

Disparity Refinement Using Merged Super-Pixels for Stereo Matching

Jianyu Heng, Zhenyu Xu, Yunan Zheng, and Yiguang Liu^(✉)

Vision and Image Processing Lab (VIPL), College of Computer Science,
SiChuan University, Chengdu 610065, People's Republic of China
liyug@scu.edu.cn, lygpapers@aliyun.com

Abstract. The traditional disparity refinement methods cannot get highly accurate disparity estimations, especially pixels around depth boundaries and within low textured regions. To tackle this problem, two novel stereo refinement strategies are proposed: (1) merging super-pixels into stable region to maintain continuity and accuracy of the same disparity; (2) optimizing the co-operative relations between adjacent regions. Then we can obtain high-quality and high-density disparity maps. The quantitative evaluation on Middlebury benchmark shows that our algorithm can significantly refine the results obtained by local and non-local methods.

Keywords: Stereo matching · Disparity refinement · Super-pixels
Region optimization

1 Introduction

Stereo matching has been one of the key problems in computer vision for years. Recently, most of publications [1–4] have been focused on solving this problem. And the segment-based methods [7–9] have attracted more and more attention due to their good performances for years.

Most segment-based stereo matching algorithms follow the four-step pipeline [5]: First, matching cost computation; Second, cost aggregation; Third, disparity computation/optimization; Fourth, disparity refinement. Traditional disparity refinement methods, involving left-right consistency checking [10], hole filling [11], and median filtering [12, 13], could not provide highly accurate disparity estimation. Yoon et al. [14] adopted adaptive supporting-weight approach for correspondence search to refine the local aggregation results. Yang [15] firstly proposed the non-local aggregation method and refined the non-local results with minimum spanning tree (MST). Based on Yang's method, Mei et al. [16] proposed a segment-tree (ST) structure for non-local cost aggregation, they enhanced the disparity values, with a depth-color segmentation method extended from a classic graph-based segmentation method [17]. The region-based methods [18, 19], presented to further improve the disparity estimation, can get better results especially in low textured regions.

In this paper, we propose a stereo refinement algorithm based on merging super-pixels (MSP). Our algorithm includes the following seven steps: First, estimating the initial disparity values with a local or non-local method and locating the

super-pixels with a depth-color segmentation method from stereo images; Second, estimating the robust information of each super-pixel by voting; Third, searching for the supporting neighbors of each super-pixel; Fourth, merging super-pixels into region based on the correlation of adjacent super-pixels; Fifth, updating the information of each region and finding out unreliable regions; Sixth, correcting unreliable region with its supporting region; Seventh, assigning disparity value for each pixel with considering the disparity of the correlative region.

In general, our paper makes these main contributions: (1) we merge super-pixels into stable region, then the disparity of each pixel can be estimated by considering the constraint on smoothness of the correlative region to maintain the continuity of the same disparity. (2) we apply the optimization of the cooperative relations between adjacent regions to reduce the unreliable disparity values and obtain the high-quality depth boundaries.

2 Obtaining Raw Cost Aggregation and Initial Disparities

2.1 Obtaining Cost and Disparity in Pixel Domain

First of all, we employ some local or non-local algorithms to obtain the raw cost aggregation and initial disparity values. These algorithms always poorly use WTA strategy to select disparities from multiple candidates and the disparity estimation obtained by these algorithms is not accurate enough. Later, the accuracy will be improved by our algorithm.

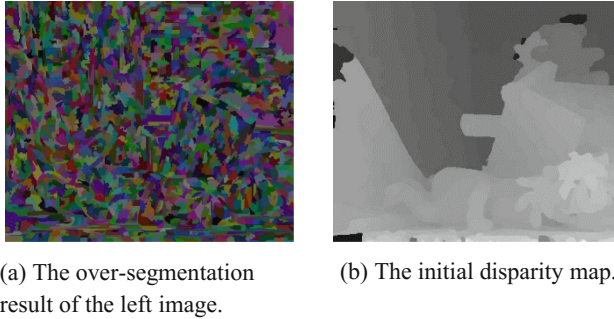
2.2 Over-Segment Based on Color-Depth

Segment-based algorithms usually assume that disparity values vary smoothly in each segment and the depth discontinuities only occur on segment boundaries. But in practice, over-segment based on color-depth is preferred and the assumption is not always met. In this paper, we use efficient graph-based image segmentation [16, 17]. Figure 1 shows the disparity map of the Teddy stereo pair and the segmentation result of the left image produced by the method in [16]. In this paper, we call the over-segmentation super-pixel.

