Joint Depth Map Upsampling via Prior Minimum Spanning Tree

Xiu Li
Graduate School at Shenzhen, Tsinghua University
Shenzhen, China
li.xiu@sz.tsinghua.edu.cn

Zhixiong Yang
Graduate School at Shenzhen, Tsinghua University
Shenzhen, China
yzx14@mails.tsinghua.edu.cn

Lulu Xie
Graduate School at Shenzhen, Tsinghua University
Shenzhen, China
xielulu2011@126.com

***Abstract*—Depth map is the fundamental of various 3D applications. While consumer-level depth cameras have gained much attention recently due to their affordable cost, the low resolution and bad quality of depth maps generated by them have obstructed their applications in practice. In this paper, we propose a novel depth map upsampling algorithm based on prior minimum spanning tree (pMST). Utilizing the registered color and depth images, we construct a prior map which indicates the probability of the co-occurrence between depth discontinuities and color image edges. Under the joint guidance of prior map and color image, a prior minimum spanning tree is extracted and the high resolution depth map is obtained via a joint bilateral filter on the pMST. Experimental results demonstrate the effectiveness of our method compared to existing local depth super resolution methods.**

Keywords—Depth upsampling; Minimum Spanning Tree; Edge-aware filter; Depth super resolution; Prior map

I. INTRODUCTION

The depth map characterizes the distance between the scene and the camera, which is the fundamental of 3D object reconstruction, gesture recognition and robotic navigation. The classical methods to obtain depth maps, such as stereo matching algorithms and laser range measurements. However, the first have many limitations in practical applications, while the latter are too expensive to be popular in our daily life.

During recent years, the consumer-level depth cameras such as Time of Flight and Microsoft Kinect have gained much attention from the industry and academia, because of its convenience to get dense depth maps and affordable cost. However, there are some drawbacks in the depth map acquired by them. The depth maps obtained by the ToF range sensors are commonly in low resolution, e.g., 176×144 or 200×200, which are less appealing in vision applications. Although the Kinect can provide depth map with a higher resolution, e.g., 640×480, it is still not comparable with the corresponding high-resolution color image, which is usually 1280×1024. Apart from that, it suffers from missing depth information due to occlusion or other degradation.

To address the problem above, depth map upsampling and inpainting are unavoidable. Depth maps have different smooth depth regions, which are often separated by sharp object edges, and the missing depth information usually occurred between different depth layers. Usually, the depth discontinuities exists along with color image edges. Considering for that, a lot of methods interpolate the depth under the guidance of the registered high resolution optical image, trying to transfer the structural information of the color image to the depth map. For instance, by utilizing the famous bilateral filter, Kopf et al. [1] proposed a joint bilateral filter to upsample the depth map while preserving its edges with aligned high-resolution optical image. Yang et al. [2] applied the joint bilateral filter to the cost volume domain iteratively to improve the resolution of the depth map. Derek et al. [3] modified the bilateral filter by adding a range term to denoise and upsample the depth map at the same time. Yang [4] proposed the tree filter to upsample the depth map, which is very efficient because of the tree structure. Taking advantage of the correlation between color and depth image, Diebel et al. [5] proposed the Markov Random Fields to estimate the missing depth values and reconstruct high resolution depth images. Extending traditional bilateral filter to the trilateral filter, Kai-Han Lo et al. [6] utilized the trilateral filter to upsample the depth images with the pre-aligned color images.

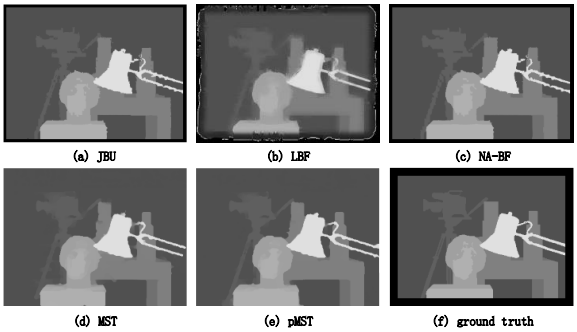


Fig.1. Visual result on the Middlebury dataset, the up-scale factor is 4. (a) JBU [1], (b) LBF [2], (c) NA-BF [3], (d) MST [4], (e) Our method, (f) ground truth. Zoomed for better view.

