LETTER

Stereo Matching Based on Efficient Image-Guided Cost Aggregation

Yunlong ZHAN[†a)], ***Student Member***, **Yuzhang GU**[†], ***Nonmember***, **Xiaolin ZHANG**[†], ***Member***, **Lei QU**[†], **Jiatian PI**[†], **Xiaoxia HUANG**[†], **Yingguan WANG**[†], **Jufeng LUO**[†], ***and*** **Yunzhou QIU**[†], ***Nonmembers***

SUMMARY Cost aggregation is one of the most important steps in local stereo matching, while it is difficult to fulfill both accuracy and speed. In this letter, a novel cost aggregation, consisting of guidance image, fast aggregation function and simplified scan-line optimization, is developed. Experiments demonstrate that the proposed algorithm has competitive performance compared with the state-of-art aggregation methods on 32 Middlebury stereo datasets in both accuracy and speed.

key words: *stereo matching, guidance image, cost aggregation, guided filter, bilateral filter*

1. Introduction

Stereo matching is a hot research topic in computer vision, and numerous papers have been proposed to achieve accuracy and speed requirements. According to one survey paper [1], most algorithms could be classified into global and local methods. Although many global stereo matching algorithms could generate accurate disparity results, they cost more time due to inevitable iterations. Instead, local algorithms are adopted in many existing implementations. However, many existing local algorithms fail to produce accurate disparity maps with low computation cost because they lack efficient cost aggregation. Therefore, efficient cost aggregation is essential for local algorithms.

To get higher-accuracy disparity results, cross-based aggregation methods [2], [3] were proposed to form a shape-adaptive support region for each pixel. This method helped improve the performance on depth discontinuities and textureless regions. However, it is an expensive cost to compute the adaptive support region for each pixel. Adaptive kernel window [4] was adopted to improve the performance on depth discontinuities. The extra cost is to compute the adaptive support arms. Full-image guided cost aggregation [5] was proposed to get the surplus support information from the whole image instead of computing the support region for each pixel. In addition, methods of filtering on the cost-volume (i.e., based on guided filter [6] and based on bilateral filter [7]) have been common in recent years. These filters perform well in preserving edges, but the performance and computation cost are affected by the filters' kernel sizes. The performance of the bilateral filter based cost-volume is excellent [7], however, the computation cost of this method is relatively high. Moreover, in order to speed up the cost aggregation, the GPU implementation was later proposed [8]. This method achieves good speed, but with limited accuracy. We also find that the performance of filters' edge-preserving and homogeneous-area-smoothing are important for improving the results accuracy [6], [7], but the computation cost of the filtering-based cost-volume is also high. Being motivated by the above methods, we propose a novel image-guided aggregation method fulfilling both speed and accuracy.

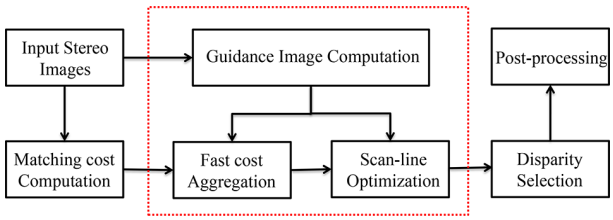

Fig. 1 Framework of local stereo matching. Our aggregation method consists of guidance image computation, fast cost aggregation and simplified scan-line optimization.

The proposed method consists of three main steps: guidance image computation, fast cost aggregation and simplified scan-line optimization, as shown in Fig. 1. The guidance image, instead of the traditional raw image, guides the cost aggregation and cost optimization, which has been little investigated in stereo vision. We mainly take the guided filter and the bilateral filter to generate the guidance images for their good edge-preserving and homogeneous-area-smoothing performance. The structure of our fast aggregation is similar to the structure in one paper [8]. But we propose a polynomial increment step strategy, which can improve the aggregation effect with fewer iterations to achieve good results. Furthermore, a simplified scan-line optimizer is applied after the fast cost aggregation to further smooth the results. By applying this novel aggregation method to local stereo matching, competitive performance can be achieved. The main features of our method, i.e., fast cost aggregation, guidance image model and simplified scan-line optimization, are detailed in Sect. 2. Experiments and discussions are presented in Sect. 3 and we conclude this letter in Sect. 4.

Manuscript received October 27, 2015.
Manuscript revised November 28, 2015.
Manuscript publicized December 9, 2015.

