Depth Map Refinement Using Reliability Based Joint Trilateral Filter

Takuya Matsuo[1], **Naoki Kodera**[2],
Norishige Fukushima[3], and **Yutaka Ishibashi**[4], Non-members

ABSTRACT

In this paper, we propose a refinement filter for depth maps. The filter convolutes an image and a depth map with a cross computed kernel. We call the filter joint trilateral filter. Main advantages of the proposed method are that the filter fits outlines of objects in the depth map to silhouettes in the image, and the filter reduces Gaussian noise in other areas. The effects reduce rendering artifacts when a free viewpoint image is generated by point cloud rendering and depth image based rendering techniques. Additionally, their computational cost is independent of depth ranges. Thus we can obtain accurate depth maps with the lower cost than the conventional approaches, which require Markov random field based optimization methods. Experimental results show that the accuracy of the depth map in edge areas goes up and its running time decreases. In addition, the filter improves the accuracy of edges in the depth map from Kinect sensor. As results, the quality of the rendering image is improved.

Keywords: Joint Trilateral Filteringn, Stereo Matching, Depth Mapn, Refinement Filter, Post FilteringJoint Trilateral Filtering, Stereo Matching, Depth Map, Refinement Filter, Post Filtering

1. INTRODUCTION

Recently, consumer-level depth sensors, e.g. Microsoft Kinect [1] and ASUS Xtion [2], are released, and then image processing methods for depth maps attract attentions. For example, pose estimation, object detection, point cloud, and free viewpoint video rendering are presented. Especially, the free viewpoint image rendering requires high quality depth maps. The free viewpoint images are synthesized by the depth image based rendering (DIBR) [3] that demands input images and depth maps.

Depth maps are usually computed by stereo matching methods. The stereo matching finds corresponding pixels between left and right viewpoint images. The depth value is calculated by the correspondence. The stereo matching is constructed from four steps that are matching cost computation, cost aggregation, depth map computation/optimization and depth map refinement [4]. The mainstream methods of stereo matching perform complex optimizations to improve the accuracy of depth map. The stereo matching with optimization methods based on Markov random field, e.g. the semi-global block matching [5], the belief propagation [6], and the graph cuts [7], generate accurate depth maps. While the complex optimization algorithms increase their computation time. In addition, the strong constrains of the smoothness consistency over the optimizations obscure local edges of the depth maps.

If we render a novel image by the depth image based rendering with the ambiguous depth maps, edges of objects in the composite image will be not accurate. Thus, it is important for the free viewpoint image synthesis to use depth maps which are accurate on the object edge. Therefore, we propose a refinement filter for depth maps. The filter enhances the accuracy of the depth maps, especially object boundaries, while their computational cost keeps low.

We organize the remainder of the paper as follows. Section 2 presents an overview of related works of this paper. Section 3 introduces the conventional refinement filter and proposes a novel refinement filter of depth maps. Section 4 shows the experimental results. Finally, we conclude this paper in Section 5.

2. RELATED WORKS

Generally speaking, depth maps are noisy. Thus the depth maps are often filtered by noise reduction filters. The bilateral filter [8] is one of the candidates, which can reduce noises with keeping edge shapes. However, the performance of edge keeping and noise reduction depends on the image conditions before filtering. When the image is so noisy, the performance of the noise reduction becomes down dramatically. In addition, only Gaussian noise can be removed by the filter, although depth map contains spike and non-Gaussian noises.

overcomes this problem in a special condition. The condition is that we can use two images which are captured at the same viewpoint but have different image characteristics. References [9, 10] use a non-flash image and a flash image as an input pair. The flash of camera reduces image noise but changes lighting

Manuscript received on March 14, 2012 ; revised on April 23, 2012.

[1,2,3,4] The authors are with Nagoya Institute of Technology Nagoya, Japan. E-mail: t-matsuo@mcl.nitech.ac.jp, n-kodera@mcl.nitech.ac.jp, fukushima@nitech.ac.jp and ishibasi@nitech.ac.jp

conditions, e.g. scene lighted by candlelight. The non-flash image keeps light conditions but contains large noise. To combine both the pros, these papers use non-flash images as a filtering target, and a flash image as a filtering kernel computation target. As the filtering result, the output image keeps lighting conditions without large noise. A key point of the filtering technique is as follows. It is effective to compute the filtering kernel by noiseless information instead of noisy filtering target images.

The knowledge of the joint bilateral filtering is applied to depth map processing. References [11-13] propose depth up sampling and super resolution methods based on the joint bilateral kernel computation. It is effective for depth maps from depth sensors because the resolution of the depth map tends to be low.

Other applications are stereo matching improvement methods [14, 15] and refinement methods of depth maps [16, 17]. References [14, 15] apply the joint bilateral filtering to depth estimation. The stereo cost volume which indicates probabilities of depth states is filtered by the joint bilateral filter. The filter uses an input natural view for the kernel computation, and then the accuracy of estimated depth maps is improved.

The refinement methods of the joint filtering are proposed in [16, 17]. These papers use depth maps as filtering targets and stereo image pairs as kernel computation targets. The filter computes filtering kernel by pixel color information and additional pixel reliability information. The reliability is computed by a L-R consistency check method [4]. The checking assumes that projected depth information from a left depth map and a projected right one should have the same value. If the left and right depth value is inconsistent, the pixels in the region are regarded as unreliable, and then the reliability becomes low. These filtering improve the quality of depth maps brilliantly, however, is not suitable for depth maps from depth sensors and for real-time computations.

It is because that these methods [16, 17] require left and right depth maps. When we use a depth sensor, we can obtain only one depth map. In addition, these joint bilateral filtering based methods [11-17] require iteration processes whose conversion time depends on ranges of depth values. Generally speaking, the ranges of the depth maps from depth sensors tend to be higher than the depth map from the stereo matching methods. For example, Kinect can capture the depth map with 11-16 bits. Thus computational costs of the conventional methods of the joint bilateral refinement tend to be high.