2.3 Cross-Checking Test

At first, a local or non-local cost aggregation method runs the left and the right image as reference images in turn to obtain two corresponding disparity maps. In order to eliminate the outlier in disparity map and obtain robust disparity estimation of each segmentation, the cross-checking test is applied. Then the occlusions and matching errors in the disparity map can be obtained, they are all called unreliable pixels in this paper. After cross-checking, the cost volume is refined according to [15]. Let D denotes the disparity map, a new cost value is computed for each pixel p at each disparity level d as:

$$C_1^n(p) = \begin{cases} |d - D(p)|, & p \text{ is stable and } D(p) > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$



(a) The over-segmentation result of the left image.

(b) The initial disparity map.

Fig. 1. The segmentation result of the left image by using color-depth based over-segmentation method and the disparity map of the Teddy stereo pair by using segment-tree stereo matching algorithm [16]. (Color figure online)

3 Robust Super-Pixels Merging

The super-pixels are sensitive to unreliable pixels and they are correlative rather than individual. If the super-pixel is handled solely, the disparity values around the boundary between adjacent regions, which have the same disparity may be discontinuous. In this paper, an effective approach of merging super-pixels to stable region is proposed to resolve this problem.

3.1 Voting the Information of Super-Pixel

Before merging, the information of super-pixels should be obtained by voting. The information contains RGB values, disparity and the message whether the super-pixel is unreliable or not. The process of voting robust information can be expressed as:

First, the RGB values of super-pixel are estimated by using RGB values of all pixels within the region. And the RGB values of each super-pixel are respectively determined by voting a one-dimensional histogram, where the x-coordinate is the value of one of the three channels, and the y-coordinate is the count number of values. After sorting the histogram and smoothing operation by a Gaussian filter, the value of each individual channel is finally estimated by the maximum of the corresponding histogram;

Second, the disparity of each super-pixel is estimated in a similar way by getting rid of unreliable pixels.

Third, if the number of unreliable pixels in a super-pixel is more than a given per-cent of the number of all pixels within the super-pixel, we regard this super-pixel as an unreliable super-pixel and assign true (denotes the super-pixel is unreliable) for the message of this super-pixel. Let W_{occ} denotes the percent.

3.2 Supporting Neighbors Selection

In order to get rid of piecewise smooth, the super-pixels should be merged to stable region by considering the supporting neighbors of each super-pixel. Let $W_i(S_p)$ denotes

the weight of the correlation between the super-pixel S_p and its neighboring super-pixel S_i . Considering the difference of disparity and color between super-pixels S_p and S_i . The ratio α , which denotes the ratio of common border lengths to perimeter, can be written as:

$$\alpha = \frac{N_i}{N_{all}} \quad (2)$$

where N_i denotes the length of the boundary between super-pixel S_p and S_i . And N_{all} denotes the perimeter of super-pixel N_i . Thus, $W_i(S_p)$ can be written as:

$$W_i(S_p) = \alpha \frac{|D_r(S_p) - D_r(S_i)| \cdot \sigma_s}{\sigma_c} = (1 - \alpha) \frac{I_r(S_p) - I_r(S_i)}{\sigma_c} \quad (3)$$

where S_i covers all neighbors of super-pixel S_p . σ_s and σ_c are two variables, which can self-adapt in terms of the disparity range and color range, to normalize I_r and D_r to the range $[0, 1]$. D_r denotes the disparity of super-pixel and I_r denotes the RGB values of super-pixel.

Here, it is worthy of attention that the proposed approach just depends on the con-textual information of the adjacent super-pixels and no ambiguity or artificial factor exists.