[†]The authors are with the Key Laboratory of Wireless Sensor Network and Communication, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China.
a) E-mail: qingqingzjin@126.com
DOI: 10.1587/transinf.2015EDL8223

2. Our Cost Aggregation Model

2.1 Guidance Image Model

In most implementations, raw input stereo image was directly (or processed with simple median filter) used in cost aggregation. There are few papers on discussing how to generate efficient enhanced content (i.e., enhanced edges) to improve the aggregation performance. Although there are some cost aggregation methods [6], [7] conducted by filtering on the 3D cost-volume, and these filters have excellent performance on edge-preserving, the computation cost is usually expensive. We found that the filters' good edge-preserving and homogeneous-area-smoothing performance could help improve results accuracy. Thus, we propose to take these filters to generate the 2D guidance image to guide the cost aggregation instead of taking the filters to aggregate the matching cost directly. In this way, the computation is reduced from 3D space to 2D space. Moreover, since the guidance image inherits the filters' performance, this image guided cost aggregation could generate relatively good results. This special image guided cost aggregation has been rarely studied in stereo vision.

To the bilateral filter [12], geometric relationship and color dissimilarity are considered in the weight strategy. Suppose pixel $q(i, j)$ is in the support window of pixel $p(x, y)$, then the filter weight $w(p, q)$ between pixel p and q is as

$$W^{bf}_{p,q}(I) = \exp(-\Delta S^2_{pq}/(2\delta_s^2) - \Delta C^2_{pq}/(2\delta_c^2)), \tag{1}$$

where ΔS_{pq} and ΔC_{pq} are the Euclidian distance of spatial distance and color dissimilarity, respectively. δ_s and δ_c are two constants. Square function is added for ΔS_{pq} and ΔC_{pq}, which is different from the weight strategy in one paper [7]. This filter weight attenuating rapidly when the spatial Euclidian distance ΔS_{pq} or the color difference in RGB space ΔC_{pq} increases, which enhances the local structures.

The guided filter [9] is widely used in stereo vision, but almost all the applications take the guided filter to process the cost-volume in three-dimensional space. Herein, we take this filter to generate the guidance image. Suppose pixels p and q are in the support window w_k centered at pixel k, and we take the gray scale guidance image I for simplified description. Then the weight between p and q is given by:

$$W^{gf}_{p,q}(I) = \frac{1}{|w|^2} \sum_{k:(p,q)\in w_k} (1 + \frac{(I_p - \mu_k)(I_q - \mu_k)}{\delta_k^2 + \epsilon}), \tag{2}$$

where δ_k and μ_k are the variance and the mean of I in the support window w_k centered at pixel k. The total number of pixels in w_k is denoted as $|w|^2$ and ϵ is a parameter to control the averaging strength.

In this letter, we mainly consider these two filters generated guidance images. The support window size of these two filters can be set relatively small, thus the computation cost, especially for the bilateral filter, can be very low.

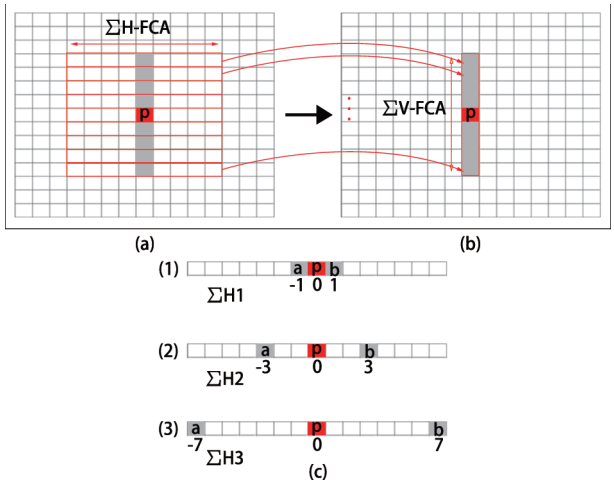

Fig. 2 The polynomial increment step cost aggregation. (a) and (b) Aggregation in horizontal and vertical directions along the 1-D space, correspondingly. (c) Aggregation in range $[-7, 7]$ with three iterations.

2.2 Fast Cost Aggregation

We define the raw matching cost as C, the guidance image as G, and the aggregated cost as C_1. As shown in Fig. 2, aggregation is decomposed into two orthogonal 1-D aggregations, namely separately aggregating in horizontal and vertical directions as the aggregation structure in one paper [8]. New step strategy is adopted to aggregate the cost with fewer iterations in our method, moreover, the guidance image is applied to provided enhanced content for the support weight.

