DEPTH MAP UP-SAMPLING USING COST-VOLUME FILTERING

Ji-Ho Cho[1]*, Satoshi Ikehata*[2]*, Hyunjin Yoo*[1]*, Margrit Gelautz*[1]*, and Kiyoharu Aizawa*[2]

[1]Vienna University of Technology, Austria [2]The University of Tokyo, Japan

ABSTRACT

Depth maps captured by active sensors (*e.g*., ToF cameras and Kinect) typically suffer from poor spatial resolution, considerable amount of noise, and missing data. To overcome these problems, we propose a novel depth map up-sampling method which increases the resolution of the original depth map while effectively suppressing aliasing artifacts. Assuming that a registered high-resolution texture image is available, the cost-volume filtering framework is applied to this problem. Our experiments show that cost-volume filtering can generate the high-resolution depth map accurately and efficiently while preserving discontinuous object boundaries, which is often a challenge when various state-of-the-art algorithms are applied.

Index Terms— Depth map super-resolution, cost-volume filtering, up-sampling

1. INTRODUCTION

Depth map acquisition is an active research area in image processing and computer vision. Various depth acquisition methods have been proposed to date; these can generally be categorized into two approaches: passive and active. Passive approaches generate a depth map of the scene by multiple image correspondences and triangulation [1], whereas active sensors, *e.g*., time-of-flight (ToF) and Kinect cameras, measure the distance from the camera to the objects directly using active infrared illumination. While those active cameras are becoming a popular alternative to passive approaches due to their simplicity, depth maps acquired by active sensors often contain considerable noise for objects with low reflectivity, and have very low-resolution.

So far, many works on depth map up-sampling have emerged and they are mainly categorized into two classes. One class is motivated by the image super-resolution literature, which explicitly considers the low-resolution image formation process. Some algorithms in this class reconstruct a high-resolution depth map of a static scene by fusing multiple low-resolution depth maps that were observed together [2, 3]. More recently, learning-based single image super-resolution techniques were integrated into depth map super-resolution to handle dynamic scenes [4]. Though they do not require a registered high-resolution texture image, a time-consuming dictionary learning for each magnification factor is often required.

On the other hand, the second class determines depth values at interpolated coordinates of the input domain in the manner of multi-modal filtering [5–8] or Markov random field (MRF) modeling [9, 10]. This class works well for noisy low-resolution depth maps by leveraging a registered high-resolution texture image assuming the co-occurrence of depth and texture structures. While general purpose multi-modal filters (*e.g*., joint bilateral up-sampling [5] or guided image up-sampling [6]) have been applied to up-sample a low-resolution depth map, they sometimes give inaccurate edges which result in considerable color bleeding artifacts. To overcome this problem, Park *et al*. [10] integrated a non-local means filter into the smoothness term of an MRF-based depth map up-sampling scheme, which reasonably preserves object discontinuities. Moreover, some recent works quantize the depth values into several discrete layers. Yang *et al*. [7] build a 3-D cost-volume using a low-resolution depth map and then perform the joint bilateral filtering [11] for a 2-D cost slice of each depth candidate. Min *et al*. [8] proposed a weighted mode filter which also finds the global mode of a filtered cost-volume yet more efficiently than [7]. These approaches use an original sparse depth map to construct a cost function for the up-sampling process, which leads to critical aliasing artifacts when the up-sampling ratio is high (*e.g*., more than $8\times$). To solve this problem, Yang *et al*. [7] iteratively perform the joint bilateral filtering. Min *et al*. [8] hierarchically iterate the up-sampling process from coarse to fine levels. While effective, these approaches sometimes give over-smoothed edges and blurred details during iterations.

In this paper, we instead leverage the cost-volume filter which relies on the guided image filter [12]. The cost-volume filtering was originally proposed to alternate time-consuming MRF-based solutions for optimizing labeling problems. The authors of [12] have shown that the framework can be applied to many different applications such as stereo matching, optical flow, and binary segmentation and demonstrated that high-quality results are obtained fast. The main differences between our method using cost-volume filtering and the method proposed by Yang *et al*. are summarized as follows:

1) We apply the guided image filtering [6] when each

This work was supported in part by the Austrian Science Fund (FWF, M1383-N23) and JSPS Research Fellowships for Young Scientists.

slice of cost-volume is filtered because the guided filtering has more potential to achieve an efficient computation than the joint bilateral filtering [11].