In fact, there is an assumption lies in the above methods: color image edges are co-occurrence with depth discontinuities, which does not hold true all the time. On one hand, there may be a large color difference in a depth soothing region. In this case, the upsampled depth map may suffer from texture copying artifacts. On the other hand, if the depth difference is large while the color difference is small in the area, it would lead to depth blurring artifacts. Actually, the gradient of color image textures are independent of depth discontinuities. If we upsample the

0054

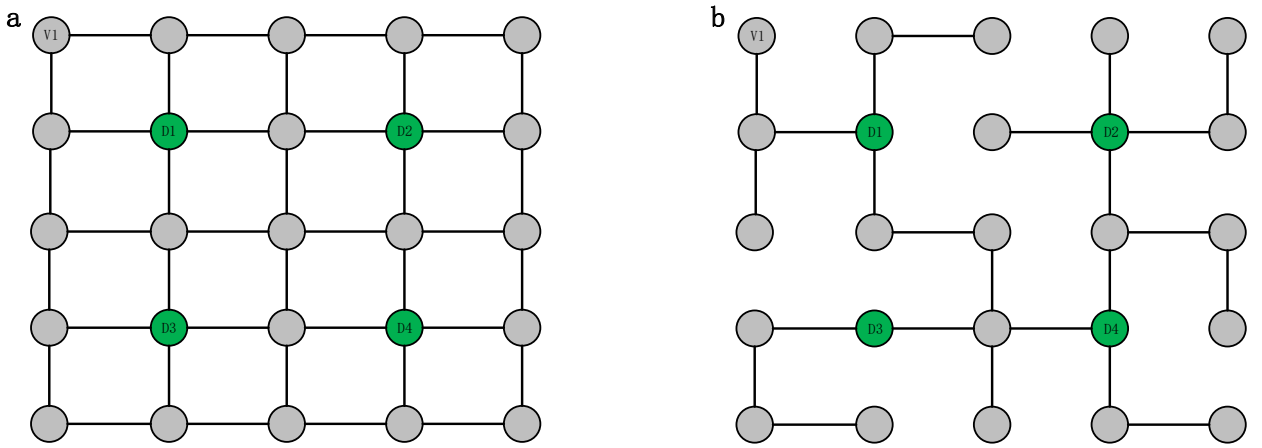


Fig.2. The illustration of MST. (a) The graph representation of guidance colour image ***I***, where the green nodes are pixels with depth values and the grey nodes denotes pixels without depth values. (b) The minimum spanning tree extracted from the graph left side, and the edges with large weight are dropped.

depth map under the guidance of the color image totally, there would be texture copying and depth blurring artifacts in the depth map.

In this paper, we present a novel algorithm for single depth map upsampling guided by utilizing the structure information between depth image and optical image. It can produce better results than previous reported local methods as shown in Fig.1. We are inspired by we could pose the prior, which color edges are most likely to coincide with depth discontinuities, to measure the relevance between color image edges and depth discontinuities. Besides, we treat the color image as an undirected, connected graph, and construct a minimum spanning tree with the aid of color image and the prior map. Then, a joint bilateral filter is adopted on the MST to upsample the depth map. Utilizing the prior map, we could avoid the artifacts caused by the inconsistent between color image edges and depth discontinuities. We will show the effectiveness of the proposed approach in the following sections.

The remainder of the paper is organized as follows. Section 2 introduces depth upsampling algorithm based on the MST. Section 3 gives details of the proposed algorithm, including the formulation of the prior map, the reconstruction of the MST and the modified bilateral filter on the MST. Section 4 demonstrates the experimental results compared with other methods. Section 5 provides concluding remarks.