The supporting neighbors are selected by minimizing the set of $W_i(S_p)$, $i = 1, 2, \dots, n$. Due to the several minimum (because of equal) at the same time, the supporting neighbors of super-pixel S_p are consist of all neighboring super-pixels, which can minimize the $W_i(S_p)$.

3.3 Merging Super-Pixels to Stable Region

This step aims to obtain stable region by merging super-pixels and it is divided into the following three cases:

- (a) If two neighboring super-pixels are both reliable super-pixel and their disparities are equal, then merge the two super-pixels;



(a) The merged super-pixels result of the left image.



(b) The first iteration disparity map.

Fig. 2. The first iteration: merging the super-pixels and then estimating the disparity map based on the merged result.

- (b) If the two super-pixels are both unreliable or one is unreliable region, the other is not and one is the supporting neighbor of the other one, then merge the two super-pixels;
- (c) The rest conditions will not be merged. If a super-pixel was not merged with any other super-pixel, it should be regarded as a stable region. We merge the super-pixels by using a forest structure. (The forest construction algorithm, which regards super-pixel as pixel, is similar to the ST structure algorithm in [16].)

Figure 2 gives the super-pixels merged result of the left image and the disparity map with first iteration. The experimental results show that the new segmentations are stable and our method performs well in disparity estimation.

4 The Principle of Unreliable Region Optimization

The unreliable pixels have great effects on disparity estimation. In this section, we propose a new method to deal with unreliable pixels by optimizing the unreliable region. As described in Sect. 3, before optimizing, the information and the supporting neighbors of each region must be updated.

The principles of unreliable region optimization are as follows:

- (a) Considering each unreliable region's supporting neighbors, if there is a supporting neighbor which is a reliable region, or an unreliable region which has already been optimized, then we regard the supporting neighbor as a supporting region;
- (b) If there is no supporting region of unreliable region S_u , we select the neighbor which can minimize $W_i(S_u)$ from all neighbors of S_u to be a supporting region;
- (c) If an unreliable region has more than one supporting region, selecting the supporting region with the minimum of disparity. And then we regard the selected supporting region as the final supporting region;
- (d) Assigning the final supporting region disparity for the correlative unreliable region disparity. And then set a label, which denotes the unreliable region has been optimized, to this unreliable region. Applying the four steps to all unreliable regions until each of them have been set an optimized label.

5 Depth Hypotheses Generation

In this section, we obtain the accurate disparity map by two steps. First, we adopt the constraint on smoothness to reduce the effect of spurious disparity estimation. Second, the iterative refinement is employed to enhance the accuracy of the disparity map.

5.1 The Constraint on Smoothness of Region

In order to reduce effects on spurious disparity estimation, we consider the smoothness of stable region. Usually, the depth discontinuity occurs around the boundaries of regions. Thus, the method, used to solve the smoothness problem, assigns the disparity value for each pixel by selecting the disparity from the correlative stable region

disparity, which can minimize the cost aggregation. The optimal disparity value of pixel p within super-pixel S_p can be written as:

$$\varphi_d(p) = \min\left(D_{d_i}^A(p)\right), d_i \in [D_r(S_p) - \Delta d, D_r(S_p) + \Delta d] \quad (4)$$

where Δd is a variable which determines the range of stable region disparity. If it is too small, the correct cost value may be excluded and if it is too large, the effects of spurious cost values may not be reduced. Thus we apply an adapting formulation for computing Δd , the formulation can be written as:

$$\Delta d = \frac{R}{\gamma} \quad (5)$$

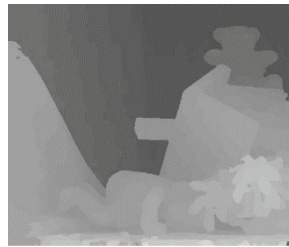
where R denotes the disparity range of image and γ is a constant which is set to six in all of our experiments. According to Eq. (4), the disparity value of pixel p is d which minimizes $D_{d_i}^A(p)$.