2) We design a new confidence measure which not only reduces aliasing artifacts but also restores the missing depth areas without time-consuming iterations.

As a result, the proposed method can up-scale the depth map fast while preserving discontinuous object boundaries and suppressing aliasing artifacts.

2. THE PROPOSED METHOD

The proposed method utilizes the cost-volume filtering framework to super-resolve the small resolution of an input image. We assume that a well-aligned high-resolution texture (RGB image) is available. The proposed method consists of three main steps: 1) construction of a cost volume, 2) filtering of a cost-volume, and 3) selection of the final label.

Cost-volume construction: the cost-volume C is constructed by using the absolute difference between the potential discrete depth label $l \in \{1, ..., N\}$ and an initially up-sampled depth $\hat{d^H}$ from the input depth map d^L with a weight value ω for each pixel p:

$$C_{p,l} = \omega_p ||l - \hat{d_p^H}||. \tag{1}$$

The initial high-resolution depth map $\hat{d^H}$ is obtained by the nearest neighbor approach. We apply this approach to avoid blurred pixels in the object boundaries because it computes new pixels as the value of the nearest pixel in the original image. However, the up-sampled result often suffers from serious aliasing artifacts. To solve this problem, previous works in [7, 8] iteratively perform the filtering process which requires more computation time and produces blurry edges. Instead, we measure the confidence ω_p of $\hat{d_p^H}$:

$$\omega_p = \begin{cases} 0 & \text{if } \hat{d_p^H} < \tau, \\ \exp(\frac{-(||I_p^H - I_{p_\downarrow}^L||)^2}{2\sigma^2}) & \text{otherwise.} \end{cases} \tag{2}$$

where I^H is a guidance high-resolution image, I^L is the low-resolution image obtained by the nearest neighbor down-sampling from I^H to align with the input low-resolution depth map, $p_\downarrow$ is the corresponding location in the low-resolution image (if $p = (x, y)$, then $p_\downarrow = ([x/s], [y/s])$, s is a scaling factor), and $\sigma = 0.1$ in all experiments. τ is a threshold value which determines the missing pixels caused by occlusions, shadows, and low-reflections.

The confidence is measured using the color difference of pixels between coordinates p in I^H at full-resolution and the corresponding down-sampled coordinates $p_\downarrow$ in I_L. It means that the up-sampled depth value has high confidence when its corresponding down-sampled color is similar to the one from the original high-resolution image. Then, low-confidence depth values are replaced by propagation from neighboring pixels which have high confidence during the following filtering and selection steps. Fig. 1 shows the confidence map containing the weights ω of the Moebius scene.

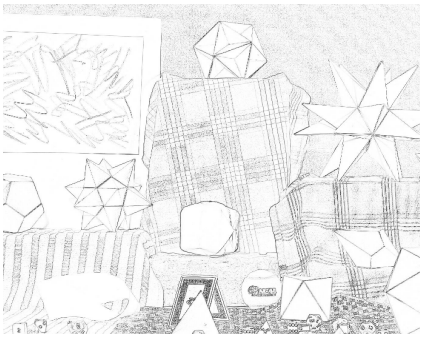


Fig. 1. Our computed confidence map of the Moebius scene ($8\times$ up-scaling). Note that bright pixels mean high confidence.

Cost-volume filtering: In this step, each piece of the cost-volume is filtered by guided image filtering. The output of the filtering at pixel p is a weighted average of pixels in the same label l:

$$C'_{p,l} = \sum_q W_{p,q}(I^H) C_{q,l} \tag{3}$$

where C' is the filtered cost volume and p and q are the pixel locations. The weight $W_{p,q}$, which is dependent on the high-resolution color image I^H, is as follows:

$$W_{p,q} = \frac{1}{|w_k|^2} \sum_{(p,q)\in w_k} (1 + (I_p^H - \mu_k)^T (\Sigma_k + \epsilon U)^{-1} (I_q^H - \mu_k)), \tag{4}$$

where Σ_k is a 3×3 covariance matrix, μ_k is a 3×1 mean vector of r, g, and b in each 3-D window w_k with dimensions $w_x \times w_y$ centered at pixel k, and U is a 3×3 identity matrix. ϵ is a user parameter and we set 0.04 for all our experiments. More details about the parameter ϵ can be found in [6].

