DEPTH ENHANCEMENT BASED ON HYBRID GEOMETRIC HOLE FILLING STRATEGY

Lu Sheng, King Ngi Ngan

Department of Electronic Engineering, The Chinese University of Hong Kong

ABSTRACT

Depth map is a crucial component in various 3D applications. However, current available depth maps usually suffer from low resolution, high noise, random and structural depth missing problems due to theoretical, systematic or hardware limitations. In this paper, we propose a novel method to enhance depth map with the guidance of aligned color image, tackling these problems in a whole framework, where a hybrid strategy on filling hole geometrically by the combination of joint bilateral filtering and segment-based surface structure propagation is introduced. Our experimental results prove the proposed method outperforms existing methods.

***Index Terms*—** Depth map, depth enhancement, hole filling, segmentation, filtering

1. INTRODUCTION

Depth acquisition technique is quite popular in recent years with the prosperity of various 3D applications in manufacture and entertainment industry. For example, virtual reality, 3D movies, game controller and so on. Recently a variety of systems have been proposed to acquire scene's depth, such as stereo vision system, real-time structured-light depth sensor (e.g., Kinect), Time-of-Flight camera or laser camera. Unfortunately most of them suffer from low quality of acquired depth maps, which typically refers to low resolution, high noise and missing regions without depth values. These drawbacks obstruct the direct usage of depth information in captured scene for different 3D applications.

Generally, errors of depth maps can be roughly classified into three categories: *Missing regions* around occlusion, texture-less regions and non-Lambertian surface, *Mismatching errors* between depth map and color image and *Random Noise*.

Due to these limitations, depth map enhancement has been extensively studied. The most studied work is the upsampling and smoothing problem. A pioneer work in this field is done by Diebel *et al.* [1]. They model it as a Markov Random Field (MRF) with the assumptions that **a)** *discontinuities in color image and corresponding depth map should be co-aligned*, and **b)** *pixels with similar texture should have similar depth.* These assumptions are reasonable and also valid in our work. Under similar fashion, many researchers

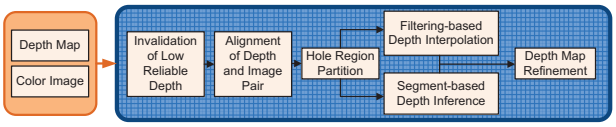


Fig. 1. Framework of proposed method.

[2, 3] also use MRF model or Auto-Regression model to upsample and/or smooth depth surface while enhance the discontinuities. Difference among their works mainly come from the design of smooth or regularization terms in objective function. But such kind of energy minimization methods are always computationally expensive, which is not reasonable in real-time applications.

Following similar assumptions, Kopf *et al.* [4] proposed Joint Bilateral Upsampling (JBU), which is an extension of famous bilateral filter and is fast and effective on upsampling low resolution or smoothing noisy depth maps. To solve problems occur in JBU, e.g., texture copying and edge blurring, and enforce its power, many modified filters have been proposed [5, 6, 7].

The missing region filling problem is related to image inpainting and occlusion handling in stereo vision. According to recent work of Richardt *et al.* [8], standard joint bilateral filter (JBF) can efficiently fill holes, but it is easy to produce artifact when the hole is too large. And filtering cannot preserve surface structure, because the extrapolated depth will always be piece-wise constant in a large hole. Many stereo algorithms simply fill the holes from the background, which all suffer from significant artifacts when the scene is complex. Our work on hole filling is related to work of Wang *et al.* [9] about stereoscopic inpainting, they over-segment stereo images and fit a plane to each segment with estimated disparity, then propagate plane into holes by matching segments in a greedy way. Their segment matching cost function heavily relies on stereo images that could not be exploited in general case. And plane fitting is not precise enough to estimate local surface structure.

