Stereo Refinement for Photo Editing

Dongwei Liu and Reinhard Klette

The *.enpeda..* Project, Department of Computer Science
The University of Auckland, New Zealand

Abstract. We present a method for refining depth information generated by a stereo-matching algorithm with the goal to provide depth-aware photo-effect applications. Our key idea is to use structural features of the base image to enhance the depth information. Our method pre-processes the original disparity map by revising the sky region and removing incorrect data (on the left-side of the disparity map) caused by occlusion. The base image is mean-shift segmented. A median filter is applied on the disparity map within each segment. Invalid step-edges in the disparity map are removed by a joint bilateral filter. Experiments show that our method can revise holes, inaccurate object edges, speckle noises, and invalid step-edges from the given depth information. Results illustrate the applicability for photo editing.

1 Introduction

Photography is an art that maps a real-world 3D scene into a planar photo. During this process, the depth information of the scene is lost. Various impressive 3D rendering techniques can be applied to photos if depth information is available, for example out-of-focus blur [19], addition of smoke or haze [20, 6], or relighting.

Depth from a single 2D picture would be convenient but is an unsolvable problem in general. Studies in this area either apply constraints on the scene or rely on massive human interaction [3, 15].

Depth sensors are more reliable, such as *binocular stereo* [5, 7], structured lighting as used in the *Kinect* [11], *depth from defocus (DFD)* [16, 17], or a *light field camera* [14]. Binocular stereo is the common sensor used for outdoor scenes, and the underlying methodology is close to human visual cognition. Generated depth maps when using stereo vision are still imperfect for photo editing, even when using a top-performing stereo matcher such as iSGM [8], for the following three reasons:

First, depth values for some pixels are unavailable in a photo, represented by *holes* in the depth map (see Fig. 1, green frame). Those holes can be due to occlusion, which is inherent for stereo vision, or to low confidence in stereo matching. For object detection or distance measurement tasks, such holes can be tolerated to some extent, but for photo editing a dense depth map is needed.

Second, the depth map may include speckle noise (see Fig. 1, red frame), and inaccurate object edges. The issue is especially noticeable for objects with

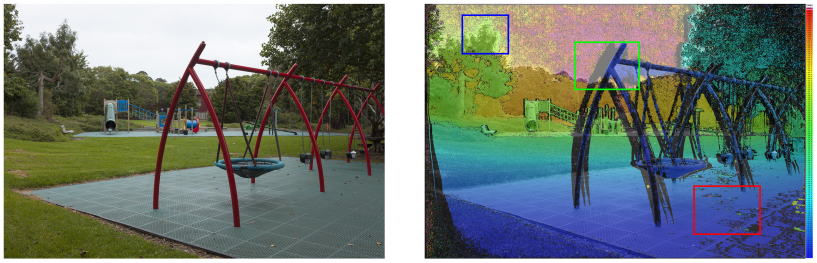

Fig. 1. The result of stereo matching is imperfect for photo editing. *Left:* Base image. *Right:* Depth map generated by iSGM [8] overlayed with the base image. See *blue frame* for inaccurate edges, *green frame* for holes, and *red frame* for speckle noise.

a detailed geometry, for example for tree crowns (see Fig. 1, blue frame). For photo editing, even slight mismatches on some clearly visible edges introduce visual discomfort.

Third, the depth map may involve invalid step-edges. Stereo matching methods detect *disparity* between base and match images. The disparity is measured in integers. The depth of a point is inversely proportional to its disparity value. Thus, depth does not change smoothly when only using integer disparities. For example, if the disparity values of two adjacent pixels are 2 and 1, their depth varies by factor 2. This issue also causes visual discomfort during photo editing.

We observe that for photo editing, the depth map should have little "collision" with human visual cognition of the base image. On the other hand, 100% accurate depth information is also not necessary. Informally speaking, a depth map needs to look "plausible". Based on this observation, we present a novel depth-refinement method to solve the three problems mentioned above. The key idea is to use structural features of the base image to enhance the depth map.

Our process takes as input a stereo pair. We generate an enhanced disparity map, fit for photo editing. The process includes four steps. First we pre-process a given disparity map in which the disparity in the sky region gets revised, and incorrect information on the left side of the base image (caused by occlusion) is removed. Then, we mean-shift segment the base image. Notable object edges are detected during segmentation. Later, we run a median filter on the disparity map within each segment. In this step, holes on the disparity map are "fixed", disparities around object edges are revised, and minor speckle noise is also removed. At last, before converting the disparity map into a depth map, a joint bilateral filter is employed to remove invalid step-edges.