To overcome the weak point, we propose a filtering method which requires only one depth map and one or two views without iteration processes. We call the filter reliability based joint trilateral filter. The proposed filter is designed to refine depth maps well within one pass processing. There are two key-points in the method; one is finding reliable pixels and filtering with the reliability as weighs, and the other is a post-processing for boundary regions, where image tend to be blurred by the joint bilateral based filtering, to recover and remove that. Main differences from the conventional approaches are;

1. The proposed refinement filter does not require iteration processes and does not require left and right depth maps, only requires one depth map.
2. The proposing post-processing, which rejects ramp edges and interpolates it, improves accuracy of object boundary. The area tends to be blurred by the kind of the joint bilateral filtering. The rejection and interpolation method is also used for undetermined depth areas.

3. DEPTH MAP REFINEMENT

Depth estimation processing has the four chains of which are matching cost computation, cost aggregation, depth map computation/optimization and depth map refinement, and we focus on the depth map refinement.

Firstly, we introduce the traditional bilateral filter and joint bilateral filter in section 3.1. Secondly, we propose a now filter of the reliability based joint trilateral filter in section 3.2. Finally, we propose a post-processing for blurred region and undetermined regions to reject and interpolate it.

3.1 BILATERAL FILTER AND JOINT BILATERAL FILTER

The proposed filter improves depth maps estimated by the block matching which is the fast but not so accurate stereo matching. The filter smooths non-uniform surfaces and corrects edges. We call this filter reliability based joint trilateral filter.

The reliability based joint trilateral filter is an extension of the bilateral filter. The bilateral filter is defined by the following formula in references [8]:

$$O_p = \frac{\sum_{S\in N} w(\boldsymbol{p},\boldsymbol{s})c(\boldsymbol{p},\boldsymbol{s})I_s}{\sum_{S\in N} w(\boldsymbol{p},\boldsymbol{s})c(\boldsymbol{p},\boldsymbol{s})}, \tag{1}$$

where I = input image, O = output image, $\boldsymbol{p}$ = coordinate of attention pixel, $\boldsymbol{s}$ = coordinate of support pixel, N = aggregation set of support pixels, w = location weight function, c = color weight function. Additionally, each weight is Gaussian distribution:

$$\begin{gathered} w(\boldsymbol{p},\boldsymbol{s}) = \exp\left(-\frac{\|\boldsymbol{p}-\boldsymbol{s}\|_2}{2\sigma_s}\right), \\ c(\boldsymbol{p},\boldsymbol{s}) = \exp\left(-\frac{\|I_p-I_s\|_2}{2\sigma_c}\right), \\ (\sigma_s,\sigma_c : const.), \end{gathered} \tag{2}$$

where $\| \cdot \|_2 =$ L2 norm. In this filter, the weight of the support pixel becomes large, when the pixel has a near intensity of the attention pixel and has a near position of the attention one. Striding edge parts have small weights due to large intensity differences. Thus the depth map is smoothed while maintains edge parts.

However, if the input depth map has widely incorrect values around object boundaries, the edges of the object are not corrected by the bilateral filter. Thus, the bilateral filter is extended into the joint bilateral filter in order to refer to the natural image for exacting edge information. We use an input image for the color weight computation instead of a depth map. The joint bilateral filter is defined by the following formula in references [9, 10]:

$$D_p = \frac{\sum_{S \in N} w(\boldsymbol{p}, \boldsymbol{s}) c(\boldsymbol{p}, \boldsymbol{s}) D_s}{\sum_{S \in N} w(\boldsymbol{p}, \boldsymbol{s}) c(\boldsymbol{p}, \boldsymbol{s})} \tag{3}$$

where D_p and D_s are depth value of the attention and the support pixel, respectively. The kernel of the color weight are also computed by input image I ,not computed by D; thus it is possible to remove noisy pixels while keep edge parts of natural image by computed color weight using natural image. However, it is a smoothing filter, ramp edges are occurred by mixed values in edge area.

3.2 RELIABILITY BASED JOINT TRILATERAL FILTER

Wherein, we add reliability information of depth maps as the third weight element to the joint bilateral filter in the proposed joint trilateral filter. The third weight element has an effect of enhancing joint bilateral filter and controlling occurred ramp edges. The reliability in a kernel is mainly calculated by the differences between the depth value of an attention pixel and support pixels. The joint trilateral filter is defined by the follow formula:

$$\begin{gathered} D_p = \frac{\sum_{S \in N} w(\boldsymbol{p}, \boldsymbol{s}) c(\boldsymbol{p}, \boldsymbol{s}) r(\boldsymbol{p}, \boldsymbol{s}) D_s}{\sum_{S \in N} w(\boldsymbol{p}, \boldsymbol{s}) c(\boldsymbol{p}, \boldsymbol{s}) r(\boldsymbol{p}, \boldsymbol{s})} \\ r(\boldsymbol{p}, \boldsymbol{s}) = \exp\left(- \frac{\|\boldsymbol{D_p} - \boldsymbol{D_s}\|_2}{2\sigma_r} \right) \\ (\sigma_r : const.), \end{gathered} \tag{4}$$

If a part has a small depth difference, its reliability is large. However, when depth of the attention pixel suffers from a large noise, it makes a problem to assign the large reliabilities for the support pixels. Additionally, the boundary parts of the depth map are not accurate and tend to be blurred. Therefore, we should to assign the parts as low reliability. Thus, the reliabilities should be adaptively determined according to the following classification approach.

The proposed classification approach requires one depth map on the target view and right and left natural images. One of which must posit on the target view and the other is optional view which does not have to be required. We assume that a depth value of an ideal result and an intensity of natural image are close between an attention pixel and a support pixel in a same object. In addition, we assume that correspondence pixels in left and right views, which are connected by the depth map, have close intensities. Therefore, reliable pixels have the following conditions in the proposed classification approach:

1. Comparing the depth value of the attention pixel D_p^l and the support pixel D_s^l in the left depth map. The difference should be below a threshold α.

$$\|D_p^l - D_s^l\|_1 \leq \alpha. \tag{5}$$

2. Comparing the intensity of the attention pixel I_p^l and the support pixel I_S^l in the left natural image. The difference should be below a threshold β.

$$\|l_p^l - l_s^l\|_1 \leq \beta. \tag{6}$$

3. Comparing the intensity of the support pixel I_S^l of the left natural image and the corresponding pixel $I_(S + D_s^l)^r$ of the natural image of then right viewpoint. The difference should be below a threshold γ.

$$\|l_s^l - l_{S+D_s^l}^r\|_1 \leq \gamma, \tag{7}$$

where $\| \cdot \|_1 =$ L1 norm, and $l =$ left viewpoint, $r =$ right viewpoint.