II. DEPTH UPSAMPLING BASED ON MST

In this section, we review depth upsampling algorithm based on MST, which is an efficient non-local method. The MST is constructed based on colour image, and a bilateral filter is applied on the MST to upsample the low resolution depth map.

A. Joint bilateral filter based on the MST

Consider the reference color image I as a connected, undirected graph $G = (V, E)$, each node in V represents a pixel in I, and each edge in E connects a pair of neighboring pixels in I. The weight $\omega_e(r,s)$ of an edge e, connecting a pair of adjacent pixels r and s, is determined by the following formula:

$$\omega_e(r,s) = |I(s) - I(r)| \quad (1)$$

Then, a minimum spanning tree T can be derived from the graph G via Kruskal's or Prim's algorithm. During the construction of MST, edges with large weight are removed, corresponding to the assumption that weight average operation should try not to cross the depth edges. Fig.2 provides a simple illustration of MST's construction.

For any two pixels p and q, where only exists one path connecting them in T, which is also the shortest one. If p and q are similar, they are close among the MST, vice versa. The distance $D(p,q)$ between them is determined by the sum of weights along MST. Utilizing Gauss function to convert distance metrics to similarity metrics, we could evaluate the similarity of p and q as follows:

$$w(p,q) = w(p,q) = \exp\left(\frac{-D(p,q)}{\sigma}\right) \quad (2)$$

Where σ is a user-specified parameter to adjust the similarity between two nodes.

Then, the joint bilateral filter can be naturally extended to MST structure:

$$\mathrm{D(p)} = \frac{\sum_{q\in\Omega} w(p,q)D(q)}{\sum_{q\in\Omega} w(p,q)} = \frac{C(p)}{W(p)} \quad (3)$$

Where Ω stands for the nodes in the MST, $C(p)$ denotes the depth cost from all other pixels, $W(p)$ is the weight normalization parameter.

If we compute the weight directly, it would be very time consuming which is unrealistic in practical applications. However, by taking advantage of the tree structure, we can compute the depth cost and weight from other nodes efficiently. Yang [7] showed that we could compute the support from all other nodes efficiently by traversing the tree structure T in two sequential passes. We will show how to compute the depth cost and weight in the next subsection.

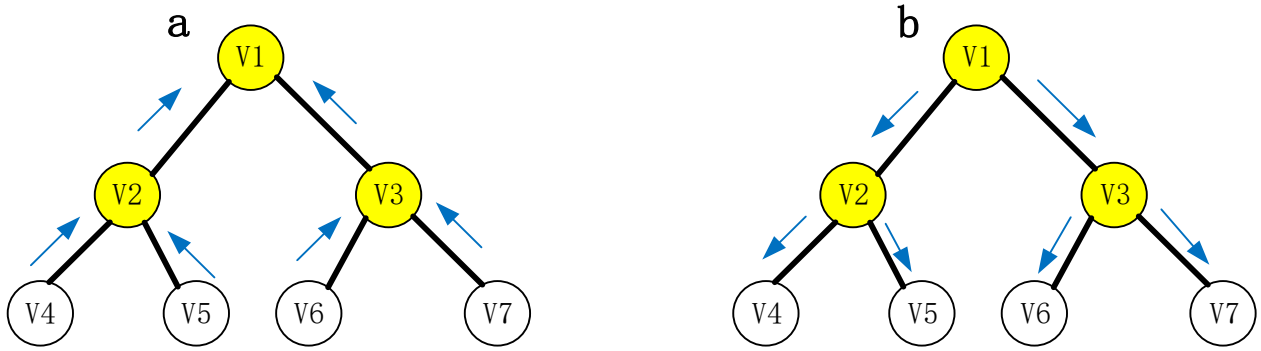


Fig.3. Two passes to compute the cost and weight along the tree. Yellow nodes represent the parent node. (a) First pass: from leaf to root. (b) Second pass: from root to leaf.