5.2 Enhancement with Iteration

After estimating the accurate disparity values, we can use iterative refinement to enhance the disparity estimation. As shown in Fig. 3, in the first iteration, disparity value with the best cost value is selected for each pixel, and then the robust typical disparity value can be voted for each stable region. In the next iteration, refining the disparity values by re-computing the steps from 2 to 7 based on the last iteration disparity map. New stable regions are determined and their information is updated. The best disparity values of pixels are selected only among the represent disparity value of the correlative stable regions. The final disparity values can be assigned after two iterations.



(a) The merged super-pixels result of the left image.



(b) The second iteration disparity map.

Fig. 3. The second iteration: merge super-pixels and then estimate the disparity map based on the merged result.

Figure 3 shows the second iteration segmentation result of the left image. Obviously, the experimental result performs better than the result in the first iteration (Fig. 2). In addition, in order to verify the robustness of the proposed algorithm, Fig. 4 shows the merged results of the rest stereo image pairs in the Middlebury data sets [6].

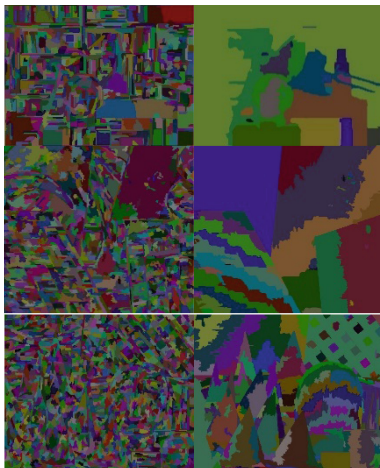


Fig. 4. The image from top to bottom is the merged super-pixels results of Tsukuba, Venus and Cones.

6 Experimental Results

The local algorithm [14] and the non-local algorithm [16] proved to be the top performer on Middlebury benchmark [6], but the results of this paper demonstrates that quantitative disparity map estimated by these algorithms can be improved by the proposed algorithm (MSP).

All experiments in this paper strictly follow a local stereo matching pipeline [5]. The specific descriptions are as follows:

- (a) Cost computation: The same cost used in the local method [14] and non-local method [16], is adopted in all our experiments. It is a blending of truncated color difference and truncated gradient difference.
- (b) Cost aggregation: Two cost aggregation methods are evaluated with various stereo data sets: local aggregation with adaptive supporting-weight (AW) [14], non-local aggregation with enhanced ST (Segment-tree) [16].
- (c) Disparity optimization: WTA (Winner-Take-All) operation is adopted in all experiments. This method simply chooses the disparity for each pixel with the minimal aggregated cost.
- (d) Disparity refinement: Based on the result of (c), applying the merged super-pixel (MSP) refinement algorithm to enhance the performance. Two parameters require to be set in this method, the parameter k is set to 0.03 and W_{occ} is set to 0.4. The final disparity map can be obtained by only iterating the proposed algorithm twice.

The disparity maps of all four stereo pairs in the Middlebury data sets computed by local method [14] are presented in Fig. 5(a). And the disparity maps obtained by the proposed algorithm, and based on the resulting disparity maps in Fig. 5(a), with different iterations, are presented in Fig. 5(b)–(c). Obviously, Fig. 5(b)–(c) show that their results are more accurate than the result in Fig. 6(a). Thus, it proves that the proposed method (MSP) is available to enhance the performance of local methods. Similarly, the proposed method (MSP) is effective to improve the performance of non-local methods. Visual comparisons in Fig. 5 show that the proposed refinement method performs better within the low textured regions. For instance, the region near the hand of teddy bear (the third row of Fig. 5) is estimated inaccurate with cost computation method (the first step of stereo matching pipeline). Both the local and non-local cost aggregation methods cannot correct these errors, but the proposed method can obtain the accurate disparity values through optimizing the unreliable region with its supporting region. Moreover, the method is more accurate around depth boundaries, such as the boundaries of the newspaper in Venus data set (the second row of Fig. 5). Errors around