4. If the above conditions are fulfilled, the reliability function r becomes valid. Otherwise, it set to 0. In other words, the reliability function is re-defined as follows;

$$r(\boldsymbol{p}, \boldsymbol{s}) = \begin{cases} \exp\left(- \frac{\|\boldsymbol{D_p} - \boldsymbol{D_s}\|_2}{2\sigma_r} \right) & (meetcond) \\ 0 & (eles). \end{cases} \tag{8}$$

As a result, depth maps are smoothed while keeping object edges. In addition, the noise of depth maps is not imparted to the reliability as much as possible. Moreover, the filter operates at high speed, because the filter is single pass. The filter consisted of the carefully selected pixels refines well at one shot. Figure 1 shows examples of the kernel weight of the proposed method. The region A and B in the input image are zoomed up, and then we can see that the kernel weights are fitted by the image edges except for unreliable regions.

3.3 POST-PROCESSING FOR OBJECT BOUNDARY

The joint trilateral filtering corrects depth maps around edge boundaries. However, the depth maps are blurred and ramp edges are generated, when depth candidates in a kernel is large. The reliability based joint trilateral filter has the smaller blurred region than the conventional joint bilateral filter, but the blurred region is still remained. Thus we find the ramp edge and enhance the edge sharply. The method of removing the ramp edge is as follows. If a focusing pixel p(x) is a ramp edge part, the relationship between the pixel and neighborhood pixels fills the following conditions:

$$\begin{aligned} \mathrm{abs}(p(x-1)-p(x)) &= 1, \\ \mathrm{abs}(p(x)-p(x+1)) &= 1, \\ \mathrm{abs}(p(x-1)-p(x+1)) &= 2. \end{aligned} \tag{9}$$

If we find the region, we re-label the region as an undetermined region, and then the region is interpolated from the neighborhood regions where are determined. The interpolation method is the joint bilateral interpolation based on the joint bilateral up sampling. Equation of this interpolation method is almost the same as the joint bilateral filter except for a set of support pixel. We can use only determined pixel, thus support pixel set M is a set of valid pixels in the interpolation. An undetermined depth value dp at pixel $\boldsymbol{p}$ is interpolated by the following equation and then we can finally obtain a refined depth map.

$$\begin{aligned} d_p &= \frac{\sum_{s\in N} w(\boldsymbol{p},\boldsymbol{s})c(\boldsymbol{p},\boldsymbol{s})d_s}{\sum_{s\in N} w(\boldsymbol{p},\boldsymbol{s})c(\boldsymbol{p},\boldsymbol{s})}. \\ w(\boldsymbol{p},\boldsymbol{s}) &= \exp\left(-\frac{\|\boldsymbol{p}-\boldsymbol{s}\|_2}{2\sigma_s}\right), \\ c(\boldsymbol{p},\boldsymbol{s}) &= \exp\left(-\frac{\|\boldsymbol{I_p}-\boldsymbol{I_s}\|_2}{2\sigma_c}\right), \\ &(\sigma_s, \sigma_c : \mathit{const.}) \end{aligned} \tag{10}$$

To apply the proposed post-filtering method to the depth map from Kinect sensor instead of the stereo matching, there are two problems. One is that Kinect cannot capture left and right images, and the other is that the depth map from Kinect has many invalid regions where depth values are not obtained. The former is solved by ignoring the reliability assumption of the γ term. The latter is solved by the joint bilateral interpolation of above of this section. Figure 2 shows the invalid regions. The region (A) is the occlusion part of an IR projector and an IR camera, and the region (B) is the warping hole when the depth map is registered to the image position. The region (C) is the saturated area because of sunlight, and the region (D) is the light reflected area. The region (E) is a black object which reduces IR light. All regions are interpolated by the joint bilateral interpolation.

4. EXPERIMENTAL RESULTS

We have two experiments; one is depth estimation experiments and the other is free viewpoint image synthesis experiments. The Middlebury's data sets [4] are used for the stereo evaluation. Data set are Tsukuba (Fig. 3(a)), Venus (Fig. 3(b)), Teddy (Fig. 3(c)) and Cones (Fig. 3(d)). Image resolutions of each data set are 384×288 (Tsukuba), 434×383 (Venus), and 450×375 (Teddy and Cones), respectively. Competitive methods are block matching (BM) as a simple stereo matching, semi-global block matching (SGBM) as an optimized method which has real-time capability, and the BM with the joint trilateral filter (C-Tri). In addition, the bilateral filter (Bi) and the median filter (Med) as the refinement filters for the depth map from the BM are used in order to reveal advantages of the proposed filter. The free viewpoint image synthesis is performed by the depth image based rendering. Depth maps are obtained by the BM, the SGBM and the BM with the joint trilateral filter. The synthesized free viewpoint images are compared with pre-captured images by means of Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) [18]. The SSIM is defined by the following formula in reference [18]:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{11}$$

where μ_x, μ_y are average of x or y, σ_x^2, σ_y^2 are variance of x or y, σ_{xy} is covariance of x and y, and C_1, C_2 are constant values to stabilize the division with weak denominator. We set (C_1, C_2)=(7.0756,58.9824) which are default parameters in reference [18]. In addition, we experiment on depth maps from Kinect. Compared depth map are without and with joint trilateral filter. Then, we compare how much correction is the edge.

