An improved semi-global matching method with optimized matching aggregation constraint

Xu Huang¹², Yilong Han³⁵ and Kun Hu⁴

¹School of Geospatial Engineering and Science, Sun Yat-sen University, Rongyuan Road 4, Zhuhai, China;
²Wuhan Engineering Science and Information Institute, Jiangda Road 30, Wuhan, China;
³School of Remote Sensing and Information Engineering, Luoyu Road 129, Wuhan, China;
⁴Key Laboratory of Technology in Geo-spatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, No. 19 North 4th Ring Road West, 100190, Beijing, China

Email: hanyl@whu.edu.cn

Abstract. Stereo dense image matching (DIM) is a key technique in generating dense 3D point clouds at low cost, among which semi-global matching (SGM) is one of the best compromise between the matching accuracy and the time cost. Most commercial or open-source DIM software packages therefore adopt SGM as the core algorithm for the 3D point generation, which computes matching results in 2D image space by simply aggregating the matching results of multi-directional 1D paths. However, such aggregations of SGM did not consider the disparity consistency between adjacent pixels in 2D image space, which will finally decrease the matching accuracy. To achieve higher-accuracy while keep the high time efficiency of SGM, this paper proposes an improved SGM with a novel matching aggregation optimization constraint. The core algorithm formulates the matching aggregation as the optimization of a global energy function, and a local solution of the energy function is utilized to impose the disparity consistency between adjacent pixels, which is capable of removing noises in the matching aggregation results and increasing the final matching accuracy at low time cost. Experiments on aerial image dataset show that the proposed method outperformed the traditional SGM method and another improved SGM method. Compared with the traditional SGM, our proposed method can increase the average matching accuracy by at most 11%. Therefore, our proposed method can applied in some smart 3D applications, e.g. 3D change detection, city-scale reconstruction, and global survey mapping.

1. Introduction

Dense image matching (DIM) is to find pixelwise correspondences between stereo pairs. DIM is a traditional but still interesting topic in remote sensing and photogrammetry communities. DIM is able to compute dense 3D point clouds at low cost, thus having fueled many smart 3D applications, e.g. global survey mapping, city-scale reconstruction, virtual reality, 3D change detection, robot navigation, and some RGB-D data-based works [1-4].

Semi-global matching (SGM) is the most popular DIM method, which formulates DIM as the optimization of a global energy function, then breaks the global optimization into multi-directional 1D
optimizations along 1D paths in image space and finally computing the approximate solutions of the global energy function by simply aggregating the matching results of the 1D optimizations [5]. In general, the workflow of SGM concludes: (1) matching cost computation, which formulates the appearance (e.g. intensities, textures) similarity of a matcher between stereo pairs as matching cost; (2) matching cost propagation, which conducts pixelwise matching by 1D optimization in a dynamic programming function; (3) matching results aggregation, which simply accumulates 1D matching results from multiple 1D paths; and (4) matching results computation and refinement, which compute the matching results with the minimum aggregation results. If epipolar image pairs are available in stereo dense image matching, the matching results of each pixel are their disparities (or the parallax in the epipolar space). Though proposed in 2007, SGM is still one of the best compromise between the matching accuracy and the time cost. Some state-of-the-art matching algorithms [6-8] achieved higher matching accuracy than SGM, but they also have much higher matching time cost. The high time cost of these state-of-the-art algorithms limit their applications in city-scale reconstruction. Therefore, most image-based 3D reconstruction software packages, e.g. Photoscan, SURE, ENVI, Erdas, Micmac, still adopt SGM as the core algorithm for city-scale 3D reconstruction [9]. However, the 1D optimization results of SGM are not robust, since they lack the matching constraints in the perpendicular direction of the 1D path.

To improve the matching accuracy of SGM, some work has been done to improve the workflow of SGM. The matching cost computation step of SGM can be improved by using convolutional neural network (CNN) [10-11], combined cost metrics [12-13], or local filtering methods [14-16] to compute more accurate matching cost. The matching cost propagation of SGM is unreliable, due to its 1D optimization paths. The unreliability problem of the 1D optimization paths of SGM can addressed by more global cost propagation paths [17-18] or more robust 1D optimization strategy [19-20]. The final matching results of SGM can also be improved through post processing. They either remove the noises on the surface of the matching results [21-22] or refine the building boundaries in the matching results [23-24]. All the above methods are able to improve the matching accuracy of SGM. However, few work payed attention to the matching results aggregation step. Since traditional SGM only keep the local disparity consistency along the 1D paths, the matching aggregation results of adjacent pixels in 2D image space may not be consistent, which will decrease the final matching accuracy.

To further improve the matching accuracy of SGM while keep its high time efficiency, this paper focuses on improving the matching aggregation of SGM, and proposed an efficient matching optimization aggregation method. The core algorithm formulates the matching aggregation problem as the optimization of a global energy function, and utilized a local solution to impose disparity consistency between adjacent pixels in 2D image space, which is able to improve the final matching accuracy at low time cost.

