 × 300), provided by [28]. The color and depth videos can be downloaded at [29]. The ground truth depth video is downsampled by a factor of 4, and AWGN was then added with a mean of 0 and a standard deviation of 10. The number of the set of neighboring frames $T(t)$ in (14) is set to 2. Namely, $(t-1)$th and $(t+1)$th frames are used for calculating $H^t_{GT}(t)$. Fig. 14 shows upsampled results given by the proposed method. As shown in Fig. 14(f) and (g), we can cope with problems caused by the error of the estimated optical flow on the depth discontinuities. The input and output depth videos are available at [30].

For objective evaluation, we measured the percent (%) of bad matching pixels with the ground truth depth video in Fig. 15.



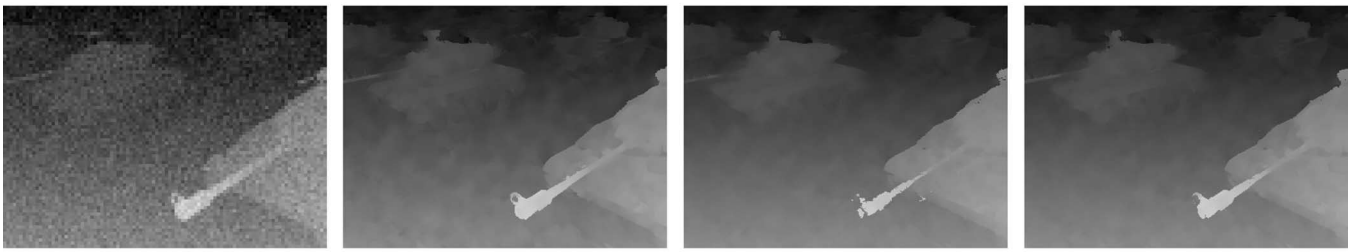
(a) Input depth maps of 'GC+occ' [2]



(b) Enhanced depth maps

Fig. 13. Depth refinement results for depth maps of stereo matching algorithm "GC+occ" [2]. (a) Input depth maps of "GC+occ." (b) Output depth maps enhanced by the WMF. The processing times for each depth map are $0.42s$, $0.64s$, $0.66s$, $0.66s$ Note that the MCM was not used for the depth refinement, so that the processing time is different from that of the depth upsampling (WMF + MCM). (a) Input depth maps of "GC+occ" [2]. (b) Enhanced depth maps.

(a) Color image (b) Ground truth depth (c) Depth discontinuities region



(d) Low-quality depth (e) Without temporal (f) With temporal/without patch sim (g) With temporal/with patch sim

Fig. 14. Temporal consistency on the depth video: (c) is used for measuring an error rate on the depth discontinuities region in Fig. 15(b). (d) Low-quality depth is obtained by applying the downsampling of a factor 4 and adding AWGN with a mean of 0 and a standard deviation 10. (e) We can find that there are some flickering on the background regions. (For a better visualization, please refer to depth videos at [30].) (f) shows the erroneous result on the depth discontinuities, which is caused by the estimated optical flow. (g) By using the patch-based similarity measure, we can handle this problem while maintaining the temporal consistency. (a) Color image. (b) Ground truth depth. (c) Depth discontinuities region. (d) Low-quality depth. (e) Without temporal. (f) With temporal/without patch sim. (g) With temporal/with patch sim.

The experiment was performed with 50 frames. The temporal consistency was not enforced at the first frame so that the results for the first frame were all same. We found that the accuracy of the temporally consistent estimate ("with temporal consistency") is superior to that of the depth upsampling on the single frame ("without temporal consistency"). One interesting observation in Fig. 15(a) is that the method that does not use the patch similarity measure ("without patch sim") has better performance than the method, which uses it ("with patch sim"). While the "with patch sim" method calculates weight w by measuring the patch similarity between two corresponding pixels on the neighboring frames with a fixed-size patch, the "without patch sim" method uses a constant weight, namely, $w_{t-1} = 0.25$, $w_t = 0.5$, and $w_{t+1} = 0.25$. Therefore, weight w on "with patch sim" may be sometimes smaller than a constant value 0.25 on the background regions, and it may result in less temporal smoothing in (14). However, a use of patch similarity measure can reduce the error of the temporal smoothing, which is caused by the erroneous optical flow on the depth boundaries, as shown in Fig. 14(f) and (g). Based on the assumption that the depth discontinuities has more important information than the background regions, the adaptive weight based on the patch similarity measure is used for the temporally consistent depth upsampling. We can also find its effect in Fig. 15(b), which shows the error rate calculated at the depth discontinuities. The example of the depth discontinuities regions is in Fig. 14(c). The "with patch sim" method provides more accurate results on the depth discontinuities. The average processing times are