This paper tries to tackle these problems together with a special treatment to large hole filling problem. In particular, we assume holes come from depth upsampling, unreliable depth removal along with missing regions. In the first step, we invalidate and remove unreliable depth, then align the depth map with the color image and map it into image's coordinate. In the second step, we propose a hybrid strategy to fill in the

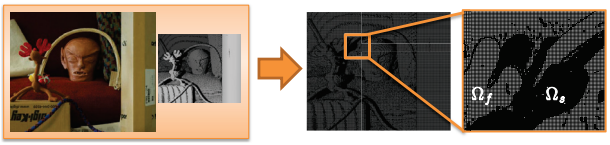


Fig. 2. Align the depth map into color image coordinate and then partition the hole region into Ω_s and Ω_f. Test depth map comes from Middlebury dataset.

hole. After that, a standard joint bilateral filter is applied to refine the depth map. Overall framework is shown in Fig 1.

2. PROPOSED METHOD

We take an image I and its corresponding depth map D as inputs. Define the set of invisible (hole) pixels as Ω, and the set of visible pixels as Ψ.

2.1. Unreliable Region Detection and Invalidation

Before transforming the depth map into image's coordinate, we need to invalidate unreliable depth pixels. The reliability can be measured by the depth gradient as mentioned in [8], because unreliable pixels always occur along the depth discontinuities or in a neighbourhood with high depth variance according to the fact that depth camera cannot accurately capture depth in such regions. What's more, most real-time depth sensors have the mismatching errors between color and depth edges due to calibration error between color camera and depth sensor. Invalidating such kinds of low reliable regions and filling in depth with the guidance of image will diminish the edge mismatching problem and increase the depth value reliability.

In detail, sobel approximation is applied to compute the depth gradient, while we invalidate pixels that have larger gradient value than a given threshold τ.

2.2. Hybrid Strategy of Geometric Hole Filling

After invalidating the unreliable regions, and transforming it into color image's coordinate, the resultant depth map contains three types of holes in the depth map: holes from occlusion and/or specular regions Ω_o, invalidation Ω_d and sparse upsampling Ω_u. Therefore, we define the hole set Ω as

$$\Omega = \{p \mid p \in \Omega_o \cup \Omega_u \cup \Omega_d\}, \tag{1}$$

where p indicates pixel coordinate. Our proposed hybrid strategy is a combination of filtering and surface structure propagation. Filtering-based approaches are quite efficient to interpolate depth values if the hole region is small, but it will possibly fail when dealing with large holes. However, we can exploit the widely used segment constraint [9] to infer the structure, i.e., segment a hole and its neighbours into several small patches according to the guided color image, and we assume each patch has a smooth surface structure without sudden depth variation. Then a patch with enough depth samples can be modelled by a plane or curved surface and it is reasonable to propagate its surface parameters into neighbor patches under similar texture in the hole.

Our hole filling process firstly partitions hole set Ω into two subset Ω_f and Ω_s, and then employs depth interpolation in Ω_f and depth inference in Ω_s, see Fig. 2.

2.2.1. *Hole Region Partition*

A pixel q is considered in the region Ω_f when its local $w \times w$ window has enough informative samples to interpolate its depth. We dilate visible region Ψ by a square of width w, as $\Psi_w = \text{Dilation}(\Psi, w)$, then pixel $p \in \Psi_w \cap \Omega$ will always have one sample at least. Then the set Ω_s and Ω_f are

$$\Omega_s = \text{Dilation}\left(\left(\Omega - \Psi_w \cap \Omega\right), w\right) \cap \Omega \tag{2}$$

$$\Omega_f = \Omega - \Omega_s \tag{3}$$

The dilation operation in eqn 2 is to safely exclude pixels that have insufficient depth samples in their neighbor from Ω_f.

2.2.2. *Depth Interpolation by filtering*

To fill Ω_f, a standard joint bilateral filtering [4] is utilized. For each pixel $p \in \Omega_f$, and its visible local neighbours $q \in \mathcal{N}_p \cap \Psi$ in a $w \times w$ window, its estimated depth is

$$D_p = \frac{1}{N_p} \sum_{q \in \mathcal{N}_p \cap \Psi} \mathcal{G}_s\left(p, q\right) \mathcal{G}_r\left(I_p, I_q\right) D_q \tag{4}$$

where $\mathcal{G}_s$ and $\mathcal{G}_r$ are Gaussian kernel functions with standard deviations σ_s and σ_r, measuring the spatial similarity and range (color) similarity, respectively. N_p is the normalization factor that ensures the summation of weights is equal to zero.