The results of the depth estimation are show in Table1 and Fig. 4. The parameters of proposed method $(\sigma_s, \sigma_c, \sigma_r, \alpha, \beta, \gamma)$ are (16.0,61.0,13.4,21,184,1) in Tsukuba dataset, (30.0,16.5,17.5,14,59,1) in Venus dataset, (19.0,14.0,255,20,59,2) in Teddy dataset, (20.0,16.9,24.0,26,75,4) in Cones dataset to maximize the accuracies. And kernel size is (15×15) in all data sets. These parameters are decided by heuristics (as will be described in the next section). The error rate of the joint trilateral filter is better than the BM for all data sets in Table 1. The improvements are 2.45% in Tsukuba, 0.53% in Venus, 0.29% in Teddy and 0.21% in Cones. Especially, the joint trilateral filter is effective as same as the SGBM with Tsukuba data set. However, the accuracy of the proposed method is worse than the one of the SGBM in any data sets. It is because that the proposed method is categorized into post filtering, the type of methods depend on an accuracy of input depth maps. These filters need at least one pixel which has an exact depth value in the

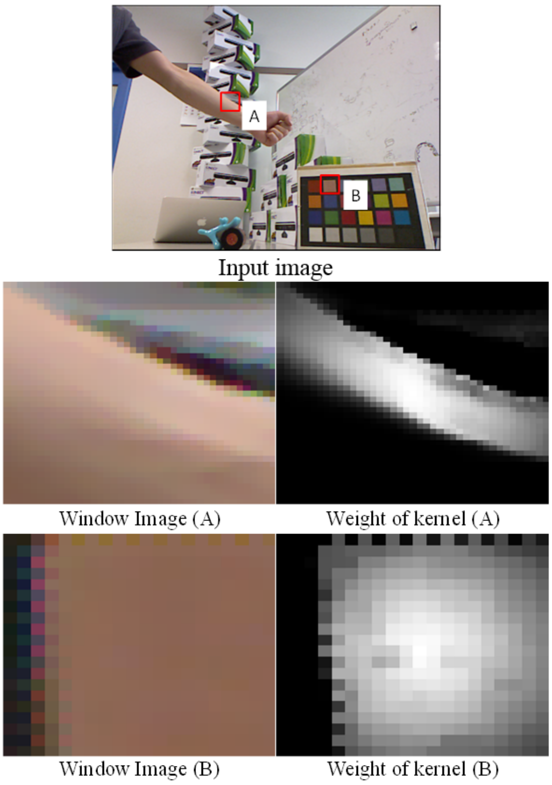


Fig.1: *Visualization of kernel weight; white means large weight and black small weight.*

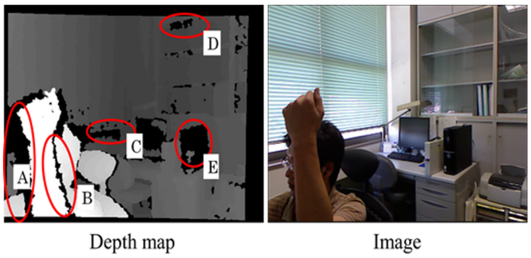


Fig.2: *Depth map from Kinect and relative view: specific invalid depth region are circles.*

filter kernel. If there is no pixel of the exact depth value in the filter kernel it is impossible that the filter refines error pixels. The depth maps from the BM have the lower accuracy in the low textured area than the one from the SGBM. As a result, if the valid range of the filter is small in the depth map from the BM, the BM with the proposed filter has less accuracy than the one of the SGBM.

Here, we define Relative Improvement Rate (RIR) in order to indicate how much the joint trilateral filter is improved from the error rate of the BM. The RIR is defined as:

$$RIR = \frac{E_{BM} - E_{C-Tri}}{E_{BM}}, \qquad (12)$$

where $E_X =$ error rate of method X. The RIR has shown a high value of 42.3% in Tsukuba and 25.6% in Venus (see in Table 2). In addition, the joint trilateral filter is highly effective compared with the bilateral filter and the median filter. Noises have been eliminated and object edges are more accurate than the depth map of the BM in Fig.4. However, RIR has shown the lower value of 4.4% in Teddy and 3.2% in Cones than Tsukuba and Venus. A difference among them is the number of gradations of depth. The number of the gradations of depth is 16 in Tsukuba, 32 in Venus and 64 in Teddy and Cones.

Thus, we have an additional experiment. We convert the depth ranges which are 16 or 32, and use a narrow baseline in Teddy data set. In the Table2, a similar improvement is seen if the number of gradation of depth is similar to Tsukuba and Venus. Here, AD (Absolute Difference) is a difference between error rates from the C-Tri and the BM. It says that this joint trilateral filter is effective when the number of the gradation of depth is low.

(a) tsukuba (b) venus

(c) teddy (d) cones

Fig.3: *Middlebury's stereo data sets.*

Table 1: *Error rate.*

Data Set, No.of gradation	BM with C-Tri	BM	SGBM	Med	Bi
Tsukuba,16	2.76	5.82	3.26	3.99	4.14
Venus,32	1.38	2.07	1.00	1.87	1.85
Teddy,64	5.50	6.55	3.26	6.44	6.33
Cones,64	5.42	6.60	3.02	6.45	6.41

(%)

In Table 3, the following is the result of the running time to get depth maps. The experimental environment is Intel Core i7-920 2.93GHz with Visual Studio 2010 Ultimate. Table 3 is shown in the running time for the BM, the SGBM and the filtering. The unit is milliseconds. As a result, the BM with the joint trilateral filter is faster than the SGBM in any

Table 2: *Improvement rate.*

Data Set, No.of gradation	BM with C-Tri	BM	AD	RIR
Tsukuba,16	2.76	5.82	3.06	52.6
Venus,32	1.38	2.07	0.69	33.3
Teddy,64	5.50	6.55	1.05	16.0
Teddy,32	4.62	7.82	3.20	40.9
Teddy,16	5.46	14.3	8.84	61.8
Cones,32	5.42	6.60	1.18	17.9

(%)

Table 3: *Running Time.*

No.of Gradation	BM	Filter C-Tri	Sum of BM and C-Tri	SGBM
16(Tsukuba)	5.8	13.7	19.5	28.9
32(Venus)	9.9	20.7	30.6	56.9
64(Teddy&Cones)	10.7	22.9	33.6	76.8

(ms)

Table 4: *Running Time.*

Methods	PSNR[dB]	SSIM
SGBM	34.50	0.9695
C-Tri	35.60	0.9701
BM	34.82	0.9667

data sets. It is because that proposed method filters depth map directly and does not have iterating process. In addition, our proposed filter is independent of the number of gradations of the depth map. In contrast, optimization methods like the SGBM depend on that. Our method only depends on image size and kernel size, and the SGBM also depends on these factors. Therefore, the number of gradations of the depth map become higher, the advantage of proposed method in the running time becomes larger than the SGBM (in Table 3).