Matching cost is aggregated in horizontal direction according to Figs. 2 (a), followed by the vertical aggregation as shown in Figs. 2 (b). In such a way, aggregation is computed in two 1-D support areas, rather than the traditional ones directly in the 2-D support areas, which could largely reduce the computation cost. Figures 2 (c) shows an example of aggregating cost in horizontal direction with three iterations in range $(-7, 7)$. Suppose the basic step size is r, then the step size in the next iteration is a polynomial increment $2r + 1$, namely offset $= 1, 3, 7, 15$ for the first four iterations. The corresponding aggregating range is 1, 4, 11, 26. This step size is designed to aggregate all the cost between the offset pixels without optimizing training. Some overlaps are allowed between different iterations for fully extracting the local information and controlling the step size not to increase too quickly, because the information near the center pixel is more representative than the far away pixels' in some way.

In each iteration, three pixels (i.e., center pixel p and pixels a and b at $-r$ and $+r$ offset) are taken into consideration. In this way, only $3 * 3 = 9$ discrete pixels are needed to compute the aggregated cost for the center pixel in $(-11, 11)$ aggregating range. Aggregation in the vertical direction can be computed in the similar way. Different iterations can be employed according to specific applications. Moreover, adaptive support weight strategy is adopted for

the aggregation between the center pixel and the offset pixels. The adaptive weight $w(p, q)$ considers the geometric relationship and color dissimilarity between the center pixel $p(x, y)$ and the offset pixel $q(i, j)$. The guidance image G, instead of the raw image, is adopted to compute this weight based on

$$w_{p,q}(G) = \exp(-\Delta S_{pq}/\lambda_s - \Delta C_{pq}/\lambda_c), \quad (3)$$

where λ_s and λ_c are two constant parameters. This weight is computed based on the guidance image G instead of the traditional raw input image to offset the sampling in the aggregation. At each iteration, the target pixel's cost at disparity d is updated as follows:

$$C_1(p, d) = C(p, d) + w(p, p + r)C(p + r, d) + w(p, p - r)C(p - r, d). \quad (4)$$

2.3 Simplified Scan-Line Optimization

To further remove the ambiguities in matching cost, a simplified scan-line optimizer with smoothness constraints is applied after the fast cost aggregation. This simplified scan-line optimization is based on Hirschmüller's semi-global methods [10]. But only four simplified scan-line processes are adopted and these processes are independently in orthogonal directions, i.e., two along vertical direction and two along horizontal direction. Suppose the scan-line direction is θ, then the cost $C_2^\theta(p, d)$ of pixel p at disparity d is updated according to:

$$C_2^\theta(p, d) = C_1(p, d) + min(C_2^\theta(p_\theta, d), C_2^\theta(p_\theta, d \pm 1) + p_1, \min_{k \in D} C_2^\theta(p_\theta, k) + p_2) - \min_{k \in D} C_2^\theta(p_\theta, k), \quad (5)$$

where D is the disparity range, and p_θ is the previous station of pixel p alone the direction θ. p_1 and p_2 ($p_1 \leq p_2$) [3] are the penalty terms for smoothness and they are defied as follows:

$$\begin{cases} p_1 = \pi_1, p_2 = \pi_2, if AD_1 < T, AD_2 < T, \\ p_1 = \pi_1/5, p_2 = \pi_2/5, else if AD_1 > T, AD_2 > T, \\ p_1 = \pi_1/3, p_2 = \pi_2/3, otherwise, \end{cases} \quad (6)$$

where π_1, π_2 and T are constant parameters. AD_1 is the color difference between the previous station pixel p_θ and current pixel p in the left guidance image G_L. Similarly, AD_2 is the color difference in the right guidance image G_R. Finally, we generate the smoothed cost $C_3(p, d)$ according to