We could see that the MST's construction acts like image segmentation to some extent. It automatically drags away the pixels which are not belong to the same region or the same structure in the color image. At last, the nodes are divided into several clusters according to their pixel values. The spatial distance between pixels are naturally determined along the MST. Therefore, joint bilateral filter based on MST throws away the spatial kernel, which adjusts the pixels' distance between them. During the filtering process, pixels segmented into the same region contribute more to the target pixel. Unlike the fixed window size of bilateral filter, it is an adaptive weighted method, which is more reliable and reasonable for image filtering and depth upsampling.

B. Compute the cost and weight through tree structure

We map the low resolution depth image $D_{\downarrow}$ into the color image I pixel grid, and the sparse grid depth image denoted as D. There are some pixels in D with known depth values from the corresponding low resolution depth image $D_{\downarrow}$, which referred as seed pixels. The initialization cost $C(p)$ and weight $W(p)$ of pixel p are as follows: if it is a seed pixel, $C_0^{\uparrow}(\mathrm{p}) = \mathrm{D}(\mathrm{p}), W_0^{\uparrow}(p) = 1$, if not, $C_0^{\uparrow}(\mathrm{p}) = W_0^{\uparrow}(p) = 0$. According to Yang [7], the support (cost and weight) can be computed through two sequential passes. The computation process is illustrated in Fig.3.

In the first pass, the tree is traced from leaf nodes to root nodes. For a node v, its cost and weight value are not updated until all its children have been visited [6]:

$$C^{\uparrow}\big(P(v)\big) = C_0^{\uparrow}\big(P(v)\big) + \omega_e(v, P(v))C^{\uparrow}(v) \tag{4}$$

$$W^{\uparrow}\big(P(v)\big) = W_0^{\uparrow}\big(P(v)\big) + \omega_e(v, P(v))W^{\uparrow}(v) \tag{5}$$

Where $P(v)$ means the parent node of v, $C^{\uparrow}(v)$ and $W^{\uparrow}(v)$ are intermediate result.

In the second pass, the tree is traced from root nodes to leaf nodes. Starting from the root node, the cost and weight are passed to the sub-trees, and the updating formulas are as follows:

$$\mathrm{C}(\upsilon) = \omega_e(v, P(v))C^{\uparrow}\big(P(v)\big) + (1 - \omega_e^2(v, P(v)))\, C^{\uparrow}(v) \tag{6}$$

$$W(\upsilon) = \omega_e(v, P(v))W^{\uparrow}\big(P(v)\big) + (1 - \omega_e^2(v, P(v)))\, W^{\uparrow}(v) \tag{7}$$

For each node, only a total of 6 addition/subtraction and 7 multiplication/division operations are needed to estimate its depth value, which is more efficient than global optimization methods and many other local methods.

III. PRIOR MAP FOR THE CONSTRUCTION OF MST

In this section, we will first introduce how to get the prior map. Then, with the prior map and high resolution colour image, we will reconstruct the MST to get more accurate result.