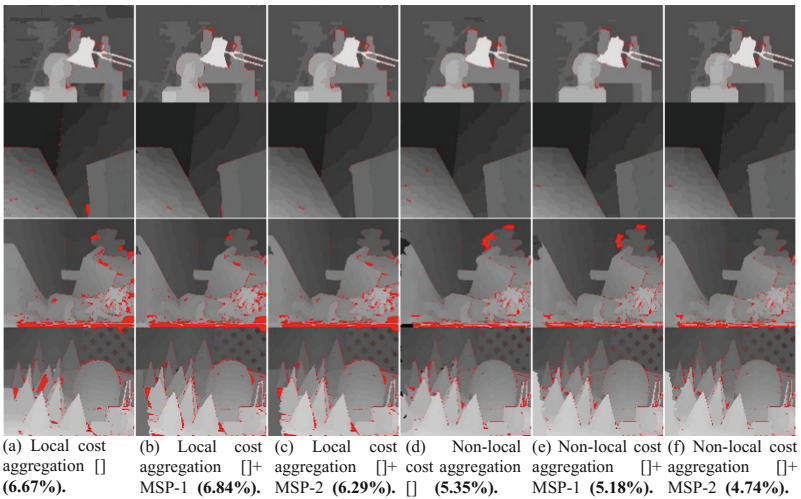


Fig. 5. Experimental results using the Middlebury data sets [6]: Tsukuba, Venus, Teddy and Cones. (a) is the disparity map obtained by using the local cost aggregation algorithm [14]. (b)–(c) are the refined results of (a) by applying MSP-1 and MSP-2 refinement method proposed in Sect. 2, respectively. (d) is the disparity map obtained by employing the non-local cost aggregation [16]. And (e)–(f) are the refined results of (d) by applying MSP-1 and MSP-2 refinement method, respectively. The bold numbers under the images are the average errors (percentages of bad pixels) which show that the significant improvement of quantitative evaluation with local and non-local stereo matching method by employing the proposed refinement method. The corresponding quantitative evaluation is summarized in Table 1. Visual comparison of the disparity maps using the local or non-local cost aggregation method without MSP or not shows that the proposed refinement method performs better around depth boundaries. For instance, the disparity estimations around the boundaries of the newspaper (the second row) in (b)–(c) or (e)–(f) are more accurate than in (a) or (d). Moreover, note that the proposed refinement can also enhance the performance in low textured regions. For example, the disparity estimations within the low texture region near the hand of teddy bear (the third row) in (b)–(c) or (e)–(f) are more accurate than in (a) or (d).

depth boundaries are mostly due to noises and would cause inconsistency, the method corrects the errors by merging super-pixels to stable region and assign the disparity value for each pixel by considering the constraint on smoothness of stable region. More details are presented in Figs. 6 and 7. According to the comparisons of the disparity estimation within zoom-in regions in Figs. 6 and 7, MSP-2 performs completely better

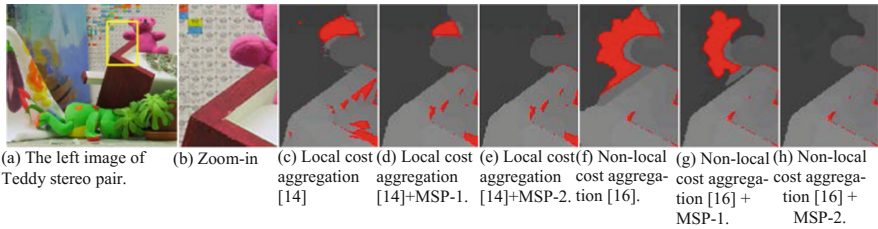


Fig. 6. (a) The left image of Teddy stereo pair from Middlebury data sets [6]. (b) The zoom-in region of yellow box. (c) The result of the local cost aggregation [14]. (d) The refined result of (c) by employing MSP once. (e) The refined result of (c) by employing MSP twice. (f) The result of the non-local cost aggregation [16]. (g) The refined result of (f) by employing MSP once. (h) The refined result of (f) by employing MSP twice. Visible comparison of the results in low textured region, (d)–(e) are more accurate than (c) and (g)–(h) are more accurate than (f), shows that the proposed refinement method is significantly available to reduce the efforts of spurious disparity values estimated by local or non-local method. (Color figure online)