The results of PSNR and SSIM of the experiments of the free viewpoint image synthesis are shown in Fig. 5 and Table 4. In this experiment, Teddy data set is used. PSNR and SSIM of the free viewpoint image using the depth map with the joint trilateral filter are better than using the depth map of the BM and the SGBM in Table3. The rate of improvement is about?0.78 dB from the BM. As a result of comparing the synthesized images visually, the object edge of the composite image of using the BM is some deficient parts. In contrast, these deficient parts are especially improved in the synthesized image of using the joint trilateral filter (Fig. 3). There are some deficient parts of the object edge of the composite image using the SGBM. PSNR difference between the joint trilateral filter and the SGBM are 1.10 dB. PSNR of the SGBM is 0.32 dB lower than the BM. The results of SSIM show similar tendency for all methods.

This is because that there are important and unimportant regions in a depth map for the free viewpoint image synthesis [19]. For example, a high frequency texture region is important and low one is not. In addition, an object boundary region is important and a region far from a boundary is not. The proposed filter can refine not only pixels on the object boundary but also one on the low frequency texture region while the SGBM smooths low textured region and oversmooths object boundary. Therefore the proposed method overcomes the SGBM in the context of the free viewpoint image synthesis. Figure 6 shows the experimental results from the Kinect depth map. The non-filtering depth map of getting Kinect has rough edges. Thus, the edge of composite image of using it is defectiveness. In contrast, the synthesized image with the filtered depth map has corrected edges. As a result, the edge of the composite image is more corrective then the non-filtering it. Figure 7 shows the depth map from proposed method without ramp erosion and ramp edge detection results. The results shows that, ramp edges tend to be emerged at area where have large depth gaps. After erosion in Fig. 6 (c), ambiguity of ramp edge is removed.

5. DISCUSSION

5.1 RELATIONSHIP AMONG PARAMETERS

Here, we explain the detail of parameters setup. Our proposed filter has seven parameters. These are the variable of Gaussian sigma σ_s of space weight, the variable of Gaussian sigma σ_c of color weight, the variable of Gaussian variable σ_r of reliability weight, the depth value threshold α, the intensity threshold β, the LR-Check threshold γ, and the kernel size. These parameters are able to be classified into four categories. These are space, color, depth, and LR-Check categories. So, the Gaussian variable σ_s and the kernel size are in the category of a space, the Gaussian variable σ_c and the threshold β are in the category of a color, the Gaussian variable σ_r and the threshold α are in the category of a depth, and the LR-Check threshold γ is in the category of LR-Check.

Except for the LR-Check threshold, space, color and depth categories' parameters shape truncated Gaussian distribution;

$$w(\boldsymbol{x},\sigma,\boldsymbol{th}) = \begin{cases} \exp\left(-\frac{\|\boldsymbol{X}\|_2}{2\sigma}\right) & (x \leq \boldsymbol{th}) \\ 0 & (else), \end{cases} \qquad (13)$$

where σ is a variable of sigma, and $\boldsymbol{th}$ is threshold values. For example, in space categories, sigma of space weight σ_s corresponds to σ and kernel radius of the filter corresponds to $\boldsymbol{th}$, and sigma of color weight σ_c corresponds to σ and the threshold β corresponds to $\boldsymbol{th}$. So we measure a ratio of using Gaussian distribution. The ratio is defined as:

$$(\textit{The ratio of using Gaussian distribution}) = 100.0\left(1.0 - \exp\left(-\frac{\|\boldsymbol{th}\|_2}{2\sigma}\right)\right) \quad (14)$$

The ratio is usage of Gaussian distribution in each kernel size or the threshold. If the ratio is very small, the threshold value is very small or the variable of Gaussian sigma is very large.

After setting up the optimal parameters shown in Section 4, we evaluate relativity between sigma and threshold in each category. When parameters in one category are evaluated, other categories parameters are set with the optimal parameters. The kernel size or threshold in the evaluating category is changed at regular intervals. At this time, the Gaussian variable of sigma is manually reconfigured at the optimal point. Then the ratio of using Gaussian distribution is calculated.

Figure 8 shows the ratio of optimal usage of the space Gaussian distribution in each kernel size. The Gaussian ration has small ratio in all data sets. To keep the ratio small, we should set the parameter of sigma large. In this case, shape of the kernel becomes nearly box kernel.

Figure 9 shows the ratio of optimal usage of the color Gaussian distribution in each intensity threshold. The ratio becomes larger as the threshold becomes larger. In addition, the optimal ratio reaches about 100% in all data sets. When the ratio is about 100%, the optimal variable of Gaussian sigma is constant. In addition, when the ratio is low, all thresholds setting reshape Gaussian distributions to have higher sigma by lift up that tail of the distribution. These facts show that the optimal shape of the distribution is Gaussian distribution, thus the intensity threshold should be set high with appropriate color weight parameter. Figure 10 shows the ratio of optimal usage of the depth Gaussian distribution in each depth threshold. The result has same trend of color category.

5.2 PARAMETER DEPENDENCY

In this subsection, we evaluate relativity between error rate and each category. Before the experiment, all parameters are set optimal, again. When one category is evaluated, other categories' parameters are fixed.

Figure 11 shows the error rate of each kernel size, when the Gaussian variable of space weight is set optimal percentage. The error rate of a number of data set, excluding Venus, has a peak position. The kernel size of the peak position is about 15. There are many objects in the Tsukuba, Teddy, and Cones dataset. Thus the kernel size is not so large. However the error rate of Venus does not have a peak position in the experiment. It is because that there are only a few large objects in Venus, and almost regions are flat. Thus the kernel size should be set larger, when refining images like this. The optimal ratio (Fig. 8) at the optimal point is about 10% in all data sets.

