Hierarchical Joint Bilateral Filtering for Depth Post-Processing

Qing-Qing Yang∗†, Liang-Hao Wang∗†, Dong-Xiao Li∗†, Ming Zhang∗†
∗Institute of Information and Communication Engineering, Zhejiang University, Hangzhou 310027, China
†Zhejiang Provincial Key Laboratory of Information Network Technology, Hangzhou 310027, China
Email: qqyang@zju.edu.cn

Abstract—Various 3D applications require accurate and smooth depth map, and post-processing is necessary for depth map directly generated by different correspondence algorithms. A hierarchical joint bilateral filtering method is proposed to improve the coarse depth map. By first carrying out depth confidence measuring, pixels are put into different categories according to their matching confidence. Then the initial coarse depth map is down-sampled together with the corresponding confidence map. Depth map is progressively fixed during multi-step upsampling. Different from many filtering approaches, confident matches are propagated to unconfident regions by suppressing outliers in a hierarchical structure. Experiment results present that the proposed method can achieve significant improvement of initial depth map with low computational complexity.

Keywords—joint bilateral filtering; hierarchical; depth post-processing

I. INTRODUCTION

As an important representation of 3D scene, dense depth map bridges the two sides of scene capturing and reconstruction. The problem of depth map acquisition is always converted into stereo correspondence problems under certain geometry constraints, e.g. epipolar geometry constraint. A vast of methods have been proposed during the last decades, [1][2], and [3] provide an excellent review in this research field, and most methods are under certain framework described in [1]. The solution to correspondence problems are always based on some assumptions, like order, uniqueness, consistency and smoothness. To generate accurate and smooth depth map required by high quality scene reconstruction, global methods utilize these assumptions in the form of certain imposed constrains, and solve the problem by optimizing the total energy. These algorithms are always complex and time-consuming. Local methods run fast by aggregating neighboring matching cost in the local window, which are sensitive to noises and mismatches. We propose a new hierarchical post-processing method for the refinement of depth map generated by local methods. Straightforward approaches work poorly in occluded regions near object boundaries and textureless regions. Rather than blurring the wrongly estimated pixels with neighbor estimations, we propagate the confident results to these unconfident regions in a manner similar to the coarse-to-fine approach. By first downsampling the initial raw depth map, hierarchical structure is applied to a set of down-sampled depth images. The confident results are propagated during the upsampling steps with a carefully designed joint bilateral filter. Significant improvement of initial raw depth map can be achieved by applying the proposed hierarchical joint bilateral filtering.

The remainder of the paper is organized as follows. A short review of related works will be presented in Section 2. In Section 3, we will describe the proposed method in detail and some experimental results will be discussed in Section 4. We conclude our work in section 5 and explore possible future work.

II. RELATED WORKS

In applications like view synthesis and scene reconstruction, the depth map must be of good quality. Influenced by the noise and occlusion, mismatches may easily take place in object boundaries and textureless regions. Boundary misalignment is another problem caused by different correspondence estimation algorithms. Window based aggregation will result in known foreground fattening effect, and pixel-based global optimization also performs poor at object boundaries. Depth post-processing is essential to remove mismatches, fix the occlude regions and align object boundaries.

Various methods are proposed to post-process the depth map. In these methods, image segmentation is widely used, aiming at utilizing local patterns to fit initial depth map to objects, which is modeled by piece-wise planes [4][5][6]. In [7], Bayesian image matting [8] is applied to handle visible artifacts caused by mixed pixel colors at edges during image-based rendering. In [6] unstable and occluded pixels are explicitly detected, and [9] introduces outlier confidence to measure the likelihood that one pixel is an outlier, both methods remove these undesired pixels by iterative optimization.

Another category of methods use some kind of neighborhood filters. Median filter is widely used in depth map refinement step to filter out spurious mismatches [1], but foreground fattening effect also occurs when larger filter windows are used. Due to the ability to blur images while preserving object edges, bilateral filter [10] is widely used in many photography applications. Joint bilateral filter [11][12]
extends its usage in a wider field. Yang et al. [6] apply the cross bilateral filter to enhance the depth map to achieve sub-pixel estimation, and later extend the application in depth map super resolution [13]. Rather than directly blur the depth map, bilateral filter is applied to filter cost slices and a better sub-pixel estimation is obtained. Recently, Mueller et al. [14] proposed an adaptive cross-trilateral median filter for depth map post-processing. Additional confidence weight is added to classical bilateral weight kernel and the final result is selected in a weighted median form, instead of weighted mean. Competitive results can be generated from a coarse initial depth map. But the calculation of confidence weight and weighted median selection in [14] are time-consuming and the whole post-processing step need several iterations. In [15], the depth map is first down-sampled to reduce the details within objects and thus some outliers are removed. Then down-sampled depth map is filtered by joint bilateral filter during multistep upsampling, which lower the computation cost.