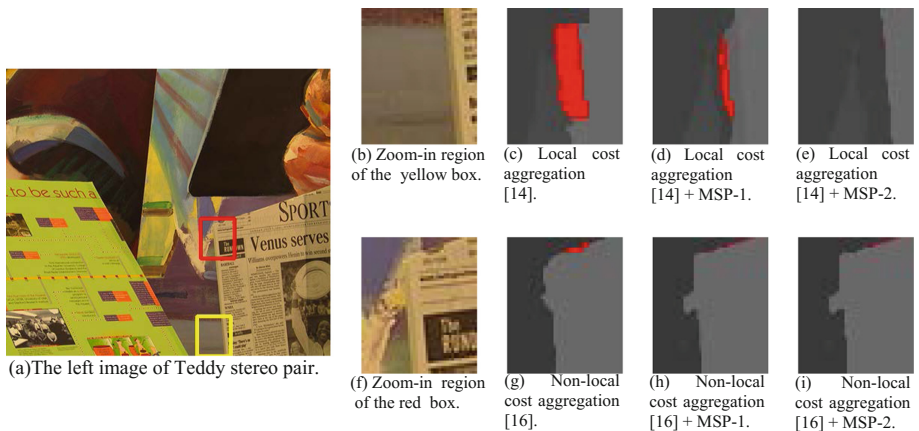


Fig. 7. (a) The left image of Venus stereo pair from Middlebury data sets [6]. (b) The zoom-in region of the yellow box. (c) The result of the local cost aggregation [14]. (d) The refined result of (c) by employing MSP once. (e) The refined result of (c) by employing MSP twice. (f) The zoom-in region of the red box. (g) The result of the non-local cost aggregation [16]. (h) The refined result of (g) by employing MSP once. (i) The refined result of (g) by employing MSP twice. Visible comparison of the results around depth boundaries, (d)–(e) are more accurate than (c) and (h)–(i) are more accurate than (g), shows that the proposed refinement method is significant available to improve the accuracy of the results estimated by local or non-local method. (Color figure online)

than local and non-local methods, with more accurate estimation both in low textured regions (shown in Fig. 6) and around depth boundaries (shown in Fig. 7).

The running time of the algorithm is related to the number of iterations. By using a PC with CPU of PM 2.5G, the total time for processing the stereo pair of Tsukuba is about 2 s. Here, the number of iterations is 2, and the time for image segmentation is about 1 s. The comparisons between the proposed refinement method and local method [14] or non-local method [16] are shown in Table 1. The average error of local method is reduced by 0.38% (from 6.67% to 6.29%) through applying the proposed method. And the rank is increased by 18.3 (from 79.5 to 61.2). The average error of non-local method [16] is reduced by 0.61% (from 5.35% to 4.74%) through using the proposed method. And the rank is increased by 13.4 (from 37.7 to 24.3). It is clear to see the significant improvement of quantitative evaluation when we replace local and non-local stereo matching method with our novel refinement method.

Table 1. Quantitative evaluation of six methods (AW [14], AW [14] + MSP-1, AW [14] + MSP-2, ST [16], ST [16] + MSP-1, ST [16] + MSP-2) on the standard middle-bury benchmark [6] with error threshold 1. The percentages of the erroneous pixels in nonocc./all/disc. regions are used to evaluate the performance of the method. The disparity estimation using local [14] or non-local [16] method with MSP-1 is slightly more accurate than only using [14] or [16], while MSP-2 outperforms the other methods with the best overall accuracy.