Figure 12 shows the error rate of each color threshold. All data sets have a peak position. The threshold value of the peak position is about 60 to 80 in the Venus, Teddy, and Cones data sets. But the threshold value of the peak position is about 180 in the Tsukuba data set. It is because the amount of noise in color images is different between these data sets. The Tsukuba data set is recorded by University of Tsukuba in Japan. On the other hand, the Venus, Teddy, and Cones data sets are recorded by Middlebury College in the United States. So the recording environment is different in each data set. The variable should be set according to the amount of noise in the joint image. The optimal ratio (Fig. 9) at the optimal point is 100% in all data sets.

Figure 13 shows the error rate of each depth threshold. All data sets have a peak position. The threshold value of the peak position is about 10 to 30. The optimal threshold value of depth weight is smaller than color's it. It is because the depth values have smaller variance than the color intensity. Variance of depth values depends on depth map estimation method. Thus when we use unstable depth estimation method, we should set larger parameter of depth categories. The optimal ratio (Fig. 10) at the optimal point is about 90% to 95% in all data sets.

Figure 14 shows the error rate of each threshold of LR-Check. All data sets have a peak position, and the threshold value of the peak position is about 1 to 4. It is small range. It is because that if the threshold value is large, the value makes no sense. Additionally, if the threshold value is set to 0, the condition of LR-Check is very hard. So the threshold value of LR-Check should be set small value, excluding 0.

6. CONCLUSION

In this paper, we proposed a depth map refinement filter called joint trilateral filter for a free viewpoint image synthesis and a point cloud rendering.

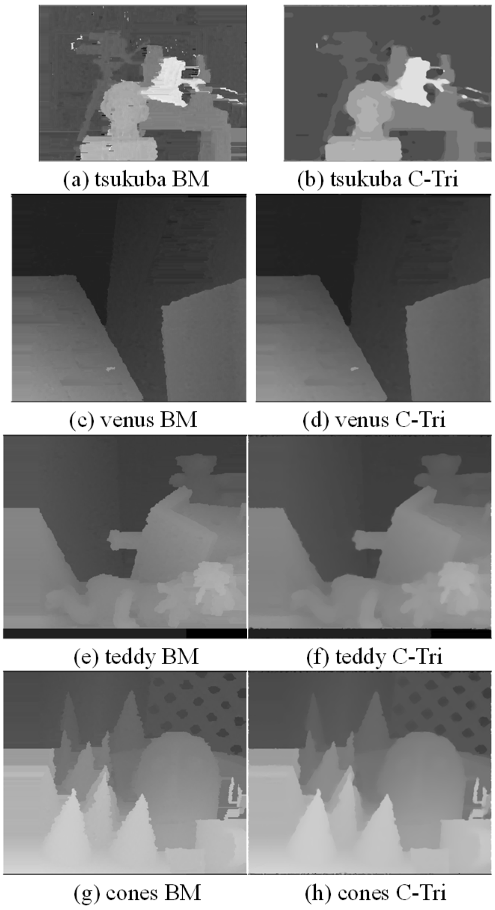
Experimental results of the depth refinement show that the error rate of the depth map is reduced up to 3.06%, and the improvement rate is 52.6% in Tsukuba. Also, when the number of gradation of the depth map is low, the accuracy of the joint trilateral filter is about the same as the SGBM. In addition, the joint trilateral filter is highly effective compared with other refinement filters. Moreover, the proposed filter is independent of the number of gradations of depth map so that computational cost becomes lower than the SGBM when the number is large. Therefore the filter is suitable for real-time application.

Experimental results of the view synthesis show that PSNR of using the joint trilateral filter is improved by 0.78 dB compared to using the BM, and

PSNR difference between the joint trilateral filter and the SGBM are 1.10 dB. The joint trilateral filter is more effective than optimization method of the SGBM in the free viewpoint image synthesis because object edges in the depth maps are corrected. In addition, it is possible that the joint trilateral filter adapts depth maps from depth sensors like Kinect.

Our future works are to extend this filter to be independent of the number of gradations of depth map and to improve accurate.

The authors would like to thank Professor Shinji Sugawara for valuable discussions, and the anonymous reviewers for their helpful comments and suggestions. This work was partly supported by the Grand-In-Aid for Young Scientists (B) of Japan Society for the Promotion of Science under Grant 22700174 and “Realization of Real-time Free Viewpoint Video Transmittion by Using Depth Map Super Resolution (AS232Z01514A)", Adaptable and Seamless Technology Transfer Program through Target-driven R&D, Japan Science and echnology Agency, and SCOPE (Strategic Information and Communications R&D Promotion Programme) 122106001 of the Ministry of Internal Affairs and Communications of Japan.


(a) tsukuba BM (b) tsukuba C-Tri

(c) venus BM (d) venus C-Tri

(e) teddy BM (f) teddy C-Tri

(g) cones BM (h) cones C-Tri

Fig.4: *Results of depth map: left side: BM without filter, right side: BM with proposed refinement filter.*