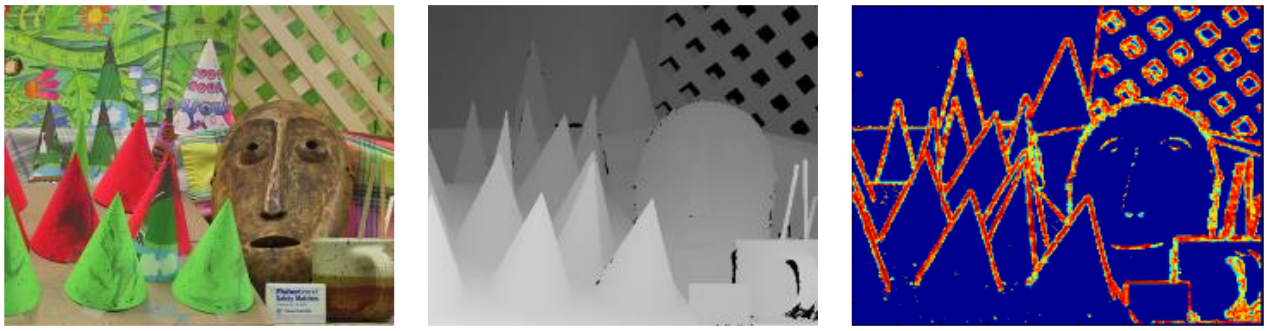
A. Prior map

Depth images are characterized as plain areas partitioned by sharp boundaries, which are not always consistent with color image edges. Therefore, the construction of MST above may introduce some false connections in the tree. Rather than simply taking the gradient magnitude of individual pixels into consideration, we integrate the color images and depth maps to examine the coincidence between local edge structures and depth layer boundaries [8]. Therefore, we construct a prior map to denote the similarity between color image edges and depth layer boundaries. Then, we utilize the prior map to guide the MST construction, alleviating the artifacts introduced by the false connections along the tree.

Low resolution depth image D_l is first upsampling to the resolution of color image I_c by bicubic interpolation. In fact, since the bicubic interpolation only utilizes the depth information in the low resolution depth image, the high resolution depth map obtained by it would be blurring but no texture copying and depth bleeding artifacts. Then, the prior map is computed as follows:

$$t(x, y) = \frac{|<G_l(\mathrm{x,y}), G_c(\mathrm{x,y})>|}{\|G_l(\mathrm{x,y})\|_2 \|G_c(\mathrm{x,y})\|_2} \tag{8}$$

In the formula above, $G_l(x, y) = \begin{pmatrix} G_x(x', y') \\ G_y(x', y') \end{pmatrix}, (x', y') \in \mathrm{N}(x, y)$, $G_x(x', y')$ and $G_y(x', y')$ denote the gradients in the x and y directions within a local window $\mathrm{N}(x, y)$ centered at pixel (x, y), respectively. $G_c(x, y)$ is the same as $G_l(x, y)$, which is defined using the maximum gradient magnitudes among the color channels. |x| means the absolute value of x, $< \mathrm{x}, \mathrm{y} >$ denotes the dot product among vector x and y, and $\| x \|_2$ is the



(a) (b) (c)

Fig.4. The comparison of colour image, ground truth depth image and prior map. (a) Colour image, also the guidance image. (b) Ground truth depth image. (c) The prior map. Note that the prior map is computed through the bicubic interpolation to upsample the depth map with 4×, and we put them together just for comparison. The red demonstrates higher value, and the blue indicates lower value, which means the small difference between colour image edges and depth discontinuities.

Euclidean L_2 norm of a vector x. Besides, $\mathrm{t}(x, y)$ is set to be zero if $\| G_l(\mathrm{x},\mathrm{y}) \|_2$ or $\| G_c(\mathrm{x},\mathrm{y}) \|_2$ is smaller than a threshold $\mathcal{E}$. Fig.4 gives an example of the prior map, where the red indicates higher values, and the blue indicates lower values.

B. The construction of pMST

The map gives us a prior that which color image edges are most likely to coincide with depth discontinuities. We take advantage of the prior to guide the MST's construction, and the weight of the edges are dependent on the color difference and the prior we employed. The weight $\omega_e(r,s)$ of an edge is rewritten as follows:

$$\omega_e(r,s) = \begin{cases} |I(s) - I(r)| \cdot \big(1 + t(r)\big) & if\ t(r) > \tau_1 \\ min(|I(s) - I(r)|, \tau_2) \end{cases} \quad (9)$$

Here, τ_1 is the threshold of the similarity prior, τ_2 is the threshold of the color difference of two neighboring pixels s and r. In the smooth region, the edge weight is built based on the color image (the prior is not taking into consideration). In order to avoid the influence of large difference in color image, the weight should be constant as the differences become large. We adopt the truncated model to address the problem above, where the edge weight increases linearly based on the difference in the color image within a limitation τ_2. When the pixel's prior is bigger than a certain threshold τ_1, we need to consider the corresponding discontinuities of depth map. The weight is larger than the original MST in the area, which only takes the color difference into consideration. As a matter of fact, the prior map serves as a more severe condition for image segmentation. The original weight segments the image under the texture difference guidance of color image, while the mixing of prior map integrates the depth map discontinuities into color image differences to provide a joint image segmentation implicitly. Our depth upsampling algorithm based on pMST is listed in Algorithm 1.