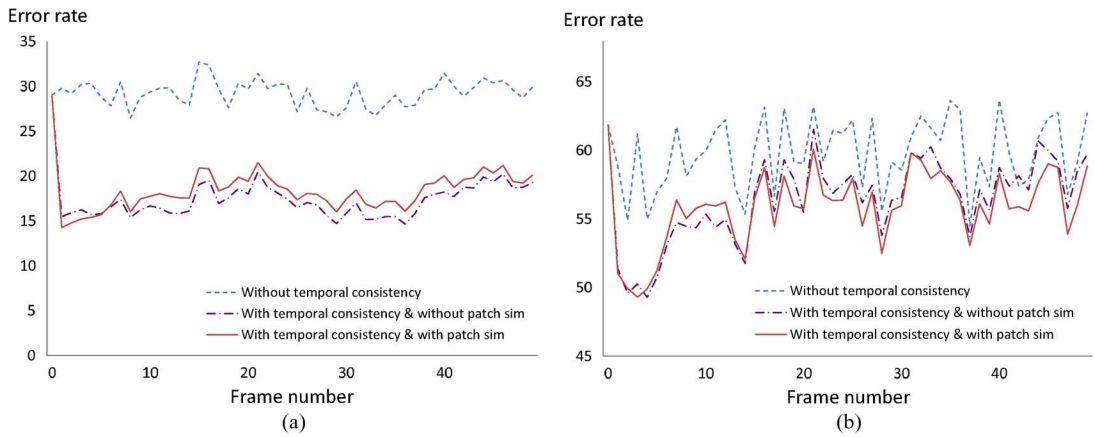


Fig. 15. Percent (%) of bad matching pixels on depth video upsampling. (a) Error rate for all pixels. (b) Error rate for pixels on the depth discontinuities.

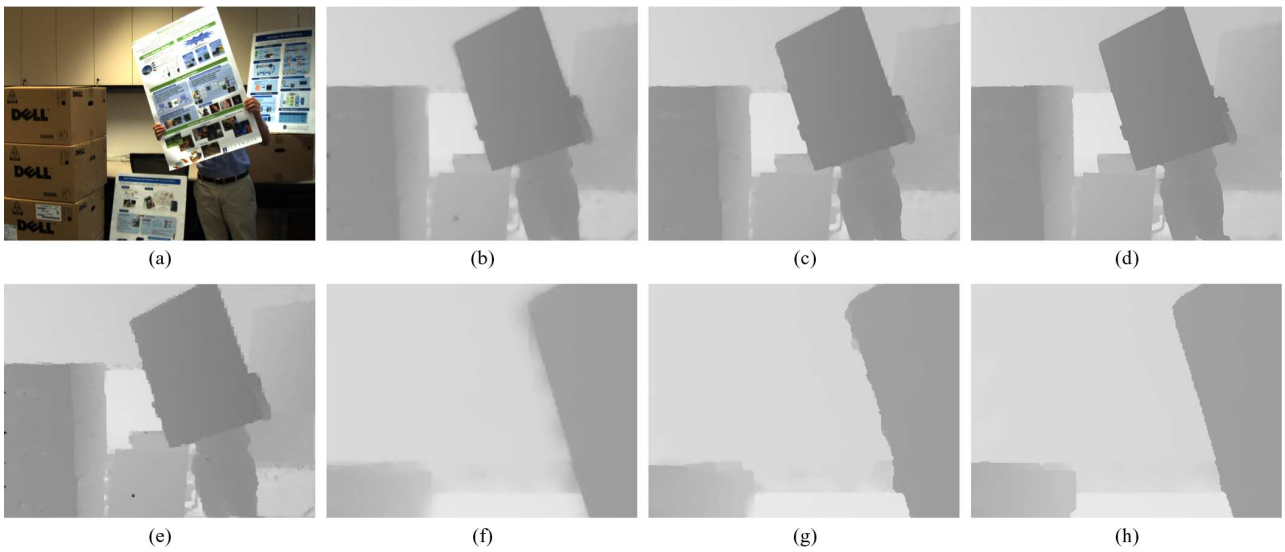


Fig. 16. Upsampling results for low-quality depth image (from "Mesa Imaging SR4000") with corresponding color image (from "Point Grey Flea"). The sizes of the input depth and color images are 176 × 144 and 1024 × 768, respectively. The depth maps, acquired by the depth sensor, were normalized between 0 and 255. (a) Color image. (b) Two-dimensional JBU. (c) Three-dimensional JBU. (d) Proposed method. (e) Initial depth map. (f) Cropped image of (b). (g) Cropped image of (c). (h) Cropped image of (d).