Cost selection: Finally, for each pixel p, the final label f_p which has the minimum cost value is selected by:

$$f_p = \arg\min_l C'_{p,l}. \tag{5}$$

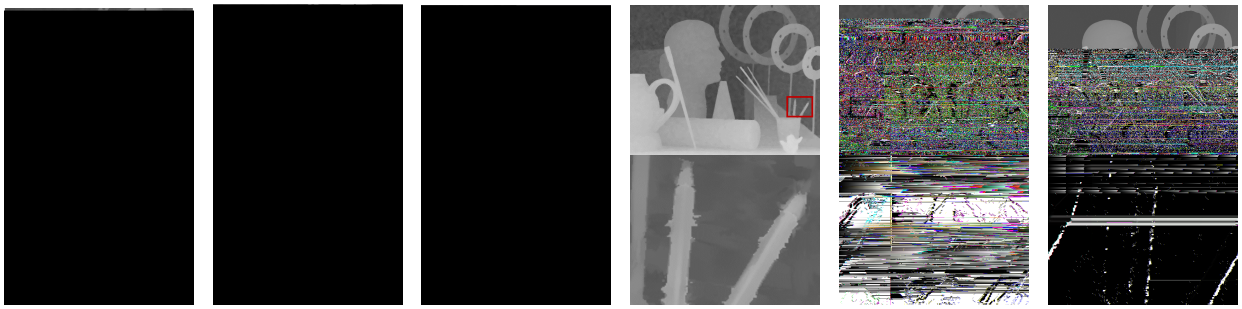
3. EXPERIMENTAL RESULTS

This section describes the evaluation of our method on various data sets. τ in Eq. (2) is 10 and the number of candidate depth values N is 256 (8 bits) for all experiments.

3.1. Middlebury stereo dataset

First, we compared our method with four different depth map super-resolution methods: joint bilateral up-sampling [5],

(a) JBU [5] (b) MRFU [10] (c) WMFU [8] (d) Iterative JBU [7] (e) our method (f) Ground-truth



(g) JBU [5] (h) MRFU [10] (i) WMFU [8] (j) Iterative JBU [7] (k) our method (l) Ground-truth

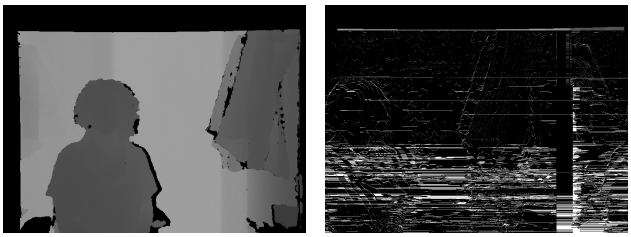
Fig. 2. Experimental results of 8× image up-sampling. The upper row represents the results on the clean Moebius set and the lower one illustrates results on a noisy Art data set from [10].

MRF-based method [6], weighted mode filtering [8], and iterative joint bilateral up-sampling [7] using the Middlebury *Art* and *Moebius* dataset. The original resolution of both datasets is 1376 ×1088.

In this experiment, the low-resolution depth image is up-sampled with a guidance of the corresponding high-resolution RGB image. As depicted in Fig. 2, our up-sampling method can super-resolve the low-resolution depth map while preserving depth discontinuities. Table 1 shows that the proposed method can generate lower error rates in terms of root means squared error (RMSE). Table 2 shows the results on a noisy dataset provided by [10]. The MRFU method produces the best result in terms of RMSE in the noisy data up-sampling case because it uses an MRF framework which is very robust against noise. Our method also produces low error rates for the noisy dataset, which indicates that our cost volume filtering approach is also able to cope with noise. Furthermore, we have achieved a very fast computation time using a GPU implementation. It takes about 0.5 seconds to up-scale the low-resolution image into 1376×1088 size. The scaling factor does not affect the computation time.