Recently, low resolution depth maps from time-of-flight(ToF) cameras are used to generate high resolution ones. Many depth super resolution methods [13][16][17][18] are proposed to enhance the initial depth map, and most of which utilize joint bilateral filter following certain upsampling step. Dense high resolution depth map can be achieved with the corresponding high resolution color image.

III. PROPOSED METHOD

Mismatches in depth map derived from local algorithms are mainly distributed throughout occluded and textureless regions, others are separate small noise pixels. Window based filters performs poor at these regions due to the limited window size, while expanding the window size will also increase the processing time tremendously, as well as ruining the object boundaries. Hierarchical structure usually performs in a coarse-to-fine manner. In the small-scale depth map, distant pixels in large-scale image can be related by a small window, and separate mismatches can be removed during the downsampling stage. Thus, depth post-processing can be carried out by starting from a relative accurate low-resolution depth map. Confident Ground Control Points(GCPs) are propagated during multi-step upsampling to fix mismatches. Though our scheme has similar skeleton with [15], our method solves the post-processing problem in a quite different way. Confidence map is introduced and structured depth refinement steps are utilized in our hierarchical joint bilateral filtering scheme. Figure 1 presents the flow chart of the whole post-processing algorithm.

After receiving a stereo depth pair, a confidence map is generated by carefully designed confidence measure. Then the initial coarse depth map is down-sampled with a box filter to form depth map pyramid, together with the registered color image. Hierarchical joint bilateral filtering is performed during multi-step upsampling stages. In each upsampled level, confident pixels are preserved and unconfident pixels are re-estimated by structured filtering, which is performed in a propagation manner. The improved depth map is derived when being restored to the original scale. In the remainder of this section, we explain key steps of our algorithm.

A. Depth Confidence Measuring

Depth map generated by local stereo correspondence method will have many mismatches, but the majority of pixels are correctly matched pixels with high confidence. The percentage of confident matches depends on utilized local algorithms. In the proposed post-processing method, these confident matches are treated as Ground Control Points, which are propagated to re-estimate unconfident pixels or outliers. A robust and efficient confidence measuring method is essential. A summary of confidence metrics for stereo matching can refer [19].

We measure the matching confidence by combining Left Right Consistency(LRC) check and matching cost Peak Ratio(PKR) calculation. And each initial depth estimation is labeled with one of following tags: confident, unconfident and occluded. For a selected pixel \((x_l, y_l)\) in the left image with depth value \(d_l\), the corresponding pixel in right image \((x_r, y_r)\) has depth value \(d_r\), LRC is measured by:

\[
C_{LRC}(x_l, y_l) = |d_l - d_r| \tag{1}
\]
Pixels with LRC value above threshold $\eta_{LRC}$ are considered as occluded pixels, otherwise, the matching cost $PKR$ is calculated. The cost of the candidate depth value is denoted by $c$, and by defining $c_1$ the minimum cost for a pixel, $c_2$ the second minimum cost, the PKR is defined by:

$$C_{PKR}(x, y) = \frac{|c_1 - c_2|}{c_2}$$

Pixels with $C_{PKR}$ under threshold $\eta_{PKR}$ are labeled unconfident, and the remaining pixels are labeled confident.

The generated confidence map is used to guide the following post-processing steps. And the GCPs, which are composed of confident pixels, are used in hierarchical bilateral filtering to re-estimate unconfident and occluded pixels.

B. Multi-step Down-sampling and Up-Sampling

The initial depth map generated by local algorithm is down-sampled in a multi-step manner. In each step, the depth map is downsampled by a factor of $2 \times 2$ with a box filter. To preserve the correct depth value, pixel in the top left corner of the $2 \times 2$ sampling window is selected. Different input stereo pairs may have different resolutions, and appropriate setting of pyramid level is needed. We calculate the pyramid level adaptively by the following formula:

$$N_{PL} = \lfloor \log_2\left(\frac{D_{\max} \cdot \alpha}{A_f} \right) \rfloor,$$

where $D_{\max}$ is the maximum depth value, $\alpha$ is a preset constant factor, and $A_f$ is the filter aperture. This adaptive value ensures that the downsampled occluded regions can be fully covered by the filter window in the bottom pyramid level. Both the confidence map and reference color image are down-sampled as well.