Fig. 14(e) 0.60, (f) 0.91, and (g) 0.98 s, respectively. In other words, an additional processing time for the temporally consistent estimate of "Tanks" depth video is about$0.38\,\mathrm{s}(=0.98\,\mathrm{s}-0.60\,\mathrm{s})$, which consists of 0.31 s for the optical flow estimation and 0.07 s for the patch similarity measure.

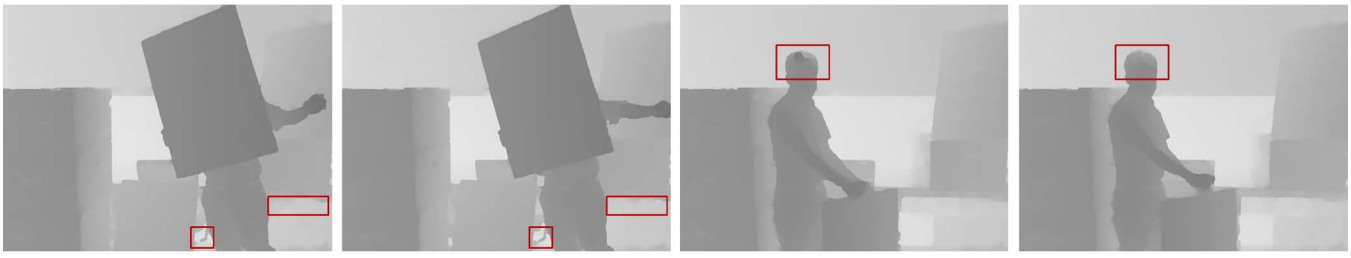
E. Experiments Using ToF Depth Camera

The experiments were also performed using depth and color videos, captured by the color camera and the depth sensor in Fig. 10. As shown in Fig. 16, the proposed method was evaluated by comparing the upsampled depth images with those of the 2-D JBU [10] and the 3-D JBU [11]. The sizes of the input depth and color images are 176 × 144 and 1024 × 768, respectively. The input depth map was normalized between 0 and 255. The temporal consistency scheme was not used in order to evaluate the performance of the upsampling methods only. We found that the proposed method provides the best edge-preserving performance on the depth discontinuities. The processing times are 6.8 s for the 2-D JBU, 1592.3 s for the 3-D JBU, and 5.6 s for the proposed method.

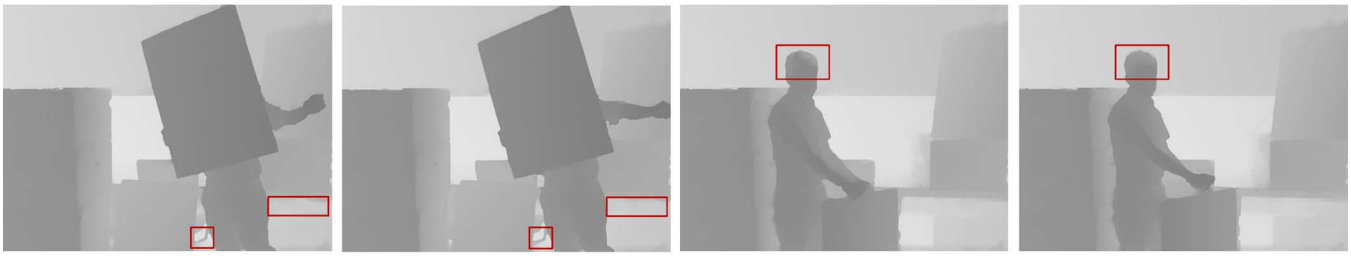
Fig. 17 shows the temporally consistent upsampled depth sequences. The sizes of the input depth and color images are the same to those of Fig. 16. The results of each row are upsampled depth maps of 107th, 111th, 307th, and 311th frames. The number of the set of neighboring frames $T(t)$ is set to 2. As shown in Fig. 17(a), there are some flickering on the background and head regions (particularly, inside red boxes) due to the noisy input depth data. The proposed method generates the temporally consistent high-quality depth maps by employing the estimated optical flow and the patch-based similarity measure. For a better visualization, please refer to depth videos at [30].