2.2.3. *Depth Inference under segment constraint*

Many successful super-pixel segmentation methods have been published recently, in this application we use a fast method called *simple linear iterative clustering* (SLIC) [10] to group pixels into a set of color patches, in which pixels share similar color or texture. Then patches that overlap Ω_s will be sorted into two sets $\mathcal{S}_v$ and $\mathcal{S}_u$, where $\mathcal{S}_v$ means set where each patch has enough visible pixels (e.g., more than 50%) to infer its surface structure and patches in $\mathcal{S}_u$ are not.

Surface model estimation for patches in $\mathcal{S}_v$. For simplicity, we can just model the surface by

$$D(u, v) = a_0 + a_1 u + a_2 v, \quad \text{or} \tag{5}$$

$$D(u, v) = a_0 + a_1 u + a_2 v + a_3 u^2 + a_4 v^2 + a_5 uv \tag{6}$$

where eqn 5 is the linear form, and eqn 6 is the quadratic form. We use RANSAC to robustly estimate each patch's surface

model. What's more, for the sake of accuracy, we can alternatively transform the depth map into 3D metric coordinate (X, Y, Z), and model function $Z(X, Y)$ under a similar way. In this case, recovering pixel p's depth is to find the intersection of the surface and the line-of-sight along p.

After estimating surface models for patches in set $\mathcal{S}_v$, their invisible pixels can be efficiently inferred. At the same time, the surface models of visible patches are also estimated. We can further refine them by merging patches with similar surface structure, and then re-calculate their surface models.

Surface propagation for patches in $\mathcal{S}_u$. It turns out to be a patch matching problem. Here we propose a greedy algorithm that robustly find two most similar patches according to a novel matching cost.

Our algorithm firstly selects candidate patches set $\mathcal{C}_{\mathcal{S}_u} = \{P_v\}$ against $\mathcal{S}_u$, where P_v has an estimated surface model and it is chosen near hole S_u, because surface structure will be more consistent and reliable near the hole boundary. Thus filling process will be under an order from outer to inner patches. In each iteration, find the best matched patch in $\mathcal{C}_{\mathcal{S}_u}$ and assign its surface model to query patch P_u and fill in the depth, then P_u will be added into $\mathcal{C}_{\mathcal{S}_u}$. This process will continue until all patches in $\mathcal{S}_u$ is filled.

Given patch $P_v \in \mathcal{C}_{\mathcal{S}_u}$, and P_u in the set $\mathcal{S}_u$, we want to measure their similarity. Since each patch has arbitrary shape, then commonly used MSE is inapplicable, and mean intensity is not enough distinctive. Our proposed method randomly selects n pixels in P_u as $p_u^j, j = 1, \ldots, n$ and k pixels in P_v as $p_v^i, i = 1, \ldots, k$, and defines a $m \times m$ square-sized sub-patch to each selected pixel, as B_v^i in P_v and B_u^j in P_u respectively. If two patches are similar, their sub-patch matching cost should be minimal. Sub-patch matching is valid because it considers the color and spatial distributions of texture while is able to handle patch with arbitrary shape.

To robustly estimate their similarity and not introduce mismatch, we propose a shape-adapted sum-of-square to measure the similarity between B_v^i and B_u^j.

$$\mathcal{E}_{B_u^j}\left(B_v^i\right) = \frac{\|\mathcal{K}_v^i \circ \left(B_v^i - B_u^j\right)\|_F^2}{N_v^i} + \frac{\|\mathcal{K}_u^j \circ \left(B_v^i - B_u^j\right)\|_F^2}{N_u^j} \tag{7}$$

where $\mathcal{K}_v^i$ and $\mathcal{K}_u^j$ are bilateral kernels centred at pixel p_v^i and p_u^j, which are similar as eqn 4, measuring the color similarity and spatial similarity of the center pixel against its neighbours. $\circ$ represents element-wise multiplication. N_v^i and N_u^j are normalization factors similar in section 2.2.2. Then cost between B_u^j and patch P_v is $E_{B_u^j}(P_v) = \frac{1}{k}\sum_{i=1}^{k} \mathcal{E}_{B_u^j}(B_v^i)$.