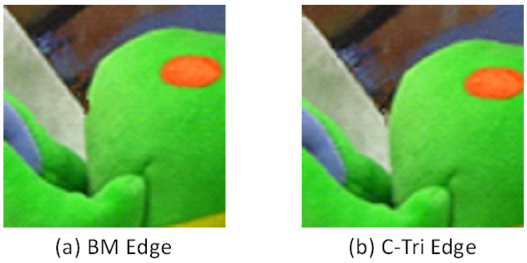
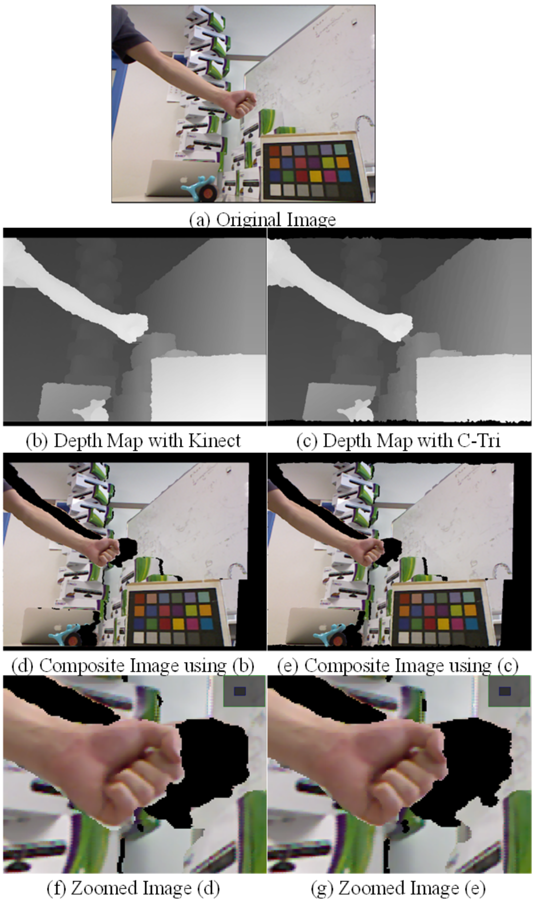

(a) BM Edge (b) C-Tri Edge

Fig.5: *Zoom up of synthesized view of Teddy.*


(a) Original Image

(b) Depth Map with Kinect (c) Depth Map with C-Tri

(d) Composite Image using (b) (e) Composite Image using (c)

(f) Zoomed Image (d) (g) Zoomed Image (e)