The rest of the paper is structured as follows. In Section 2 we recall algorithms as used in our method. Section 3 provides details for our depth-refinement method. Experimental results are shown and discussed in Section 4. Section 5 concludes.

2 Basic Concepts and Notations

This section briefly recalls algorithms as used in our approach.

Stereo Matching. Research on stereo matching has a long history already. Good performing algorithms are, for example, based on belief propagation [5] or on *semi-global matching* (SGM) [7]. A variant of SGM, *iterative SGM* (iSGM) [8], was winning the Robust Vision Challenge at ECCV 2012, and we use iSGM in this paper. Stereo matching aims at solving an ill-posed problem (to identify exactly one matching pixel in a match image starting with one pixel in a base image [9]. Difficulties for solving this problem arise for many reasons, and one is occlusion. Algorithms (including iSGM) use a smoothness constraint which also causes "blurred" disparities at occlusion edges. For our purpose we like to "sharpen" at those edges; for this reason we "merge" results of iSGM with segmentation results in the base image.

Mean-Shift Image Segmentation. *Mean-shift segmentation* [4] is an iterative steepest-ascent method that detects peaks in the density function defined in a feature space; e.g., see [9]. Mean-shift requires to specify a window of defined size, and weights on this window (the kernel) in feature space.

Median Filter and Bilateral Filter. We also apply the median and bilateral filter, which are both known as edge-protecting image smoothing methods. The *median filter* is a nonlinear operation which runs through an image I and replaces each pixel value $I(p)$ by the median value of neighboring pixels within a $(2k+1) \times (2k+1)$ window W_p:

$$I_{median}(p) = \text{median}\{I(p_i) : \ p_i \in W_p\} \tag{1}$$

The *bilateral filter* [18], also known as "surface blur", is a selective mean filter for image smoothing or noise reduction. The filter does a weighted average for each pixel p in image I in a window W_p considering both spatial distance and color intensity distance of pixels:

$$I_{bilateral}(p) = \frac{1}{\omega_p} \sum_{p_i \in W_p} I(p_i) \cdot f_c(\|I(p_i) - I(p)\|) \cdot f_s(\|p_i - p\|) \tag{2}$$

where f_c is the kernel for color-intensity distances of pixels, f_s is the kernel for the spatial distance of pixels, ω_p is a normalization parameter defined by

$$\omega_p = \sum_{p_i \in W_p} f_c(\|I(p_i) - I(p)\|) \cdot f_s(\|p_i - p\|) \tag{3}$$

Here, f_c and f_s can be Gaussian functions. A variation of a bilateral filter has been developed for depth refinement, called *joint bilateral filter* [12, 10]. This filter uses the original color image I to specify the kernel, and then refines the corresponding depth map D:

$$D_{J\_bilateral}(p) = \frac{1}{\omega_p} \sum_{p_i \in W_p} D(p_i) \cdot f_c(\|I(p_i) - I(p)\|) \cdot f_s(\|p_i - p\|) \tag{4}$$

3 Depth Refinement

Given a pair of rectified stereo images, base image I_L and match image I_R, we calculate a disparity map D_0 using a stereo matching method, where D_0 is given in coordinates of I_L defined on Ω.

3.1 Sky Region and Left-side Occlusions

Sky is typically shown in landscape photographs. Since a sky region is often large and has little texture, the stereo matching results in such a region are often not accurate. We detect the sky region in I_L, and set the corresponding values in D_0 uniformly to 0, which marks the sky region as being at infinity.

Blue sky and cloud regions have both high values in the blue channel of I_L. Thus, we define a pixel p is a *sky pixel* if p is in the upper half of I_L, and its blue channel value $B(p)$ is larger than a threshold T_{sky}:

$$D(p) = 0 \quad \text{if} \quad p \in \Omega_{upper} \wedge B(p) > T_{sky} \tag{5}$$

We used $T_{sky} = 0.8 \cdot G_{max}$ in our experiments, where G_{max} is the maximum level in each color channel of I_L.

Due to the nature of stereo vision, a part of the scene on the left-side (in Ω_{left}) of image I_L is not included in I_R.[1] Thus, accurate depth information of this region cannot be generated by stereo matching. In general, a usual practice is to discard I_L on Ω_{left}. For photography this operation may destroy the composition of an artwork. Thus, we aim at repairing data in Ω_{left} by given information.