Therefore, given $\mathcal{C}_{\mathcal{S}_u}$ and a query patch P_u, to each B_u^j in P_u, we can find the best patch P_{v^*} that has the smallest cost. Then we can form a histogram that each bin indicates a candidate patch, whose bin value is the number of sub-patches in P_u that matches referred candidate patch. Then the bin with largest value refers to the most similar patch. We normalize the histogram and denote it as $H_{P_u}(P_v)$, where $P_v \in \mathcal{C}_{\mathcal{S}_u}$.

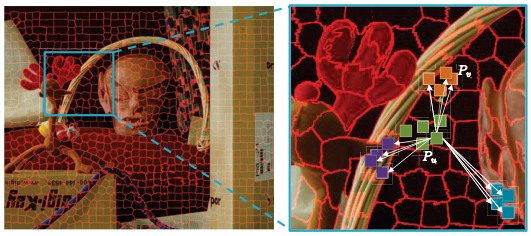


Fig. 3. Illustration of patch matching process.The left image is segmented color image, the right one is a close-up of local region marked blue in left image. P_u is query patch, P_v is in candidate patch sets. Detail description is in the text.

It is possible to find more than one patches that similar in color, we further add spatial constraint into our framework. In detail, we measure the Euclidean distance between center pixels of two patches $d(P_u, P_v)$, and normalize the distance by exponential function, then the overall cost function is

$$\mathcal{T}_{P_u}(P_v) = H_{P_v}(P_u) \cdot \exp\left(-d(P_u, P_v)^2 / \left(2 \times \sigma_d^2\right)\right) \tag{8}$$

The maximum value of $\mathcal{T}_{P_u}(P_v)$ shows the optimal patch pair. Because patches under similar texture may have different surface structures, just choose the best matched patch may inevitably introduce errors. To eliminate it, we fill the query patch from the most similar one to the least one. Once the filled patch is consistent with local neighbours, this process will stop.

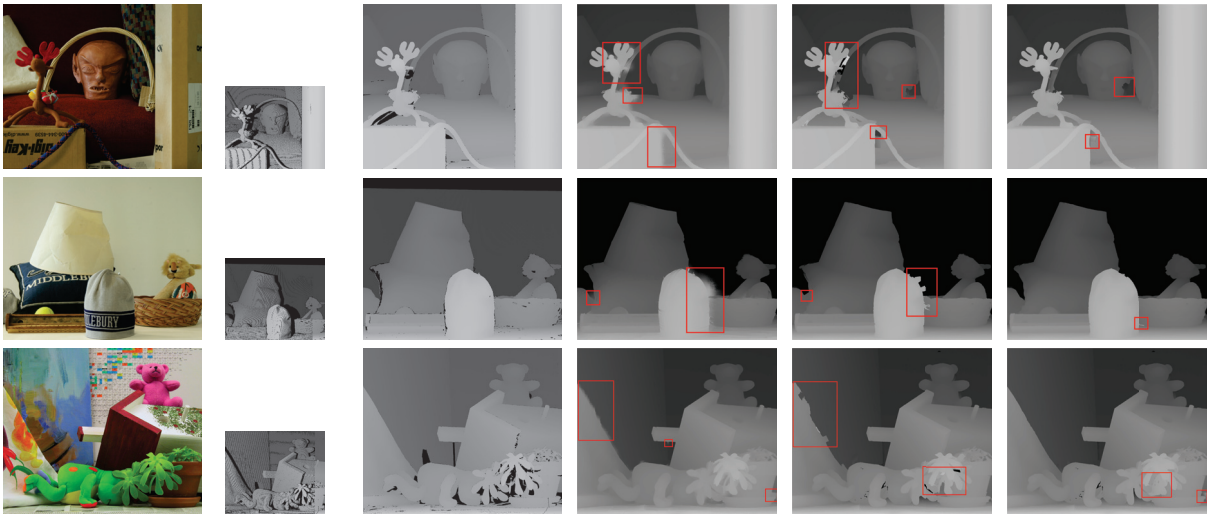
2.2.4. Depth Map Refinement

After filling in all the missed pixels in depth map, we can further refine it to reduce noise and artifact, as well as enhance depth structure according to the guided image. Recently we find standard joint bilateral filtering is sufficient to provide effective and efficient results.