The remaining of the paper is organized as follows. Section 2 introduces the methodology of the proposed method, Section 3 show the comparison experiments between SGM and the proposed method, and Section 4 concludes the whole paper.

2. Methodology

2.1. Brief introduction to sgm

SGM[5] is one of the most famous matching methods in remote sensing and photogrammetry communities. It formulates the matching problem as the optimization of a global energy function, and breaks the global optimization into sub-optimization in multi-directional 1D paths, as shown in Figure 1. In each 1D path, it applies a dynamic programming in Equation (1) for matching cost aggregation, where some window-based matching cost metrics (e.g. census [25] in this paper) can be used as matching cost due to their robust performance against luminance variation.
Figure 1. 1D optimizations of SGM in multiple paths. \( P \) is a pixel in image, which has eight cost propagation results from the eight-directional paths [5].

\[
L_r(p, d) = C(p, d) + \min \left\{ \begin{array}{l}
L_r(p - r, d) \\
L_r(p - r, d - 1) + P_1 \\
L_r(p - r, d + 1) + P_1 \\
\min_{i} L_r(p - r, i) + P_2
\end{array} \right\} - \min_{k} L_r(p - r, k)
\]  

where, \( L_r(p, d) \) is the cost propagation results of pixel \( p \) at disparity \( d \) in the direction \( r \); \( d \) is a disparity in epipolar space. Given epipolar stereo pairs, the matcher only have differences in column coordinates: \( x_r = x_l - d \) with \( x_l, x_r \) being the column coordinates in the left and the right images. \( C(p, d) \) is the matching cost of pixel \( p \) at disparity \( d \); \( p - r \) is the previous pixel of \( p \) along the 1D path in direction \( r \); \( P_1 \) and \( P_2 \) are penalty terms that penalize disparity variations between adjacent pixels, where \( P_2 > P_1 \); \( i, k \) are any disparity values.

After the matching cost propagation in multi-paths, SGM simply aggregates the cost propagation results \( L_r(p, d) \) in all directions \( r \), as shown in Equation (2). The pixelwise matching results are then derived from the matching aggregation results by using winner-takes-all strategy.

\[
S(p, d) = \sum_r L_r(p, d)
\]  

where, \( S(p, d) \) is the matching aggregation results of pixel \( p \) at disparity \( d \).

2.2. Matching aggregation optimization

2.2.1. Problem formulation. SGM only imposes local disparity consistency between adjacent pixels along 1D paths, therefore, it often meets streak problems in the matching results of each single path, as shown in Figure 2. Figure 2 shows that the direction of the streaks is consistent with the cost propagation directions, which will greatly decrease the matching accuracy. In the matching aggregation step, SGM assign equal weights to the cost propagation results of all directions. Therefore, the final disparity of adjacent pixels may not be consistent in the 2D image space, due to the ambiguity among the cost propagation results of all directions.

Figure 2. Matching results in different directions.
To solve the ambiguity in the matching aggregation step, we impose disparity smoothness constraints over the whole image, and formulate the matching aggregation problem as the optimization of a global energy function, as shown in Equation (3).

\[ E(D) = \sum_p S(p, d) + \sum_{p,q \in N} P \cdot |d_p - d_q| \]  

(3)

where, \( E(D) \) is the global energy function with \( D \) being the set of disparities of all pixels; \( N \) is the set of all adjacent pixels in 2D image space; \( q \) is an adjacent pixel of \( p \); \( d_p \) is the disparity of \( p \); \( P \) is a penalty term. The global energy function means that the final matching results should have small aggregation results and consistent disparities with their adjacent pixels.

2.2.2. Approximate solution. The global energy function in Equation (3) is a typical NP-hard problem. We can use bilateral filtering [16] to compute its local optima. Mozerov et al. [26] also proved such local filter is the approximate solution of the global energy function in Equation (3). However, the performance of the bilateral filtering depends on the filtering window sizes. Large window will bring high computational cost, while small window may not improve the final matching accuracy. To compute the solution of Equation (3) in a polynomial time while obtain high matching accuracy, we finally utilize a fast bilateral filtering strategy [27] in the matching aggregation optimization. In general, fast bilateral filtering formulates the traditional bilateral filtering as the non-local cost propagations in orthogonal scanning lines. The cost in the fast bilateral filtering is the matching aggregation results of SGM. In each scanning line, the cost propagation is conducted with the underlying assumptions that adjacent pixels with similar intensities should have the same disparities. Fast bilateral filtering only considers the intensity factor in the matching optimization. However, in intensity homogenous but disparity inconsistent regions, such assumption of the fast bilateral filtering will cause mismatches. Therefore, we add a spatial control factor in the propagation function of the fast bilateral filtering so that the matching of each pixel will not be influenced by the ones that are far away. The fast bilateral filtering function with the spatial control factor is shown as follows.

\[ L_r'(p, d) = S(p, d) + w_S \cdot w_I(p, p - r) \cdot L_r(p - r, d) \]  

(4)

\[ w_I(p, q) = \exp(-|I_p - I_q|/\sigma_l) \]

where, \( r \) is the direction of propagation path (e.g. the horizontal direction); \( L_r' \) is the propagated cost in direction \( r \); \( p \) is a pixel in the basic image; \( p - r \) is the previous pixel of \( p \) in direction \( r \); \( w_I(p, p - r) \) is the intensity weight, which is inversely proportional to the intensity difference between \( p \) and \( p - r \). \( w(p, p - r) \) is normally defined as Gaussian function. \( w_S \) is a spatial control factor, which is set as 0.7 in all experiments.