3.2. Kinect dataset

Kinect captures a registered RGB image and depth map. The maximum resolution of depth is 640×480 while the maximum resolution of RGB is 1280×960. The captured depth often suffers from missing data caused by occlusion and low reflectivity as shown in Fig. 3. Since our weight for these corrupted pixels is 0, their depth values are restored by high confident neighboring pixels. As a result, our method can generate a high-resolution depth map while preserving depth discontinuities.



(a) Nearest neighbor up-sampling (b) our method

Fig. 3. Comparison of up-sampling results for the low-resolution depth from Kinect.

4. CONCLUSIONS

In this paper, we have proposed a cost-volume filter based depth map up-sampling. We have demonstrated that the proposed method can generate a high-resolution depth with discontinuous object boundaries being preserved without an iteration, while suppressing aliasing artifacts. A certain limitation of our method is that we need a well-aligned high-

Table 1. Quantitative Evaluation of Depth Map Super-resolution (RMSE)

	Art				Moebius			
Method	2×	4×	8×	16×	2×	4×	8×	16×
JBU [5]	**0.3538**	0.6250	1.1327	2.0394	0.1886	0.4270	0.9546	1.6708
MRFU [10]	0.4306	0.6745	1.0734	2.2117	0.1795	0.2965	0.5218	0.8965
WMFU [8]	0.6521	0.9037	1.7460	3.2991	0.4672	0.6416	1.0044	1.7402
Iterative JBU [7]	0.5708	0.7002	1.5046	3.6903	0.3868	0.4760	0.6893	1.3660
Our method	0.3699	**0.5408**	**0.8371**	**1.7101**	**0.1423**	**0.2252**	**0.4165**	**0.8107**

Table 2. Quantitative Evaluation for Noisy Depth Map Super-resolution (RMSE)

	Art				Moebius			
Method	2×	4×	8×	16×	2×	4×	8×	16×
JBU [5]	1.5069	1.9484	2.9241	4.6926	1.7015	1.9206	2.3483	**3.0232**
MRFU [10]	**1.2401**	**1.8159**	2.7047	**4.3940**	**1.0343**	**1.4894**	2.1289	3.0910
WMFU [8]	1.9708	2.3191	3.3818	5.1306	1.8693	2.0430	2.5612	3.4895
Iterative JBU [7]	1.3592	1.9315	**2.4535**	4.5192	1.2506	1.6334	**2.0559**	3.2054
Our method	1.2903	1.9689	3.1132	4.6957	1.0633	1.5357	2.4692	3.0372

resolution guidance image to achieve the best performance.

5. REFERENCES

[1] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. of Computer Vision and Pattern Recognition*, 2006.

[2] C. Theobalt, J. Davis, and S. Thrun, "High-quality scanning using time-of-flight depth superresolution," in *Proc. of Computer Vision and Pattern Recognition Workshops*, 2008.

[3] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "Lidarboost: Depth superresolution for tof 3d shape scanning," in *Proc. of Computer Vison and Pattern Recognition*, 2009.

[4] O. M. Aodha, N. D. F. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image super-resolution," in *Proc. of European Conference on Computer Vision*, 2012.

[5] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 96, 2007.

[6] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. of European Conference on Computer Vision*, 2010.

[7] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. IEEE Computer Vison and Pattern Recognition*, 2007.

[8] D. Min, J. Lu, and M. N. Do, "Depth video enhancement based on weighted mode filtering," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1176–1190, 2012.

[9] J. Lu, D. Min, R. S. Pahwa, and M. N. Do, "A revisit to mrf-based depth map super-resolution and enhancement," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2011.

[10] J. Park, H. Kim, Y-W. Tai, M. S. Brown, and I. Kweon, "High quality depth map upsampling for 3d-tof cameras," in *Proc. of International Conference on Computer Vision*, 2011.

[11] 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, 2004.

[12] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 504–511, 2013.