A REVISIT TO MRF-BASED DEPTH MAP SUPER-RESOLUTION AND ENHANCEMENT

Jiangbo Lu[*], *Dongbo Min*[*], *Ramanpreet Singh Pahwa*[†], *and Minh N. Do*[†]

[*]Advanced Digital Sciences Center, Singapore
[†]University of Illinois at Urbana-Champaign, USA
[*]{jiangbo.lu, dongbo}@adsc.com.sg, [†]{pahwa1, minhdo}@illinois.edu

ABSTRACT

This paper presents a Markov Random Field (MRF)-based approach for depth map super-resolution and enhancement. Given a low-resolution or moderate quality depth map, we study the problem of enhancing its resolution or quality with a registered high-resolution color image. Different from the previous methods, this MRF-based approach is based on a novel data term formulation that fits well to the unique characteristics of depth maps. We also discuss a few important design choices that boost the performance of general MRF-based methods. Experimental results show that our proposed approach achieves high-resolution depth maps at more desirable quality, both qualitatively and quantitatively. It can also be applied to enhance the depth maps derived with state-of-the-art stereo methods, resulting in the raised ranking based on the Middlebury benchmark.

Index Terms— Time-of-flight sensor, MRFs, depth super-resolution, depth-enhancement, global optimization

1. INTRODUCTION

Accurate depth at high resolution is required in many applications such as interactive view interpolation, 3D television, robot navigation. Unlike obtaining high-resolution color images, the acquisition process of an accurate depth map at high resolution is never trivial, e.g., laser range scanners or active illumination with structured lights. However, these accurate depth measurement techniques are only applicable to the static environments.

For the purpose of acquiring depth maps for dynamic scenes at video rate, passive stereo and recent active depth sensors based on the time-of-flight (TOF) principle [1] are actually the primary choices. Unfortunately, the quality of depth maps obtained with these techniques is often not at a level desired by the high-level applications, due to inherent physical limitations or real-life constraints. For instance, depth maps returned by TOF sensors are typically of low resolution and also noisy, e.g., 176×144 for Mesa Imaging SR4000 [1]. Depth maps estimated by stereo algorithms are, however, not accurate enough, especially when they are estimated under real-time constraints. This paper hence focuses on a post-processing step that enhances the resolution or quality of a given non-ideal depth map.

Research supported by the Advanced Digital Sciences Center (ADSC) under a grant from the Agency for Science, Technology and Research of Singapore. ADSC is a related organization in Singapore of the College of Engineering at the University of Illinois at Urbana-Champaign.

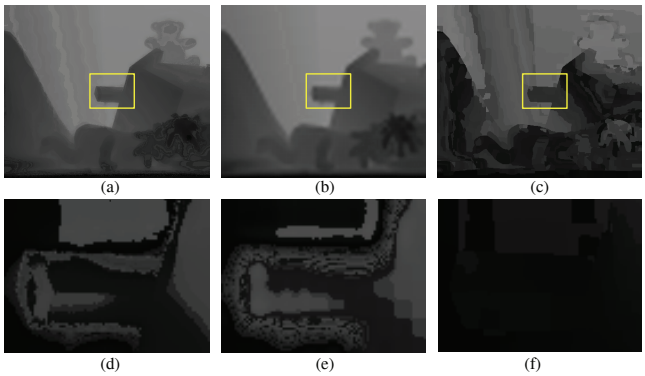

Fig. 1. Depth map super-resolution on the same input low-resolution depth map with the upsampling factor of 8. Depth maps generated by (a) Yang *et al*. [3] (b) Diebel and Thrun [4] (c) Our proposed method. (d)-(f) Close-ups of (a)-(c).

Different from some existing work on probabilistic fusion of active and passive sensors [2], our approach takes only one color image in addition to the depth map to be processed. The motivations are two-fold. First, we want to investigate how far we can go with depth enhancement using only one single high-resolution color image, which has not been adequately addressed so far. Second, for the application scenarios where network bandwidth and processing power are constrained at the content capturing end, the input to the post-processing module may only contain a single-view color image plus a registered depth map either estimated by passive stereo under the stringent real-time constraint or measured by a TOF sensor.