Therefore, the pixel to be filtered out receives weight support within planar surface with the same depth values and preserves the depth boundaries. Theoretically, it is the most reasonable way to get the accurate depth values for local weight average methods.

Algorithm 1 Joint depth upsampling with pMST

Input:
LR depth image $D_{\downarrow}$.
HR colour image I_c.
Output:
HR depth image D_h.

1: Upsample $D_{\downarrow}$ to the resolution of I_c via bicubic interpolation and get $D_{\uparrow}$.
2: **for** each pixel in $D_{\uparrow}$
3: Compute the similarity t with Equation (8)
4: **end for**
5: Check the graph $G = (V, E)$ from sparse grid depth image D.
6: Edge weights are given by Equation (9).
7: Extract the MST T from G via Kruskal's algorithm.
8: **for** each node in T
9: Compute the cost $\mathrm{C}(v)$ by Equation (4)
10: Compute the weight $\mathrm{W}(v)$ by Equation (5)
11: **end for**
12: Update the intermidiate result $C^{\uparrow}(v)$ and $W^{\uparrow}(v)$.
13: **for** each node in T
14: Compute the cost $\mathrm{C}(v)$ by Equation (6)
15: Compute the weight $\mathrm{W}(v)$ by Equation (7)
16: **end for**
17: **for** each pixel in D
18: Obtain the final depth value by Equation (3)
19: **end for**
20: **return** HR depth image D_h.

In reality, one pixel also receives the weight support from different depth regions, but when the weight propagates across the depth boundaries, it would reduce to be negligible rapidly. Traditional local filtering methods, such as bilateral filter and guided filter, compute the weight support within a fixed size window, while the pixels outside of the local window contribute