Algorithm	Avg. rank	Avg. error	Tsukuba			Venus			Teddy			Cones		
			Nonocc	All	Disc	Nonocc	All	Disc	Nonocc	All	Disc	Nonocc	All	Disc
AW [14]	79.5	6.67	1.38	1.85	6.90	0.71	1.19	6.13	7.88	13.3	18.6	3.97	9.79	8.26
AW [14] + MSP-1	76.9	6.84	1.44	1.69	7.52	0.28	0.54	3.38	7.49	15.0	18.9	4.28	11.4	10.1
AW [14] + MSP-2	61.2	6.29	1.43	1.69	7.49	0.17	0.32	1.89	7.30	12.6	18.4	4.09	10.5	9.70
ST [16]	37.7	5.35	1.25	1.68	6.69	0.20	0.30	1.77	6.00	11.9	15.0	2.77	8.82	7.81
ST [16] + MSP-1	32.1	5.18	1.09	1.48	5.83	0.15	0.28	1.97	5.75	12.8	14.0	2.57	8.92	7.37
ST [16] + MSP-2	24.3	4.74	1.09	1.48	5.85	0.15	0.27	1.90	4.76	9.98	12.9	2.49	8.84	7.11

7 Conclusion

This paper proposed a novel refinement algorithm for stereo matching, permits us to obtain the high-quality and high-density disparity map of a scene from its initial disparity estimation. Its novelty is reflected in the following two aspects: Novelty 1, presenting the method of merging super-pixels into stable region. Novelty 2, dealing with unreliable pixels by optimizing the unreliable region.

The advantage of this algorithm lies in that it is able to restrain and correct errors both in low textured regions and around depth boundaries, making us obtain the high-quality and high-density disparity map.

In the near future, we will focus on testing the algorithm with more challenging stereo data sets and various local or non-local cost aggregation methods.

References

1. Zbontar, J., LeCun, Y.: Computing the stereo matching cost with a convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1592–1599 (2015)
2. Sinha, S.N., Scharstein, D., Szeliski, R.: Efficient high-resolution stereo matching using local plane sweeps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1582–1589 (2014)
3. Zhang, K., Fang, Y., Min, D., et al.: Cross-scale cost aggregation for stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1590–1597 (2014)
4. Shi, C., Wang, G., Yin, X., et al.: High-accuracy stereo matching based on adaptive ground control points. *IEEE Trans. Image Process.* **24**(4), 1412–1423 (2015)
5. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms
6. Scharstein, D., Szeliski, R.: Middlebury stereo evaluation
7. Mei, X., Sun, X., Dong, W., et al.: Segment-tree based cost aggregation for stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 313–320 (2013)
8. Muninder, V., Soumik, U., Krishna, G.: Robust segment-based stereo using cost aggregation. In: Proceedings of International Conference on British Machine Vision Conference (2014)
9. Wang, H.W., Chang, M.W., Lin, H.S., et al.: Segmentation based stereo matching using color grouping. In: ACM SIGGRAPH 2014 Posters, p. 73. ACM (2014)
10. Cochran, S.D., Medioni, G.: 3-D surface description from binocular stereo. *TPAMI* **14**, 981–994 (1992)
11. Birchfield, S., Tomasi, C.: A pixel dissimilarity measure that is insensitive to image sampling. *TPAMI* **20**, 401–406 (1998)
12. Mühlmann, K., Maier, D., Hesser, J., Männer, R.: Calculating dense disparity maps from color stereo images, an efficient implementation. *IJCV* **47**, 79–88 (2002)
13. Rhemann, C., Hosni, A., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. In: CVPR (2011)
14. Yoon, K.J., Kweon, I.S.: Adaptive supporting-weight approach for correspondence search. *PAMI* **28**(4), 650–656 (2006)
15. Yang, Q.: A non-local cost aggregation method for stereo matching. In: CVPR, pp. 1402–1409 (2012)
16. Mei, X., Sun, X., Dong, W., et al.: Segment-tree based cost aggregation for stereo matching. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 313–320. IEEE (2013)
17. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph based image segmentation. *IJCV* **59** (2), 167–181 (2004)
18. Klaus, A., Sormann, M., Karner, K.: Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In: 18th International Conference on Pattern Recognition, ICPR 2006, vol. 3, pp. 15–18. IEEE (2006)
19. Wang, Z.F., Zheng, Z.G.: A region based stereo matching algorithm using cooperative optimization. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
20. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph based image segmentation. *IJCV* **59** (2), 167–181 (2004)