Specifically for this one color image plus one depth map setup, Yang *et al*. [3] used an iterative joint bilateral filtering scheme to build a cost volume for the final depth hypothesis selection. Though this method better preserves the depth discontinuities than other previous methods that directly apply a joint bilateral filter to the depth image [5], it does not explicitly enforce the spatial smoothness constraint when searching through all the depth hypotheses. Without a global regulation term, such a soft-weighting based local method still makes the resulting depth map fuzzy, especially along depth discontinuities [see Figure 1(a)]. Diebel and Thrun [4] have instead proposed using a MRF method to fuse low-resolution depth maps with high-resolution color images. Depth value assignment is modeled as

an energy minimization problem based on MRF formulations, and the authors used the quadratic distance in both data term and smoothness term in the energy function. This choice enables using the conjugate gradient algorithm for the least square optimization problem, but it results in much worse solution quality than what is typically expected for a MRF-based approach [6], as shown in Figure 1(b).

Therefore, the first goal of this paper is to revisit the MRF-based depth enhancement problem. We formulate the energy function more rigorously in Section 2, and solve it with loopy belief propagation (LBP) [7]. Secondly, a novel scheme is proposed to construct the data term in Section 3 allowing for pixel-wise adaptive selection of an appropriate depth reference value.

2. MRF FORMULATION FOR DEPTH INFERENCE

Given a depth map D^0, which is potentially to be up-sampled by a factor of s for each image dimension, our goal is to enhance its resolution and/or quality by using an aligned high-resolution color image I. The resulting enhanced depth image at the same resolution of I is denoted as $D = \{d_p\}$, where $p = (i, j)$ is an integer pixel on a 2D image grid.

We formulate the depth super-resolution and enhancement problem using a MRF-based depth inference framework in a unified manner. MRF is an elegant model that can be justified in terms of maximum a posteriori estimation [6]. The basic idea is to assign a depth label to every pixel p, which is the most likely estimate of the unknown true depth value, given the coarse map $D^0 = \{d_p^0\}$ and I. In the MRF framework, such a pixel-labeling problem is solved by minimizing the Gibbs energy E on a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$:

$$E = \sum_{p \in \mathcal{V}} U(d_p) + \lambda \sum_{(p,q) \in \mathcal{E}} V(d_p, d_q) \ , \tag{1}$$

where $\mathcal{V}$ is the set of all pixels, $\mathcal{E}$ is the set of graph edges connecting adjacent pixels. $U(d_p)$, known as the data energy, measures similarity between the estimated depth value and the input depth value from D^0. $V(d_p, d_q)$, known as the smoothness energy, is a regularization term that encourages neighboring pixels to have similar depths. Here we use the standard 4-connected neighborhood system. λ is a parameter to balance the two terms.

Different functional forms can be used to define the data term U and the smoothness term V, but they eventually lead to very different results in terms of solution quality and efficiency [6]. Rather than using the quadratic distance [4], we propose to use robust distance functions such as truncated absolute difference (TAD) in U and V. TAD is good at preserving discontinuities, and can also be efficiently computed via distance transform in LBP. More specifically, given a truncation threshold σ,

$$U(d_p) = \min(|d_p - d_p^0|, \sigma) \ . \tag{2}$$

When the upsampling factor $s > 1$, the input depth map D^0 is of a low resolution with regard to I. Therefore, (2) is only applied to a sparse grid of pixels $\mathcal{V}^s = \{(i, j) | i\%s = (s - s\%2)/2, j\%s = (s - s\%2)/2\}$. We set the data energy of the rest of the pixels, i.e., $p \in \mathcal{V} \backslash \mathcal{V}^s$, to zero. We have also performed experiments to test other image upsampling strategies to upsample D^0 to the full resolution of I, such as bicubic interpolation and bilinear interpolation. Consistent with the findings in [8], we find that the nearest neighbor sampling is preferable, which does not introduce unwanted blurring due to interpolated sampling. In regular depth enhancement, when $s = 1$, then $\mathcal{V}^s = \mathcal{V}$, meaning that the data cost is computed with (2) for all the pixels.

Next, we define the smoothness term V as follows,

$$V(d_p, d_q) = w_{p,q} \cdot \min(|d_p - d_q|, \tau) \ . \tag{3}$$

where τ is a truncation threshold. $w_{p,q}$ is a spatially varying weight defined using the color image I and the parameter γ,

$$w_{p,q} = \exp(-\Delta I_{p,q}/\gamma) \ . \tag{4}$$

$w_{p,q}$ serves as a soft constraint that encourages the depth discontinuities to be aligned with color edges. Unlike the previous method [3] calculating $w_{p,q}$ based on the average absolute difference of different color channels, we define $\Delta I_{p,q}$ to be the maximum absolute difference among the three color channels between the pixel p and q. With such a change, the case when the average absolute difference is mild but one of the color channels is distinctive is more appropriately handled.