0057

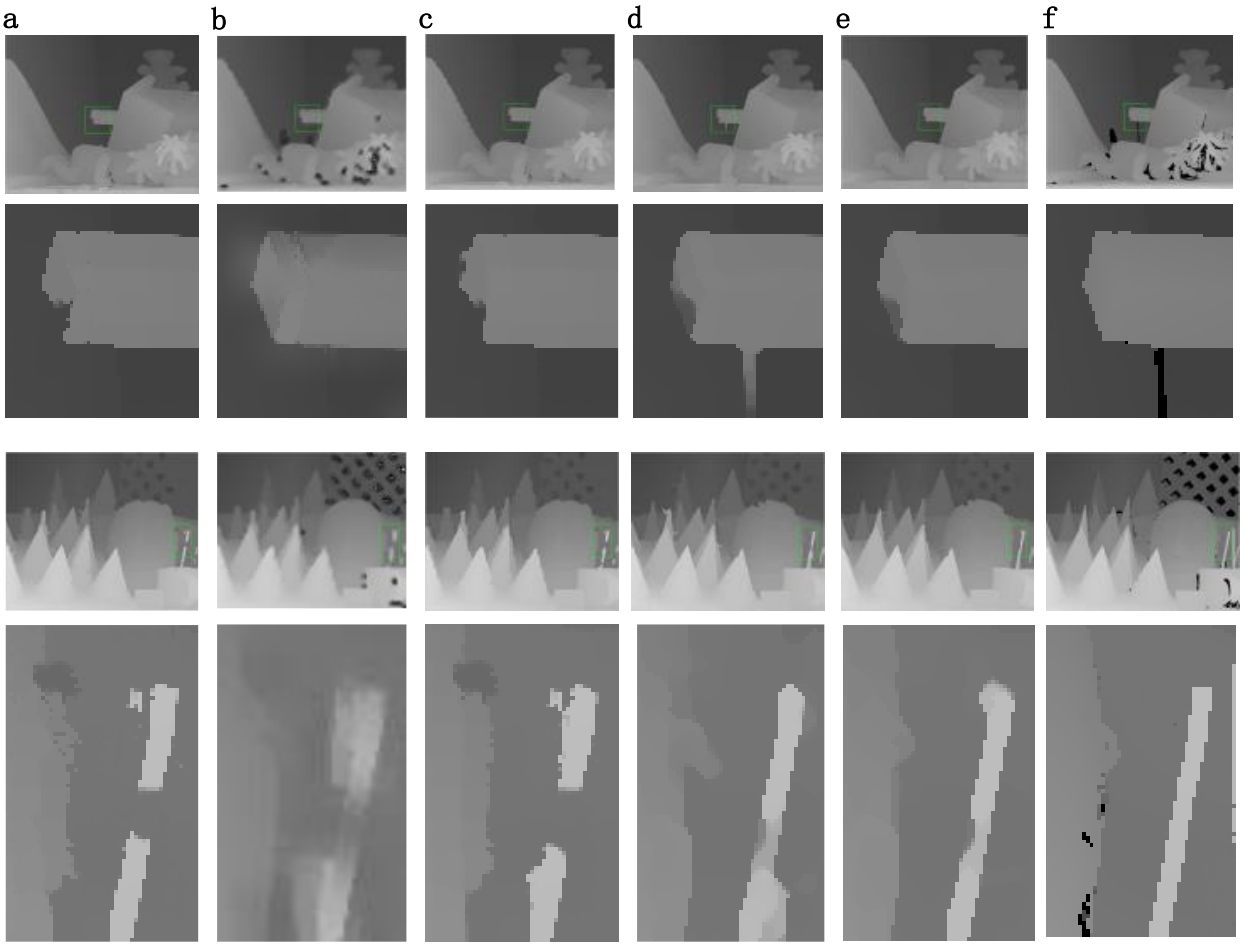


Fig.5. Visual comparison of deferent depth upsampling methods at 8×. From left to right, the result is (a) JBU [1], (b) LBF [2], (c) NA-BF [3], (d) MST [4], (e) our method: pMST and (f) ground truth [6]. From the experiment result, our method avoids the depth bleeding artifacts and preserves the thin depth structures better.

Table 1. Quantitative evaluation of five methods.

		Teddy		Cones		Venus	
		4X	8X	4X	8X	4X	8X
JBU	all	0.069	0.116	**0.033**	0.104	0.005	0.041
	disc	0.152	0.255	0.069	0.173	**0.026**	0.076
LBF	all	0.158	0.283	0.175	0.333	0.036	0.123
	disc	0.422	0.649	0.393	0.615	0.271	0.579
NABF	all	0.066	0.112	0.033	0.259	0.004	0.048
	disc	0.153	0.262	0.069	0.450	0.031	0.083
MST	all	0.055	0.088	0.045	0.082	0.018	0.032
	disc	0.124	0.193	0.084	0.114	0.049	**0.049**
pMST	all	**0.051**	**0.087**	0.034	**0.079**	**0.003**	**0.031**
	disc	**0.112**	**0.192**	**0.056**	**0.108**	0.028	0.050

zero support to the target pixel to be filtered out. The traditional methods above are easy to understand and implement, but they ignore the fact that the pixels inside a fixed window may belong to different depth region, and if we use the pixels' depth value directly, there would be artifacts in the final depth map. Deferent from traditional methods, by cutting the weight propagation on different paths, our method can estimate weight support regions with arbitrary shapes and sizes, which is an adaptive scheme to avoid the artifacts in the depth map maximally.

0058

IV. EXPERIMENTAL RESULTS

We present experiment on the standard Middlebury datasets [9, 10]. We draw a comparison of our method with other three existing local filtering methods. We implement the algorithm with an Intel Core i5 2.3-GHz CPU and 16-GB RAM. We will show the quantitative and visual evaluation of depth upsampling results.