VI. Conclusion

In this paper, we have presented a novel approach for providing high-quality depth video in a system that consists of a color and a depth camera. First, the low-quality depth maps, which are of low-resolution and noisy, are upsampled by the



(a) Without temporal consistency



(b) With temporal consistency/with patch sim

Fig. 17. Temporal consistency on depth video: Results for (from left to right) the 107th, 111th, 307th, and 311th frames. The sizes of the input depth and color images are the same to those of Fig. 16. (a) There are some flickering on the background and head regions (particularly, inside red boxes) due to the noisy input depth data. (b) The proposed method generates the temporally consistent high-quality depth maps. (a) Without temporal consistency. (b) With temporal consistency/with patch sim.

proposed WMF method. It provides the results that has better edge-preserving performance. The MCM was also proposed for suppressing the aliasing effect on the depth upsampling. Next, the proposed method was extended into the depth video for obtaining temporally consistent and improved results. The temporally neighboring pixels are estimated by the simple optical flow estimation, and the temporal smoothing is adaptively performed by using the patch similarity measure. The experimental results show that the performance of the proposed method is superior to the existing methods. Since the computational complexity does not depend on the number of depth candidates, the proposed method is very efficient. In further research, we will implement the proposed method with GPUs for a real-time performance. The proposed method is a noniterative scheme so that it is easy to implement on GPUs. Moreover, we will develop a hybrid system that provides more reliable and accurate results by combining the proposed method with stereo matching algorithms.

REFERENCES

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, Apr.–Jun. 2002.

[2] [Online]. Available: http://vision.middlebury.edu/stereo

[3] K. Yoon and I. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.

[4] D. 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.

[5] R. Larsen, E. Barth, and A. Kolb, "Special issue on time-of-flight camera based computer vision," *Comput. Vis. Image Understand.*, vol. 114, no. 12, p. 1317, Dec. 2010.

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

[7] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proc. NIPS*, 2005, pp. 291–298.

[8] J. Lu, D. Min, R. S. Pahwa, and M. N. Do, "A revisit to MRF-based depth map super-resoultion and enhancement," in *Proc. IEEE ICASSP*, 2011, pp. 985–988.

[9] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon, "High quality depth map upsampling for 3D-TOF cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011.

[10] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," in *Proc. ACM SIGGRAPH*, 2007, p. 96.

[11] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," *Proc. IEEE Comput. Vis. Pattern Recognit.*, pp. 1–8, 2007.

[12] Q. Yang, K.-H. Tan, B. Culbertson, and J. Apostolopoulos, "Fusion of active and passive sensors for Fast 3-D capture," in *Proc. IEEE Int. Workshop MMSP*, 2010, pp. 69–74.

[13] B. Huhle, T. Schairer, P. Jenke, and W. Straer, "Fusion of range and color images for denoising and resolution enhancement with a non-local filter," *Comput. Vis. Image Understand.*, vol. 114, no. 12, pp. 1336–1345, Dec. 2010.

[14] M. F. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters," in *Proc. IEEE ICCV*, 2003, pp. 900–906.

[15] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," *Proc. IEEE Comput. Vis. Pattern Recognit.*, pp. 1–8, 2008.

[16] 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.

[17] J. Weijer and R. Boomgaard, "Local mode filtering," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2001, pp. II-428–II-433.

[18] S. Paris, P. Kornprobst, J. Tumblin, and F. Durand, "Bilateral filtering: Theory and applications," *Foundations Trends Comput. Graph. Vis.*, vol. 4, no. 1, pp. 1–73, 2008.

[19] S. Gokturk, H. Yalcin, and C. Bamji, "A time-of-flight depth sensor—System description, issues and solutions," in *Proc. IEEE CVPRW*, 2004, p. 35.

[20] D. Barash, "A fundamental relationship between bilateral filtering, adaptive smoothing, and the nonlinear diffusion equation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 6, pp. 844–847, Jun. 2006.

[21] M. Elad, "On the origin of the bilateral filter and ways to improve it," *IEEE Trans. Image Process.*, vol. 11, no. 10, pp. 1141–1151, Oct. 2002.