3. EXPERIMENTS

In this section, we evaluate the performance of our proposed algorithm, and compare it with other existing methods. Since the main contribution of our work is the hole filling strategy, we compare its performance with other hole filling methods, e.g., algorithms presented by Richardt *et al.*[8] so-called multi-resolution joint bilateral upsampling (MR-JBU), and Wang *et al.* [9]. Test scenes are from the Middlebury datasets. We choose linear form to model the surface similar as that in [9] for fair comparison. The noisy depth map is construct by introducing occlusion according to cross-checking of stereo images, down-sampling($2\times$) and adding Gaussian noise.

Visual comparison is present in Fig.4. Obviously, JBU undergoes texture mapping and blurring artifact, while Wang's greedy patch matching algorithm produces apparent mismatching errors as well since stereo constraint is not applicable. Representative artifacts are shown in red boxes.



Color image Depth map Ground truth MR-JBU [8] Wang's [9] Proposed method

Fig. 4. Visual comparison on the Middlebury datasets. From left to right: input color image, input depth map, ground truth, results by [8], [9] and proposed method. Test scene (from top to bottom) are *Raindeer*, *Midd2* and *Teddy*.

	MR-JBU [8]	Wang's [9]	Ours
Raindeer	8.35	3.65	**3.33**
Midd2	14.10	3.10	**2.51**
Teddy	7.23	4.09	**3.66**

Table 1. Comparison of bad pixel rate (%)

	MR-JBU [8]	Wang's [9]	Ours
Raindeer	1.13	0.98	**0.47**
Midd2	1.67	0.62	**0.31**
Teddy	0.68	0.64	**0.40**

Table 2. Comparison of mean absolute difference

Quantitative comparisons are done via measuring the average percentage of bad pixels (BPR, error $\geq$ 1) and mean absolute difference (MAD), results on three test scenes are listed in table 1 and 2 and our method outperforms the rest algorithms with least BPR rate and MAD score (in bold font). According to quantitative and qualitative comparisons, our proposed method performs satisfactory and better than the other methods.

4. CONCLUSION

In this paper, we have proposed a new depth map enhancement approach based on a hybrid strategy combining filtering and segment-based structure propagation. Specifically, we have presented a new arbitrary-shape patch matching method to robustly extend neighbour patch's structure into query patch. Experiments explicitly show that the proposed method outperforms other methods with respect to depth hole filling problem. In the future, we will pay more attention on improving robustness of the depth inference model so that the filled regions will be seamless and contain fewer mismatching errors.

5. REFERENCES

[1] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," in *NIPS*. 2006, vol. 18, p. 291, MIT.

[2] J. Yang, X. Ye, K. Li, and C. Hou, "Depth recovery using an adaptive color-guided auto-regressive model," in *ECCV*. 2012, pp. 158–171, Springer.

[3] J. Park, H. Kim, Y.W. Tai, M.S. Brown, and I. Kweon, "High quality depth map upsampling for 3d-tof cameras," in *ICCV*. IEEE, 2011, pp. 1623–1630.

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

[5] B. Huhle, T. Schairer, P. Jenke, and W. Straßer, "Fusion of range and color images for denoising and resolution enhancement with a non-local filter," *CVIU*, vol. 114, no. 12, pp. 1336–1345, 2010.

[6] D. Chan, H. Buisman, C. Theobalt, S. Thrun, et al., "A noise-aware filter for real-time depth upsampling," in *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2*, 2008.

[7] F. Garcia, B. Mirbach, B. Ottersten, F. Grandidier, and A. Cuesta, "Pixel weighted average strategy for depth sensor data fusion," in *ICIP*. IEEE, 2010, pp. 2805–2808.

[8] N. A. Dodgson H.-P. Seidel C. Richardt, C. Stoll and C. Theobalt, "Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos," *Computer Graphics Forum (Proceedings of Eurographics)*, vol. 31, no. 2, May 2012.

[9] L. Wang, H. Jin, R. Yang, and M. Gong, "Stereoscopic inpainting: Joint color and depth completion from stereo images," in *CVPR*. IEEE, 2008, pp. 1–8.

[10] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *TPAMI*, vol. 34, no. 11, pp. 2274 –2282, nov. 2012.