A. Quantitative evaluation

We compare our method with four other methods: JBU [1], LBF [2], NA-BF [4] and MST [5] respectively. The test images are Teddy, Cones and Venus of Middlebury datasets. We use the left disparity map as the ground truth depth image. For each test depth image, the low resolution depth map is obtained by downsampling the high resolution depth map. We empirically set the parameters as follows: σ=0.05, threshold ε=5,$\tau_1 = 0.5$, $\tau_2 = 10$. Other methods' parameters are determined by the corresponding paper's default parameter settings. We apply the percentage of bad matching pixels (PBP) [9] as the evaluation metrics for assessing the performance of the algorithm. We report the PBP at discontinuity regions and the whole image. The discontinuity regions are obtained by dilating 1-pixel wide range of ground truth.

The results of different methods with different upsampling factors are listed in Table 1. In general, our method achieves the best performance among all of the algorithms. Thanks to the non-local tree structure and the prior map, we find that our method outperforms other methods at DISC indices except the "Venus" dataset, which reflects the edge preserving ability. The "Venus" dataset has most smoothing regions with few sharp edges, and it cannot test the algorithm's edge preserving ability well. Nevertheless, our method archives the best overall performance of PBP on "Venus" dataset.

B. Visual evaluation

We compare our results with others on Middlebury dataset as shown in Fig.5 for visual performance. From the result, we can see that for traditional methods as bilateral filter [1] based on the fixed size window, if the window size is less than the depth hole, it cannot fill the hole nicely, just as LBF [2]. As for the traditional depth upsampling method based on MST [4], there would be artifacts around the depth edges, such as depth bleeding, due to its' mixed different depth regions. Our method, however, reduces the depth artifacts greatly because of the edge-preserving weighting scheme adopted in the prior MST's construction. Besides, benefitting from the implicit image segmentation provided by the prior map, our method is better at preserving small thin image structures (Fig.5 (c)). Apart from that, the prior map gives a metrics on the similarity between color and depth edges, which alleviates the texture copying artifacts in depth maps.

V. CONCLUSION AND FUTURE WORK

In this paper, we present a novel prior minimum spanning tree to upsample the depth map. And the experimental results show that our work improves the performance of the original depth upsampling algorithm. Given the low resolution depth map and the pre-aligned color image, we first obtain the prior map and then integrate the high resolution color image and prior map to reconstruct the MST to upsample the depth map. We have validated our approach on the standard Middlebury dataset, yielding the significant improvements of our method.

Although our approach achieves better results than other local methods, it dependents on many parameters. In the future work, the image segmentation will be expressed explicitly to upsample the depth map. A joint image segmentation and depth upsampling framework is expected to achieve promising result.

ACKNOWLEDGMENT

The work in this paper is supported by National Natural Science Foundation of China (Grant No. 71171121/61033005) and National 863 High Technology Research and Development Program of China (Grant No. 2012AA09A408).

REFERENCES

[1] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," ACM SIGGRAPH, 2007.

[2] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth super resolution for range images. In CVPR, 2007

[3] D. Chan, H. Buisman, C. Theobalt, and S. Thrun. A noise-aware filter for real-time depth upsampling[C] //Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008. 2008. F. De Lillo, F. Cecconi, G. Lacorata, A. Vulpiani, EPL, 84 (2008)

[4] Q. Yang. Stereo matching using tree filtering. Pattern Analysis and Machine Intelligence. IEEE Transactions on 37(4)(2015), 834-846.

[5] J. Diebel, S. Thrun. An application of Markov random fields to range sensing, in: Proceedings of Advanced Neural Information Processing Systems, 2005, pp. 291–298

[6] K. Lo, Y. F. Wang, K. H, "Joint Trilateral Filtering For Depth Map Super-Resolution," VCIP 2013.

[7] Q. Yang. "A non-local cost aggregation method for stereo matching. " Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.

[8] M.-Y. Liu, O. Tuzel, Y. Taguchi, Joint geodesic upsampling of depth images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 169–176

[9] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo corresponsence algorithms. Internatinal Journal of Computer Vision(IJCV), 47(1/2/3) : 7-42, April-June 2002.

[10] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), volume 1, pages 195-202, Madison, WI, June 2003.