In each upsampling step, the sampling points in high-resolution unconfident depth map are replaced by the corresponding pixels from low-resolution confident depth map, which is considered fixed. The upsampling factor in each step is also set to 2. Different from many depth super resolution methods [17][18], in which the remaining pixels in each $2 \times 2$ window are all unconfident except the derived one, in the proposed method, confident matches in the initial depth map are preserved and can contribute to the structured filtering. The details will be described in the following subsection.

C. Hierarchical Joint Bilateral Filtering

Joint bilateral filter is widely used in the depth map refinement stage. For a point $p$ in the original depth map with depth value $D(p)$, which is to be filtered, and $q$ is another point in the neighborhood $N(p)$ of $p$. Let $I_{ref}$ denotes the reference color image. Then the joint bilateral filter can be expressed as:

$$D_{\text{new}}(p) = \frac{1}{\sum_{q \in N(p)} W_q} \cdot \sum_{q \in N(p)} W_q \cdot D(q),$$

$W_q$ weights the contribution of the depth value $D_q$ of neighborhood pixel $q$, which is defined as:

$$W_q = F_{\sigma_s}((|p - q|) \cdot G_{\sigma_r}(|I_{ref}(p) - I_{ref}(q)|),$$

where $F(\cdot)$ and $G(\cdot)$ are spatial and range weighting function of the joint bilateral filter, and $\sigma_s$ and $\sigma_r$ are corresponding parameters.

Kopf et al. [16] utilize joint bilateral filter in depth upsampling by applying a spatial filter to the low resolution, while applying a similar range filter to the full resolution image. And in [13], joint bilateral filter is applied to cost volume for predicting more accurate depth values in upsampled depth map. And Riemens et al. [17] point out that a multi-step approach can reduce the computation cost, and Yang et al. [18] also improve their original approach by utilizing a hierarchical structure.

A hierarchical joint bilateral filter is designed to re-estimate the unconfident and occluded pixels. Followed by the latest upsampling step, fresh and more confident depth estimations are derived from the fixed low-resolution depth map. Rather than applying the filter to each pixel, we perform filtering in a structure similar to the one proposed in [18]. Figure 2 illustrates our hierarchical bilateral filtering algorithm on a $3 \times 3$ fixed depth map. Subject to the figure size, we demonstrate our approach with a $3 \times 3$ filter window, while in real application, a larger window of $5 \times 5$ is used. The unconfident pixels surrounded by 4 confident neighbors are firstly fixed as shown in Figure 2(b), then the remaining unconfident pixels are re-estimated. This two-step approach ensures that in each step, the unconfident pixel to be re-
estimated is surrounded by at least 4 confident neighbors.

To make the filter be more content adaptive, a simplified variant of original joint bilateral filter is utilized. For confident pixels, the spatial function $F()$ represents a box filter, which returns weight 1 if the pixel is within the filter aperture and weight 0 outside. Since the weighting pixels in the filter aperture can be derived from many ways, e.g. some are preserved confident depth, some are fixed pixels from the previous scale, a confidence function $C()$ defined by equation (6) is added to adjust the range weight:

$$C(x, y) = \begin{cases} 
1.0, & \text{confident pixel}, \\
\exp\left(-\frac{|x|}{\sigma_x}\right), & \text{fixed pixel}, \\
0, & \text{unconfident pixel}, 
\end{cases}$$

(6)

where $x$ is the distance to the pixel in filter center. Let $D^i$ denotes the depth map in the $i$th scale, and $I^i$ the registered color image, the designed bilateral filter can be expressed as:

$$D^{i,\text{new}}(p) = \frac{1}{\sum_{q \in N(p)} W^i(q)} \cdot \sum_{q \in N(p)} W^i(q) \cdot D^i(q),$$

(7)

where the $W^i$ is the weighting function in the $i$th scale:

$$W^i(q) = C^i(q) \cdot G_{\sigma_r}(|I^i(p) - I^i(q)|).$$

(8)

The confident weight plays a significant role during the hierarchical filtering approach. By assigning high weight to confident pixels in the original depth map, reliable pixels contribute mostly in depth re-estimation. The values of fixed pixels from the previous level are less confident, thus low weight is assigned. And outliers in filter window are suppressed by assigning zero weight to these pixels.