We remove erroneous information in Ω_{left}; resulting gaps are repaired by the process discussed in Section 3.3. If the disparity value at a pixel $p \in \Omega_{left}$ deviates significantly from the median value of the row, we identify it as being an outlier, and invalidate the value of $D(p)$:

$$D(p) = \text{NA} \quad \text{if} \quad p \in \Omega_{left} \wedge |D(p) - \text{median}(\text{Row}_y)| > T_{outlier} \tag{6}$$

where $p = (x, y)$ and Row_y denotes all the disparity values in row y.[2] We denote the pre-processed disparity map as D_1.

3.2 Image Segmentation

For analyzing the structural features of I_L, we employ mean-shift image segmentation which clusters the pixels of I_L into a family $\mathbf{S}$ of segments. Figure 2, top-right, shows a segmentation result. Each segment is labelled constantly with its mean color.

In the segmentation procedure, we convert I_L from *RGB* into the *Lab* color space which is closer to human visual cognition principles. Then we run mean-shift on the 3D feature space formed by the *Lab* color values using a uniform kernel function. The window size $2k + 1$ is selected to be $0.125 \cdot G_{max} - 1$ which is a reasonable window size due to our experiments (a smaller value clusters the image into more segments, and clearly visible edges should also not be ignored). Generally we prefer over-segmentation to under-segmentation.

[1] We select Ω_{left} to be the left 10% of Ω.

[2] We decided for threshold $T_{outlier} = 0.3 \cdot d_{max}$, where d_{max} is the maximum disparity.

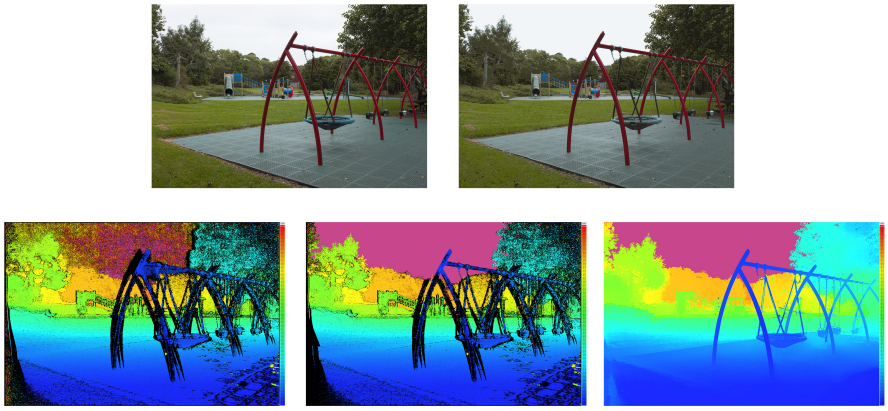

Fig. 2. Disparity refinement. *Upper row:* Base image and mean-shift segmentation result. *Bottom row:* Original disparity map, after processing the sky region and occlusions on the left-hand side, and final result.

Compared with region-based segmentation methods [2] or a superpixel technique [1], mean-shift appears to be better at detecting complex edges, and even discontinuous objects.

3.3 Holes, Speckle Noises, and Inaccurate Edges

Similar to the joint bilateral filter, we define a *joint median filter* to be a median filter applied on a disparity map D with a family $\mathbf{S}$ of segments as masks:

$$D_{J\_median}(p) = \text{median}\{D(p_i) : \ p_i \in W_p \cap S_p\} \tag{7}$$

W_p is a window around p, and $S_p \in \mathbf{S}$ is the segment which contains pixel p.

We use the joint median filter as the main step of our refinement. The size of window W_p remains a user-defined parameter which should be decided based on the scale of inaccurate edges and speckle noise.

Due to larger gaps, the joint median filter may not cover all the unavailable pixels unless when using a very large window W_p, which may lead to an over-smoothed result and low efficiency. Thus, we first run the joint median filter on every pixel in D_1, and then we run the filter iteratively on those pixels still unavailable until all the gaps are closed, resulting in a processed depth map D_2.

Taking advantage of image segments in $\mathbf{S}$ which represent the edge information of base image I_L, inaccurate depth information around object edges and speckle noise are revised by the voting mechanism of the median filter. See Fig. 2.

Note that the joint bilateral filter does not discard invalid information (e.g. inaccurate edges or speckle noise). Instead, it spreads inaccurate values to adjacent regions. In other words, the joint bilateral filter is affected by outliers. This effect is shown in Section 4. In contrast, our joint median filter is robuster on outliers if window W_p is of sufficient size.