3. PROPOSED DATA TERM CONSTRUCTION

The MRF formulation proposed in Section 2 generally works well for depth super-resolution and enhancement. Nevertheless, sometimes it does not reconstruct very accurate depth values for depth discontinuities. The main reason is that the data cost $U(d_p)$ associated with a depth hypothesis d_p at pixel p is solely defined with respect to d_p^0, whereas d_p^0 is known to be less reliable along depth discontinuities. This observation holds for the coarse depth map measured by TOF sensors, as well as the one estimated from passive stereo algorithms. In this section, a novel method is proposed to construct the data term.

Our key idea is to limit the negative impact of inaccurate depth estimates or measurements over the depth inference. Therefore, rather than always trusting d_p^0 as a good reference value when defining $U(d_p)$, we also include the initial depth values $N_d = \{d_{q_0}^0, d_{q_1}^0, d_{q_2}^0, d_{q_3}^0\}$ of neighboring pixels $N = \{q_0, q_1, q_2, q_3\}$ for the pixel p. As shown in Figure 2(a), q_i is of a distance of s pixels to the pixel p under consideration, when measured according to the full resolution grid. The rationale is that the current depth hypothesis d_p does not necessarily need to fit well to d_p^0 (if it is likely unreliable), but instead it is allowed to decide the best reference depth value d_p^* from a union of N_d and d_p^0. Conceptually, this idea is similar to shifted windows or non-centered windows used to improve performance at object boundaries in stereo matching.

Now, we present how to decide the depth reference value d_p^* adaptively for different pixels p, which will replace d_p^0 in (2). Considering possible geometric configurations around p, we explicitly tackle four different conditions. Let $h_p = |d_{q_0}^0 - d_{q_2}^0|$ and $v_p = |d_{q_1}^0 - d_{q_3}^0|$ denote the absolute difference between

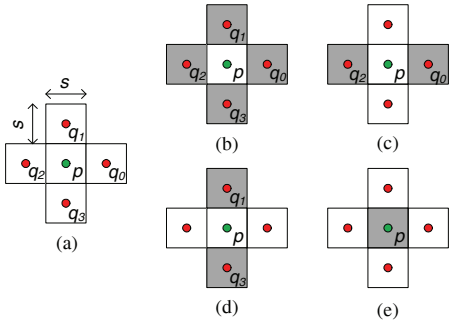

Fig. 2. Proposed locally adaptive selection of the depth reference value for the anchor pixel p. (a) Besides p, four depth measurements of the neighboring pixels $\{q_0, q_1, q_2, q_3\}$ are also considered. (b) p is near general depth discontinuities. (c, d) p is near an approximately vertical or horizontal depth edge. (e) p is in homogeneous depth regions or indefinitive regions. Shaded blocks denote those neighboring depth values that d_p^* can be selected from, in addition to d_p^0.

horizontal and vertical neighbors, respectively. Based on the strength of h_p and v_p, the four cases are categorized as follows.

1) *p is near general depth discontinuities* [Fig. 2(b)]. This corresponds to the case when $h_p > t_l$ and $v_p > t_l$. In this case, $\forall d_{q_i}^0 \in \{d_{q_0}^0, d_{q_1}^0, d_{q_2}^0, d_{q_3}^0\}$, if $|d_{q_i}^0 - d_p^0| \leq t_s$, then we set $d_p^* = d_{q_i}^0$. This means if one neighboring depth measurement can support the center depth value, then we trust the initial depth measurement more, and use it for d_p^*. Otherwise, d_p^0 is very likely to be a transitional depth value (mixing foreground and background depths), so we rely on the color cue to disambiguate which depth value to select to define d_p^*. The pixel that has the closest average color distance to the center pixel p is selected as the winner, so its depth value is assigned to d_p^*. Here we first apply a 3×3 median filter to I, suppressing the impact of image noise as well as subtle non-Lambertian effects.

2) *p is near an approximately vertical depth edge* [Fig. 2(c)]. This corresponds to the case when $h_p > t_l$ and $v_p \leq t_s$. In this case, d_p^* is selected from $\{d_p^0, d_{q_0}^0, d_{q_2}^0\}$, using the same strategy discussed above, i.e., depth proximity check followed by color distance check.

3) *p is near an approximately horizontal depth edge* [Fig. 2(d)]. This is the case when $h_p \leq t_s$ and $v_p > t_l$. Similar to the case 2), d_p^* is selected from $\{d_p^0, d_{q_1}^0, d_{q_3}^0\}$.