[22] J. Lu, G. Lafruit, and F. Catthoor, "Anisotropic local high-confidence voting for accurate stereo correspondence," in *Proc. SPIE-IST Electron. Imag.*, Jan. 2008, vol. 6812, p. 68 120J.

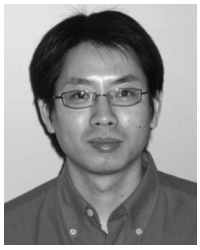
[23] A. Buades, B. Coll, and J.-M. Morel, "Nonlocal image and movie denoising," *Int. J. Comput. Vis.*, vol. 76, no. 2, pp. 123–139, Feb. 2008.

[24] M. N. Do, Q. H. Nguyen, H. T. Nguyen, D. Kubacki, and S. J. Patel, "Immersive visual communication," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 58–66, Jan. 2011.

[25] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intell.*, 1981, pp. 674–679.
[26] C. Tomasi and T. Kanade, Detection and tracking of point features Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-91-132, Apr. 1991.
[27] [Online]. Available: http://vision.middlebury.edu/flow
[28] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *Proc. ECCV*, 2010, pp. 510–523.
[29] [Online]. Available: http://www.cl.cam.ac.uk/research/rainbow/projects/dcbgrid/datasets/
[30] [Online]. Available: http://diml.yonsei.ac.kr/~forevertin/
[31] [Online]. Available: http://www.mesa-imaging.ch/
[32] [Online]. Available: http://www.ptgrey.com/

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.

From 2009 to 2010, he worked with Mitsubishi Electric Research Laboratories as a Postdoctoral Researcher, where he developed a prototype of 3-D video system (3DTV). Since July 2010, he has been working with Advanced Digital Sciences Center, Singapore, which was jointly founded by University of Illinois at Urbana-Champaign, Urbana, and the Agency for Science, Technology and Research (A*STAR), which is a Singaporean government agency. His research interests include 3-D computer vision, graphics-processing-unit-based real-time system, 3-D modeling, hybrid sensor system, and computational photography.



Jiangbo Lu (M'09) received the B.S. and M.S. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering, Katholieke Universiteit Leuven, Leuven, Belgium, in 2009.

From April 2003 to August 2004, he was with VIA-S3 Graphics, Shanghai, China, as a Graphics Processing Unit (GPU) Architecture Design Engineer. In 2002 and 2005, he conducted visiting research at Microsoft Research Asia, Beijing, China. Since October 2004, he has been with the Multimedia Group, Interuniversity Microelectronics Center, Leuven, Belgium, as a Ph.D. Researcher, where he pioneered and led research on real-time stereo matching and view synthesis. Since September 2009, he has been working with the Advanced Digital Sciences Center, Singapore, which is a joint research center between the University of Illinois at Urbana-Champaign, Urbana, and the Agency for Science, Technology and Research (A*STAR), Singapore, where he is currently a Research Scientist. His research interests include multimedia signal processing, computer vision, visual computing, video coding, interactive multimedia applications and systems, GPU-based computing, and embedded computer vision.

Minh N. Do (M'01–SM'07) was born in Vietnam in 1974. He received the B.Eng. degree in computer engineering from the University of Canberra, Canberra, Australia, in 1997 and the Dr.Sci. degree in communication systems from the Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland, in 2001.

Since 2002, he has been with the faculty of the University of Illinois at Urbana-Champaign UIUC, Urbana, where he is currently an Associate Professor with the Department of Electrical and Computer Engineering and holds joint appointments with the Coordinated Science Laboratory, Beckman Institute for Advanced Science and Technology, and the Department of Bioengineering. His research interests include image and multidimensional signal processing, wavelets and multiscale geometric analysis, computational imaging, augmented reality, and visual information representation.

Prof. Do is a member of the IEEE Signal Processing Theory and Methods and Image, Video, and Multidimensional Signal Processing Technical Committees. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He was the recipient of the Silver Medal in the 32nd International Mathematical Olympiad in 1991, a University Medal from the University of Canberra in 1997, a Doctorate Award from the EPFL in 2001, a CAREER Award from the National Science Foundation in 2003, the Xerox Award for Faculty Research from the College of Engineering, UIUC, in 2007, and a Young Author Best Paper Award from IEEE in 2008. He was named a Beckman Fellow in the Center for Advanced Study, UIUC, in 2006.