3.4 Invalid Step-Edges

A step-edge in the depth map is *invalid* iff it does not correspond to an actual depth discontinuity in the scene. The disparity map D is converted into a depth

map d as follows:

$$d(p) = \frac{f \cdot b}{D(p) + \Delta}, \text{ for } D(p) \in \mathbb{N} \tag{8}$$

where Δ is a small value to avoid division by zero, f is the focal length, and b is the base distance of the stereo cameras.

Since $D(p) \in \mathbb{N}$, the depth value $d(p)$ does not transit smoothly. A median filter does not create a new value, and it does not remove those invalid step-edges in D_2. Thus, we employ the joint bilateral filter defined in Equ. (4), which removes such edges while keeping valid edge sharp. We use a Gaussian kernel and the Euclidean metric for both spatial distance and color intensity distance. The filtered result is denoted as D_3.

Though the joint bilateral filter may spread inaccurate information, it works well as an edge-protecting smooth filter on D_2, where inaccurate information has been revised.

4 Experiments

We test our refinement method on disparity maps generated by iSGM [8]. The original resolution of the images in Figs. 2, 3, and 4 is 2400×1600. The refinement window W_p is set to be 80×80.

Figure 3 compares the original disparity maps and the results of the joint bilateral filter with our final results. Our method generates dense disparity maps with clear object edges. Compared with the joint bilateral filter, our method is robuster on outliers (for example, see the speckle noise on ground regions and inaccurate edges of tree crowns). Our method also achieves correct disparity in sky regions, and avoids the interference of left-hand side inaccurate information caused by occlusion.

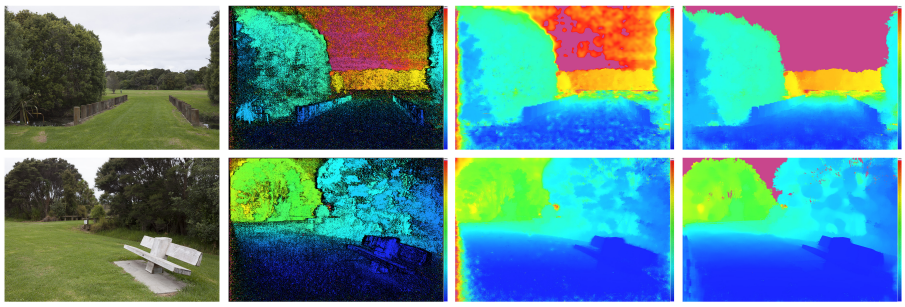

Fig. 3. Comparisons between joint bilateral refinement and our method. From left to right: base image, original disparity map, joint bilateral filter result, our final result.

For illustrating the usability of our method for photo editing, we implemented two simple depth-aware photo effects: darken either the foreground or the background of a photo. See Fig. 4. The darkening effect changes smoothly and naturally with depth. Edges of objects (e.g. tree crowns) also feel natural and comfortable under the darkening effect. Our results meet the demands of depth-aware photo effects.

Fig. 4. Application examples. Darken the foreground or background of the photos shown in Fig. 3 using our refined disparity map.

We also tested our method on standard examples provided by the *Middlebury Stereo Vision Page* [13], see Fig. 5. The resolution of the upper and lower examples are 434×383 and 384×288, respectively. We set the refinement window W_p to be 15×15. The refined disparity maps are better than the original disparity maps, but not perfect compared to the ground truth, possibly induced by the low resolution of the input data. The core mechanism of our method is a selective median filter, which is reliable only on a large refinement window. Low-resolution input data limit the size of the window, thus constraining positive effects of our method.

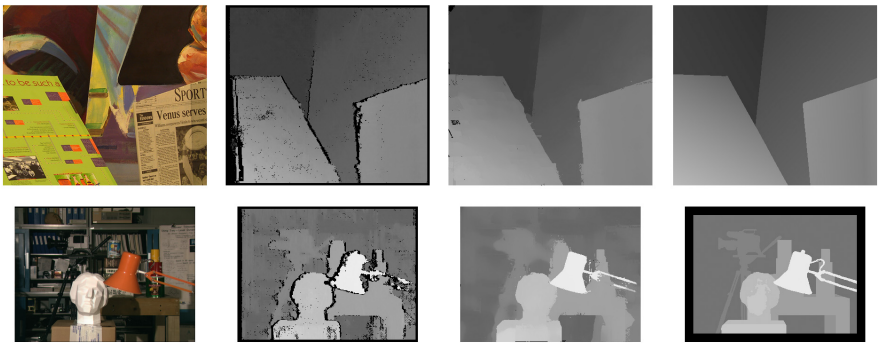

Fig. 5. Results on benchmark data of the *Middlebury Stereo Vision Page* [13]. From left to right: Base image, original disparity map, our refined result, ground truth.