4) *p is in homogeneous depth regions or indefinitive regions* [Fig. 2(e)]. This corresponds to all the other cases not covered in the previous classes. In this case, $d_p^* = d_p^0$, and the scheme degenerates to the original data term construction in (2).

Figure 3 illustrates the proposed idea when applied to super-resolve the *Teddy* depth map with $s = 8$. Appropriate depth values can be decided to define d_p^* for the anchor pixel p, adapting to difference local scene structures. Based on the Middlebury stereo evaluation [9], it is found that the proposed new data term construction performs better than the baseline method of (2). With little overhead, the error rates for the *non-occluded*, *all*, and *depth discontinuities* regions for the given *Teddy* image are reduced by 1.5%, 1.5%, and 4.3%, respectively.

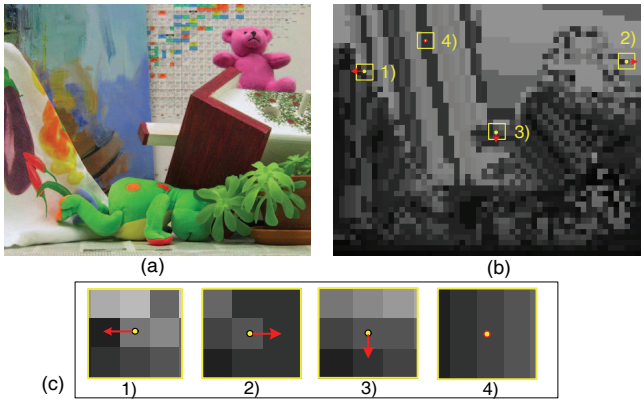

Fig. 3. (a) Full-resolution color image I. (b) Upsampled depth map with a factor of 8 using the nearest neighbor sampling. Four different pixel locations have been shown, corresponding to the four cases discussed, i.e., 1) near general depth discontinuities, 2-3) near a vertical/horizontal depth edge, and 4) in homogeneous regions. The red arrow indicates the neighboring pixel's depth value that has been selected. (c) Close-ups of (b).

4. EXPERIMENTAL RESULTS

We have implemented the proposed MRF formulation and the new data function based on the Middlebury MRF minimization source code. The LBP software [7] is used to minimize the global energy function in (1). To enable quantitative evaluation of different techniques based on the ground-truth depth maps, the Middlebury dataset [9] has been used in our experiments. Some parameters in the proposed approach are empirically set as follows, $\gamma = 20.0$, $\lambda = 2$, $t_s = 4$. For *Tsukuba* and *Venus* images containing a relatively small number of intermediate depth layers (normalized to grayscale), we set σ, τ, and t_l all to 64, while for *Teddy* and *Cones* images, σ, τ, and t_l are all set to 32.

First, we compare different depth super-resolution methods by using the same set of input low-resolution depth maps [3], with a upscaling factor s of 8. Based on the Middlebury online evaluation, Table 1 reports the disparity error rates measured against the ground-truth disparities. Compared to the previous methods such as the prior MRF formulation [4] and iterative joint bilateral filter (JBF) [3], our proposed methods have significantly improved the depth map accuracy, particularly for challenging depth discontinuities. Also, *our method-2* presented in Section 3 improves the overall performance over *our method-1* presented in Section 2, thanks to the new data term construction scheme. A visual comparison of the depth maps is given in Figure 1. Clearly, our method-2 yields piecewise smooth depth maps with cleaner and sharper depth edges.

We have also compared the performance of our method-1 and our method-2 with low-resolution depth maps of different blurring effects, which were created by varying the standard deviation of the Gaussian kernel. We find that the improvement in depth map accuracy due to applying our method-2 is more pronounced, when the blurring effect is strong or the upsampling factor s is big (as shown in Figure 3). In fact, even when

Table 1. Quantitative evaluation results for the Middlebury stereo database [9]. The image upscaling factor is 8.