$$C_3(p, d) = \frac{1}{4} \sum_\theta C_2^\theta(p, d). \quad (7)$$

3. Experiment Results

Experiments are carried out on the Middlebury stereo benchmark [1] to demonstrate the performance of proposed cost aggregation. Four standard images pairs (i.e., Teddy, Cones, Venus, and Tsukuba) (defined as $M4$) are taken into consideration. Furthermore, except *Midd*1, *Midd*2, *Monopoly* and *Plastic*, 28 more datasets (defined as $M28$) in 2001, 2003, 2005 and 2006 datasets on Middlebury stereo benchmark are also all considered. Thus, experiments are carried out on total 32 stereo datasets ($M32$). Our method is implemented with both bilateral filter and guided filter generated guidance images, i.e., OurBF and OurGF methods, respectively. Moreover, four state-of-art cost aggregation functions, i.e., aggregation based on cross-skeleton structure (Cross) [2], aggregation based on exponential step structure (ExpStep) [8], aggregation based on guided filter based filtering (GFiltering) [6], and aggregation based on full-image guided strategy (FullImage) [5], are also implemented in our experiments for comparison. During the evaluation, all these functions are performed in the cost aggregation step, and all the other steps (i.e., cost computation, disparity selection and post-processing) are kept the same. Cost computation is computed based on the AD-gradient measure as defined in the paper [6]. The 'winner-takes-all' strategy is used to select the disparity. The cross checking and a weighted median filter are performed as the post-processing.

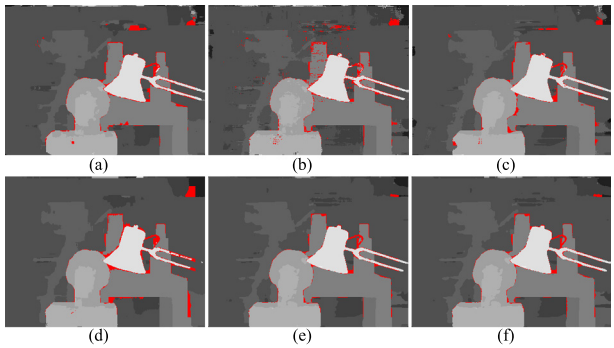

Fig. 3 Experimental results on Tsukuba dataset. Red pixels are the mismatched ones in *all* region. (a) Based on Cross [2], (b) based on ExpStep [8], (c) based on GFiltering [6], (d) based on FullImage [5], (e) based on OurBF, (f) based on OurGF.

Parameters for our aggregation functions are set as $\{\lambda_s, \lambda_c\}$={14/255,14/255}. We set $\{\delta_s, \delta_c\}$={3, 0.3} for the bilateral filter and $\epsilon = 0.10$ for the guided filter. The parameters for simplified scan-line optimization, i.e., $\{T, \pi_1, \pi_2\}$, have some slight adjustments according to different guidance images. We set them as {11/255, 0.5/255, 15/255} for the guided filter and {11/255, 0.8/255, 17/255} for the bilateral filter. Our approach has been implemented on a 3.0GHz CPU processor with MATLAB implementation.

Figure 3 and Fig. 4 show the generated disparity maps based on different cost aggregation functions. We can observe that the low-texture regions in our disparity maps are filled with the exact values and smoothed well, e.g., the background on Tsukuba and Art, while these regions are filled with some outlier blocks based on the other methods. Edge preserving is also difficult for many local stereo matching methods. We can see that the edges in our proposed methods are also preserved well, e.g., the edges of the desk and the carton on Tsukuba, however, the other methods exist some irregularities along the edges. These results

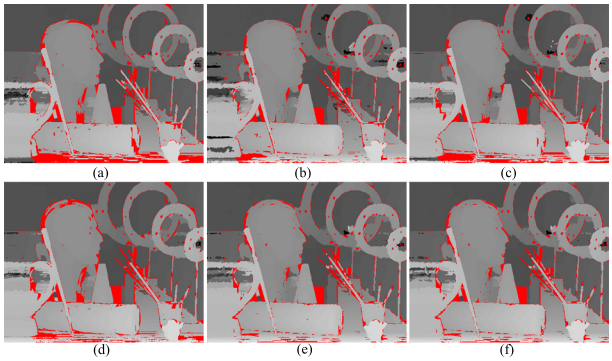

Fig. 4 Experimental results on Art dataset. Red pixels are the mismatched ones in *nonocc* region. (a) Based on Cross [2], (b) based on ExpStep [8], (c) based on GFiltering [6], (d) based on FullImage [5], (e) based on OurBF, (f) based on OurGF.

Table 1 Experimental results of different cost aggregation methods

Aggregation Method	M4 Avg. Error	M32 Avg. *All* Error	M32 Avg. *Nonocc* Error	Avg. Time
OurBF	5.26%	9.03%	5.60%	1.00
OurGF	5.34%	9.05%	5.62%	1.02
SSMP [11]	5.38%	9.56%	6.13%	1.04
GFilter [6]	5.82%	9.41%	6.02%	2.64
Cross [2]	6.37%	11.07%	7.79%	1.98
FullImage [5]	6.43%	10.13%	6.68%	0.21
ExpStep [8]	8.51%	10.51%	6.47%	0.42

demonstrate that the proposed method is good for preserving the edges and smoothing the homogeneous areas in disparity maps. The guidance image partly contributes to this performance, because our guidance image inherits some properties of the bilateral filter and the guided filter.