5 Conclusions

In this paper we present a stereo refinement method for photo editing. The input is a stereo pair. We calculate the disparity map and pre-process it by revising the sky region and by removing incorrect data on the left of the disparity map. Then, the base image is mean-shift segmented in order to detect edge information. Later, a median filter is applied on the disparity map within each segment. Finally, invalid step-edges in the disparity map are removed by a joint bilateral filter.

Experiments show that our method can revise holes, inaccurate object edges, speckle noise, and invalid step-edges from disparity maps generated by a stereo matcher. The results are suitable for photo editing.

References

1. Achanta, R., Shaji, A., Smith, K,. Lucchi, A., Fua, P., and Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods, *IEEE Trans. Pattern Analysis Machine Intelligence*, 34, 2274–2282 (2012)
2. Deng, Y., Manjunath, B.S.: Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans. Pattern Analysis Machine Intelligence*, 23, 800–810 (2001)
3. Fattal, R.: Single image dehazing. *ACM Trans. Graphics*, 27, No. 3 (2008)
4. Fukunaga, K. and Hostetler L. D.: The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Information Theory*, 21, 32–40 (1975)
5. Felzenszwalb, P.F. and Huttenlocher, D.P.: Efficient belief propagation for early vision. *Int. J. Computer Vision*, 70, 41–54 (2006)
6. Fedkiw, R., Stam J. and Jensen, H.W.: Visual simulation of smoke. In Proc. *ACM SIGGRAPH*, pp 15–22 (2001)
7. Hirschmüller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In Proc. *Computer Vision and Pattern Recognition*, Vol.2, pp 807–814 (2005)
8. Hermann, S. and Klette, R.: Iterative semi-global matching for robust driver assistance systems. In Proc. *Asian Conf. Computer Vision*, 465–478 (2013)
9. Klette, R.: *Concise Computer Vision - An Introduction into Theory and Algorithms.* Springer, London (2014)
10. Kopf, J., Cohen, M. F., Lischinski, D. and Uyttendaele, M.: Joint bilateral upsampling. *ACM Trans. Graphics*, 26, no. 96 (2007)
11. Khoshelham, K. and Elberink, S.-O.: Accuracy and resolution of Kinect depth data for indoor mapping applications. *Sensors*, 12, 1437–1454 (2012)
12. Matsuo, T., Fukushima, N. and Ishibashi, Y.: Weighted joint bilateral filter with slope depth compensation filter for depth map refinement. In Proc. *Int. Conf. Computer Vision Theory Applications*, (2013)
13. Middlebury Stereo Vision Page. vision.middlebury.edu/stereo/data/
14. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M. and Hanrahan, P.: Light field photography with a hand–held plenoptic camera. *Computer Science Technical Report*, 2, No. 11 (2005)
15. Oh, B. M., Chen, M., Dorsey, J. and Durand, F.: Image-based modeling and photo editing In Proc. *ACM SIGGRAPH*, pp 433–442 (2001)
16. Schechner, Y. Y., and Kiryati N.: Depth from defocus vs. stereo: How different really are they?. *Int. J. Computer Vision*, 39, 141–162 (2000)
17. Subbarao, M. and Surya, G.: Depth from defocus: a spatial domain approach. *Int. J. Computer Vision*, 13, 271–294 (1994):
18. Tomasi, C. and Manduchi, R.: Bilateral filtering for gray and color image. In Proc. *Int. Conf. Computer Vision*, pp 839–846 (1998)
19. Wu, J., Zheng, C., Hu, X., Wang, Y. and Zhang, L.: Realistic rendering of bokeh effect based on optical aberrations. *The Visual Computer*, 26, 555–563 (2010)
20. Zhou, K., Hou, Q., Gong, M., Snyder, J., Guo, B. and Shum, H.-Y.: Fogshop: Real–time design and rendering of inhomogeneous, single-scattering media In Proc. *Pacific Conf. Computer Graphics Applications*, pp 116–125 (2007)