### IV. Experimental Results

The quality of initial depth map seriously depends on the utilized stereo correspondence algorithm. Pixel based local matching algorithm results in noisy depth map, while smoother result can be generated by aggregating costs with fixed box window, but known foreground fattening effect occurs, which is hard to fix. And variable window approaches are considered the best in local algorithms so far, but mismatches are inevitable in object boundaries and textureless regions. In this section, the proposed method is applied to initial depth map generated by two different local methods to test the robustness and generality. Both the experiments are taken under the same parameters shown in table I.

<table>
<thead>
<tr>
<th>Table I</th>
<th>Experimental Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\eta_{LR}$</td>
<td>$\eta_{KR}$</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Supplied with high quality initial depth map, better post-processing result can be achieved. The proposed method is first applied to depth map(Fig. 3(b)) generated by cross-based adaptive window approach [20]. The filtered result is presented in Figure 3(d). Compared with results by directly joint bilateral filtering(Fig. 3(c)), our scheme achieve clearer edges as well as fixing mismatch regions. A general fix window approach is employed to test the robustness of the proposed method, raw costs are aggregated by a $9 \times 9$ window, which fattens the foreground greatly. Figure 4 illustrates the initial depth map and the result after post-processing. Most mismatches are fixed in filtered image, but due to the fattened foreground, edges are blurred to a certain extent. Since the initial depth map is rigorous, our method achieves comparable results with low calculation amount.

A further comparison is carried out by comparing the result of the proposed method with the ones generated by other known filters. Figure 5 illustrates the comparative result. To better demonstrate the differences, details of se-
lected region from depth maps are presented. Corresponding regions from synthesized views are also compared to show the benefit of the proposed method in virtual view synthesis. We can find that by direct joint bilateral filtering, though the object boundaries are well preserved, mismatches are spread; median filter fattens the depth edges and Gaussian filter smooth the whole depth map as well as blurring the edges, which result in distortion in synthesized view; the result of our method present smooth depth map, and the occluded regions are well predicted.

A. Analysis of Algorithm Complexity

The brute force implementation of bilateral filter has complexity of $O(|S|^2)$, where $|S|$ is the resolution of input image. Typically, only neighbor pixels in filter aperture $A_f$ are considered, which lead to a reduced complexity of $O(A_f^2|S|)$. Small aperture will make object edges not properly aligned, while undesirable details appears if large aperture is utilized. A empirical value of filter aperture is the depth level ($N_d$) of the depth map.

The proposed method takes three main steps, depth confidence measuring, multi-step downsampling and upsampling, and structured joint bilateral filtering. Since the depth confidence measuring operated in pixel-wise manner, so the total operation needed is $|S|$. Both downsampling and upsampling require operation $\sum_{i=1}^{n} \frac{1}{2^i} \cdot |S| < \frac{1}{4}|S|$, where $n$ is the number of pyramid levels, then the complexity of sampling step is on the order of $O(|S|)$. In each scale, a fixed filter window of $5 \times 5$ is used. If brute force implementation is applied to each level, as carried out in [15], and take the weight calculation of each neighbor pixel as a meta operation, the total operations required is

$$5^2 \cdot \sum_{i=0}^{n} \frac{1}{2^i} \cdot |S| < \frac{100}{3}|S|.$$

But in hierarchical joint bilateral filtering, confident pixels are preserved, and only occluded and unconfident pixels are re-estimated in each scale. Fixed pixels derived from low scale are also preserved, as illustrated in figure 2. The proposed method actually requires total $25\beta|S|$ defined meta operations, where $\beta$ is the unconfident ratio of initial depth map. Our scheme greatly reduces the calculation amount by preserving the complexity on the order of $O(|S|)$, and generates better depth map.

V. Conclusion and Future Work

A hierarchical joint bilateral filter is presented for depth map post-processing. Smooth results can be achieved from coarse initial depth map generated by different fast local correspondence algorithms. A confidence map is suggested to assist the depth propagation. The hierarchical structure ensures that unconfident pixels are re-estimated based on more believable pixels. Compared with straightforward joint bilateral filtering, the proposed method generates better filtered
depth map, and the amount of calculation is greatly reduced by applying small aperture as well as preserving confident matches. Despite of these improvements, our method heavily depends on the confidence map, more robust method for depth confidence measuring should be further studied. And the regular structure of the proposed method made it suitable for parallel implementation, GPU implementation will be carried out in future work.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 60802013, 60971061, 61072081), and the National Science and Technology Major Project of the Ministry of Science and Technology of China (Grant No. 2009ZX01033-001-007), Key Science and Technology Innovation Team of Zhejiang Province, China (Grant No. 2009R50003) and China Postdoctoral Science Foundation (Grant No. 20110491804).

REFERENCES