Quantitative evaluation results on both $M4$ and $M32$ are provided in Table 1. M4 error is the average error on *all*, *nonocc* and *disc* regions. The average aggregation time is a relative time compared with OurBF method. All the methods are experimented with the same conditions except the aggregation functions. Moreover, the results of recent proposed state-of-art method (SSMP) [11] are also provided for comparison. According to Table 1, the proposed methods achieves the lowest error percentages compared with all the other methods, which demonstrates that our method is competitive with all the other methods in accuracy. OurBF even achieves higher accuracy compared with OurGF, from which we can see that the bilateral filter is a little more robust in extracting local information. Our method is competitive with SSMP [11]. More importantly, the computation cost of the proposed methods is also competitive with the other methods [2], [6], [11]. The computation cost of bilateral filter based cost-volume filtering method [7] was expensive with relatively large kernel window size, thus, the guided filter based cost-volume filtering method [6] was proposed later to reduce the computation cost. OurGF method (with guided filter) has sharply reduced the time cost compared with the guided filter based cost-volume filtering method [6] on CPU. Moreover, the speed of the bilateral filter based OurBF method is even a little faster than the guided filter based OurGF method. In this way, our method has some competitiveness on speed. These results demonstrate the effectiveness of the proposed method. Furthermore, parallel optimization could also be taken to further speedup our method.

4. Conclusion

A new cost aggregation method is proposed for local stereo matching. This new solution consists of guidance image, fast cost aggregation and simplified scan-line optimization. The guidance images inherit the performance of the guided filter and the bilateral filter, which fulfill the requirements of edge preserving with low complexity. The fast cost aggregation could aggregate the cost with an efficient step strategy. The simplified scan-line optimization could further smooth the disparity maps. We have applied this method in local stereo matching. Experimental results demonstrate that our method is competitive with all the other methods in both accuracy and speed. Moreover, our method is also suitable for parallel. Efforts will be made to speed up this method to achieve real-time stereo matching on GPU.

References

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," International Journal of Computer Vision, vol.47, no.1-3, pp.7–42, 2002.
[2] K. Zhang, J. Lu, and G. Lafruit, "Cross-based local stereo matching using orthogonal integral images," IEEE Trans. Circuits Syst. Video Technol., vol.19, no.7, pp.1073–1079, July 2009.
[3] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang, "On building an accurate stereo matching system on graphics hardware," Proc. IEEE ICCV Workshops, pp.467–474, Nov. 2011.
[4] Q. Yang, P. Ji, D. Li, S. Yao, and M. Zhang, "Fast stereo matching using adaptive guided filtering," Image and Vision Computin, vol.32, no.3, pp.202–211, 2014.
[5] Q. Yang, D. Li, L.H. Wang, and M. Zhang, "Full-Image Guided Filtering for Fast Stereo Matching," IEEE Signal Process. Lett., vol.20, no.3, pp.237–240, March 2013.
[6] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," Proc. IEEE, CVPR, pp.3017–3024, June 2011.
[7] K.-J. Yoon and I.S. Kweon, "Adaptive support-weight approach for correspondence search," IEEE Tran. Pattern Matching and Machine Intelligence, vol.28, no.4, pp.650–656, April 2006.
[8] W. Yu, T. Chen, and J.C. Hoe, "Real time stereo vision using exponential step cost aggregation on gpu," Proc. IEEE ICIP, pp.4281–4284, Nov. 2009.
[9] K. He, J. Sun, and X. Tang, "Guided image filtering," IEEE Trans. Pattern Anal. Mach. Intell., vol.35, no.6, pp.1397–1409, June 2013.
[10] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," IEEE Trans. Pattern Anal. Mach. Intell., vol.30, no.2, pp.328–341, Feb. 2008.
[11] B. Ham, D. Min, C. Oh, M.N. Do, and K. Sohn, "Probability-based rendering for view synthesis," IEEE Trans. Image Process., vol.23, no.2, pp.870–884, Feb. 2014.
[12] C. Tomasi and R. Manduchi, "Bilateral Filtering for Gray and Color Images," Proc. Int'l Conf. Computer Vision, pp.839–846, 1998.