Fig.6: *Results of refined depth map and warped view from Kinect depth map.*

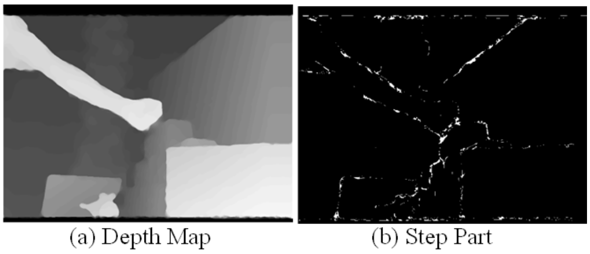

(a) Depth Map (b) Step Part

Fig.7: *Results of proposed filtering and detection result of ramp edge.*

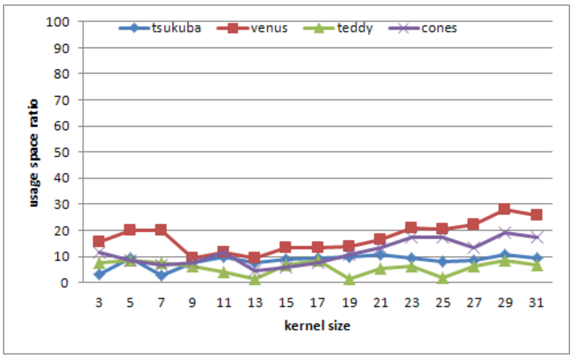


Fig.8: *Results of usage of Gaussian distribution (space).*

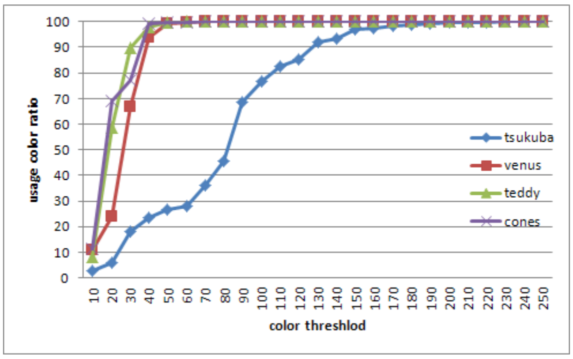


Fig.9: *Results of usage of Gaussian distribution (color).*

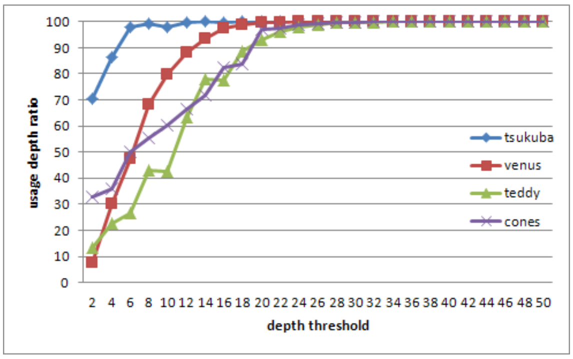


Fig.10: *Results of usage of Gaussian distribution (depth).*

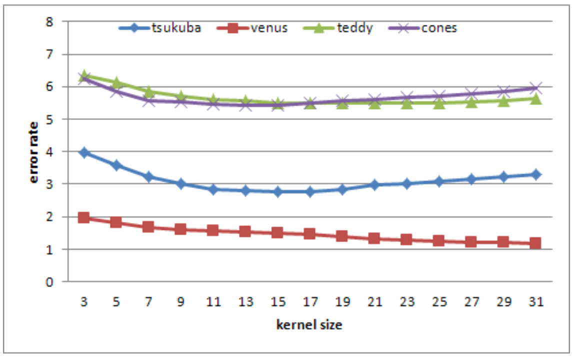


Fig.11: *Results of error rate at each kernel size.*

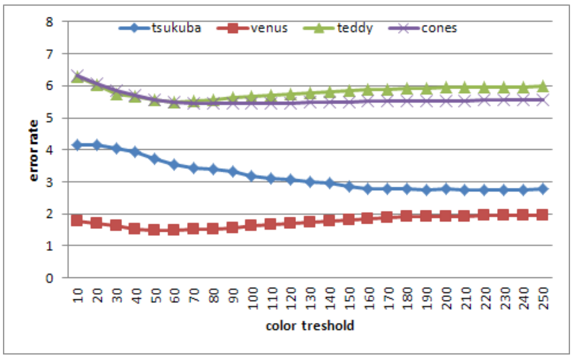


Fig.12: *Results of error rate at each color threshold.*

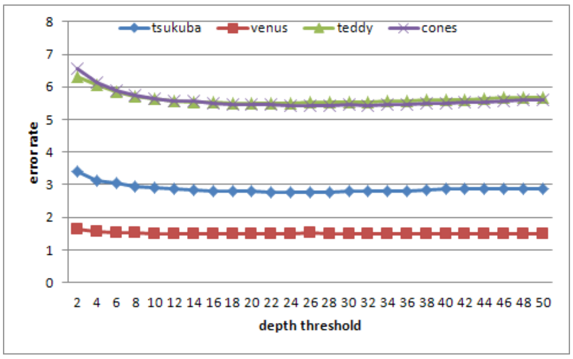


Fig.13: *Results of error rate at each depth threshold.*

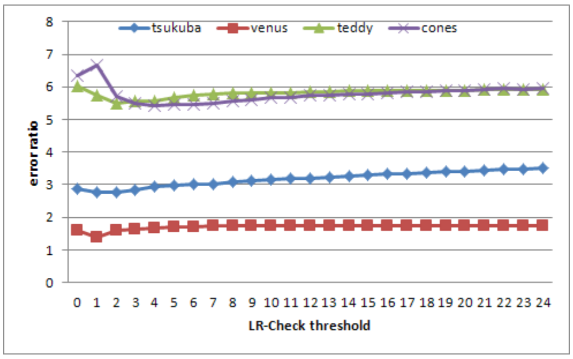


Fig.14: *Results of error rate at LR-Check threshold.*

References

[1] Kinect, "`http://www.xbox.com/`".

[2] Xtion, "`http://event.asus.com/wavi/product/WAVI_Pro.aspx`".

[3] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View Generation 3D Warping Using Depth Information for FTV," *Signal Processing: Image Communication*, Vol. 24, Issues 1-2, pp. 65–72, Jan. 2009.

[4] D. Scharstein, and R. Szeliski, "A Taxonomy and Evaluation of Depth Two-Frame Stereo Correspondence Algorithms," *International Journal of Computer Vision*, Vol. 47, Issues 1-3, pp. 7–42, Apr.-June 2002.

[5] H. Hirschmuller, "Stereo Processing by Semi-

global Matching and Mutual Information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 2, pp. 328–341, Feb. 2008.

[6] J. Sun, N. N. Zheng, and H. Y. Shum, "Stereo Matching Using Belief Propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 7. pp. 787–800, July 2003.

[7] Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, Issue 11, pp. 1222–1239, Nov. 2001.

[8] C. Tomasi, and R. Manduchi, "Bilateral Filtering for Gray and Color Image," *Proceedings of IEEE International Conference on Computer Vision (ICCV'98)*, pp. 839–846, Jan. 1998.

[9] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama, "Digital Photography with Flash and No-Flash Image Pairs," *ACM Transactions on Graphics*, Vol. 23, No. 3, pp. 664–672, Aug. 2004.

[10] E. Eisemann, and F. Durand, "Flash Photography Enhancement via Intrinsic Relighting," *ACM Transactions on Graphics*, Vol. 23, No. 3, pp. 673–678, Aug. 2004.

[11] J. Kopf, M.F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint Bilateral Upsampling," *ACM Transactions on Graphics*, Vol. 26, No. 3, pp. 96, July 2007.

[12] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth Super Resolution for Range Images," *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR'07)*, June 2007.

[13] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A Noise-aware Filter for Real-time Depth Upsampling," *Proceedings of European Conference on Computer Vision (ECCV'08) Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, Oct. 2008.

[14] K.-J. Yoon, and I. S. Kweon, "Adaptive Support-weight Approach for Correspondence Search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 4, pp. 650–656, Apr. 2006.

[15] Q. Yang, L. Wang, and N. Ahuja, "A Constant-space Belief Propagation Algorithm for Stereo Matching," *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR'10)*, pp. 1458–1465, June 2010.

[16] M. Mueller, F. Zilly, and P. Kauff, "Adaptive Cross Trilateral Depth Map Filtering," *Proceedings of 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON'10)*, June 2010.

[17] J. Jachalsky, M. Schlosser, and D. Gandolph, "Confidence Evaluation for Robust, Fast-converging Disparity Map Refinement," *Proceedings of IEEE International Conference on Multimedia and Expo (ICME'10)*, pp. 1399–1404, July 2010.

[18] Z. Wang, "Image Quality Assessment: from Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600–612, Apr. 2004.

[19] K. Takahashi, "Theoretical Analysis of View Interpolation with Inaccurate Depth Information", *IEEE Transactions on Image Processing*, Vol. 21, No. 2, pp. 718–732, Feb. 2012.

Takuya Matsuo received a B.E. degree from Nagoya Institute of Technology, Japan, in 2011. Since 2011, he is master student in Graduate School of Engineering, Nagoya Institute of Technology, Japan. His research interests are depth estimation and refinement.

Naoki Kodera received a B.E. degree from Nagoya Institute of Technology, Japan, in 2012. Since 2012, he is a research student in Faculty of Engineering, Nagoya Institute of Technology, Japan. His research interest is free viewpoint image synthesis.

Norishige Fukushima received a B.E., M.E., and Ph.D. degree from Nagoya University, Japan, in 2004, 2006, and 2009, respectively. Since 2009, he has been an assistant professor at Graduate School of Engineering, Nagoya Institute of Technology, Japan. His research interests are multi view image capturing, calibration, processing, and coding.

Yutaka Ishibashi received the B.E., M.E., and Dr.E. degree from Nagoya Institute of Technology, Nagoya, Japan, in 1981, 1983, and 1990, respectively. In 1983, he joined the Musashino Electrical Communication Laboratory of NTT. From 1993 to 2001, he served as an Associate Professor of Department of Electrical and Computer Engineering, Faculty of Engineering, Nagoya Institute of Technology. Currently, he is a Professor of Department of Scientific and Engineering Simulation, Graduate School of Engineering, Nagoya Institute of Technology. His research interests include networked multimedia, QoS (Quality of Service) control, and media synchronization. He is a fellow of IEICE and a member of IEEE, ACM, IPSJ, ITE, and VRSJ.