Algorithm	*Tsukuba*			*Venus*			*Teddy*			*Cones*		
	nonocc.	all	disc.	nonocc.	all	disc.	nonocc.	all	disc.	nonocc.	all	disc.
Before refinement	8.14	9.74	43.4	2.18	2.79	30.5	13.7	14.7	41.9	12.0	15.1	34.1
MRF-previous [4]	8.20	9.74	44.0	2.11	2.69	29.6	13.5	14.5	41.5	11.8	14.9	33.8
Iterative JBF [3]	6.27	7.23	33.6	1.20	1.50	16.7	10.7	11.5	32.1	8.83	11.0	25.3
Our method-1	4.86	5.60	24.4	0.79	1.00	10.5	9.26	9.96	25.9	8.70	11.4	24.3
Our method-2	4.35	5.09	22.2	0.79	1.00	10.5	9.33	9.87	26.3	8.79	11.3	24.5

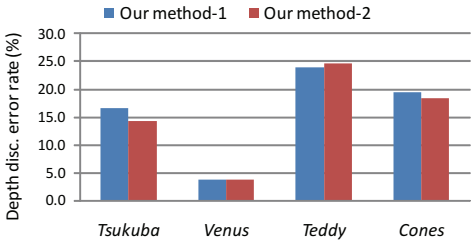

Fig. 4. Comparison between our method-1 and our method-2 near depth discontinuities regions for an upscaling factor of 4.

the standard deviation is as low as 1.5, we can still observe an appreciable accuracy improvement, as shown in Figure 4.

Lastly, we show that our methods can also enhance the quality of a depth map estimated by stereo algorithms, taking a left-view color image as the input at the post-processing end. Figure 5 shows that our method-2 corrects the matching errors and also tackles the foreground over-fattening effects seen in the original depth maps, which are generated by our previous GPU-based stereo method under the real-time constraint [10]. Our experiments also show that the proposed technique can improve the global optimization based stereo methods. For instance, it raises the Middlebury rank of *RealTimeBP* [9] by 11 slots, and it further enhances the depth accuracy achieved by the top-ranking *AdaptingBP* method [9].

5. CONCLUSIONS

In this paper, we have studied the MRF-based depth super-resolution and enhancement problem, given a coarse-resolution or coarse-quality depth map plus one registered full-resolution color image. Different from the previous methods, we have formulated the energy function for depth inference more rigorously in the MRF framework. Then, we further propose a novel data term construction scheme, which allows pixelwise adaptive selection of an appropriate depth reference value. Experimental results illustrate the effectiveness of our methods when compared to traditional approaches in enhancing the resolution and quality of the input depth maps. The proposed method also improves the depth maps estimated by existing stereo algorithms noticeably, when applied as a post-processing step. With the same MRF-based depth inference framework, our method can also work against additive image noise. Our future work includes approximating and accelerating BP-based depth inference on GPUs, as BP is known to be friendly for fast parallel processing. We also plan to model the depth image formation process for TOF sensors, and such a model will be useful for defining the energy function. Occlusion and non-perfect registration between color and depth sensors will also be tackled.

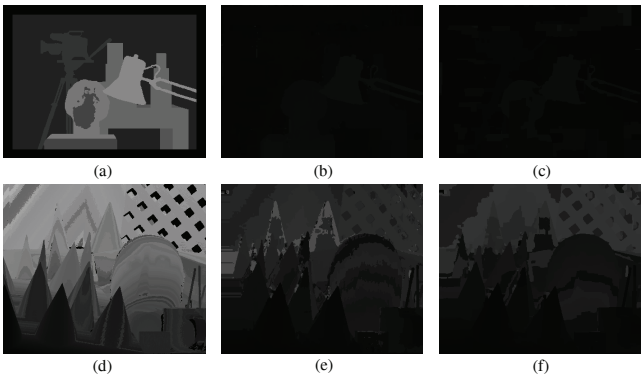

Fig. 5. Improving existing stereo results by our proposed method. From left to right: group-truth depth map, depth map by *RealTimeCross* [10], and improved depth map with our method. Top row: *Tsukuba* images. Bottom row: *Cones* images.

6. REFERENCES

[1] Mesa Imaging AG, http://www.mesa-imaging.ch/.

[2] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *CVPR*, 2008.

[3] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *CVPR*, 2007.

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

[5] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," in *ACM SIGGRAGH*, 2007.

[6] R. Szeliski et al., "A comparative study of energy minimization methods for markov random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068–1080, Jun. 2008.

[7] M. F. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters," in *ICCV*, 2003, pp. 900–907.

[8] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "LidarBoost: depth superresolution for ToF 3D shape scanning," in *CVPR'09*.

[9] Middlebury stereo page, http://vision.middlebury.edu/stereo/.

[10] J. Lu, K. Zhang, G. Lafruit, and F. Catthoor, "Real-time stereo matching: a cross-based local approach," in *ICASSP*, 2009.