In fast bilateral filtering, the cost of each pixel is propagated through non-local, orthogonal scanning lines. The cost propagation firstly proceeds in both horizontal (\( r = 0 \)) and the corresponding inverse direction (termed as -\( r \)) in Equation (4), and then the propagated cost in these directions are aggregated to get the sub-optimization in horizontal directions, as shown in Equation (5).

\[ S_r'(p, d) = L_r'(p, d) + L_{-r}'(p, d) - S(p, d) \]  

(5)

where \( L_r \) and \( L_{-r}' \) are propagated cost in \( r \) and -\( r \) directions; \( S_r' \) is the aggregated cost which is the sum of propagated cost in both directions. Since both \( L_r'(p, d) \) and \( L_{-r}'(p, d) \) contain the cost of \( p \), \( S(p, d) \) is subtracted in Equation (5) for eliminating the redundancy.

Similar to the cost propagation in \( r \) and -\( r \) directions in Equation (4), the aggregated cost \( S_r' \) then is regarded as the new cost for propagation in the perpendicular direction \( r_\perp \) and -\( r_\perp \). The propagated cost in \( r_\perp \) and -\( r_\perp \) are finally summed to yield the aggregated cost \( S_{r_\perp}'(p, d) \) in a similar manner of Equation (5). \( S_{r_\perp}'(p, d) \) is the final aggregation result in the first \( r \) and then \( r_\perp \) directions.
After the optimization of the fast bilateral filtering, the final matching results of each pixel can be derived from the optimized aggregation results $S_{ij}(p,d)$ through a winner-takes-all strategy.

3. Experiment

Our proposed method was tested on aerial dataset in Vaihingen, which consists of 6 stereo pairs, as shown in Figure 3. The dataset was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) in the international society for photogrammetry and remote sensing (ISPRS) benchmark, captured by an Intergraph / ZI digital mapping camera (DMC) (GSD: 0.08m) and a Leica airborne laser system (ALS)50 system (point density: 4 points/m2). True disparities of corresponding stereo pairs were generated by filtering inconsistent parts of light detection and ranging (LiDAR) points (e.g. occlusions, temporal changes) in a dense matching framework [28] and then projecting the consistent LiDAR points onto epipolar space.

In general, we respectively used the traditional SGM [5], an improved SGM which provided more global propagation paths for SGM [18], and the proposed method to estimate the disparities of the six stereo pairs. We then evaluated their matching performances by (1) computing the percentage of matches whose matching errors are smaller than 2 pixels (also termed as Acc-2), (2) averaging these matching errors (also termed as Acc-avg), and (3) counting their running time. The matching accuracy of the three methods are shown in Figure 4. In all experiments, the above three methods proceeded with only one single CPU configuration (i7 @2.60GHz).

![Figure 3](image_url) Testing data in our experiment.

Acc-avg measures the average matching errors for each stereo pair, thus lower Acc-avg represents higher matching accuracy. On the other hand, Acc-2 measures the percentages of good pixels whose matching errors were smaller than 2 pixels, thus higher Acc-2 means higher matching accuracy. In Figure 4, SGM performed worst in both Acc-avg and Acc-2 metrics, since it only used 1D optimization paths and did not consider the ambiguity in the matching aggregation. The more global SGM used more global propagation paths for matching, thus achieving higher matching accuracy than SGM. However, it still assign equal weights to all propagation directions in the matching aggregations. The ambiguity in cost propagation results of each direction will decrease the final matching accuracy. The proposed method adopted the same propagation paths as SGM (i.e. the eight cost propagation paths), while it efficiently optimized the matching aggregation results to reduce the ambiguity, which is able to improve the final matching accuracy. In both metrics, the proposed method outperformed
SGM and the more global SGM. Compared with SGM, the proposed method can improve the average matching accuracy by at most 12.93%.

![Graph](image)

**Figure 4.** Matching accuracy comparisons.

The time cost of different matching algorithms is also an important factor in city-scale reconstruction. The DIM algorithms with high matching accuracy but high time cost are not fit for the city-scale reconstruction. Thus, we also compared the time cost of the above three DIM methods. The time cost on different stereo pairs is shown as follows.

![Graph](image)

**Figure 5.** Time cost of different DIM algorithms.

Figure 5 shows that SGM achieved the lowest time cost. Since the proposed method additionally optimized the matching aggregation results through the fast bilateral filtering technique, it had a little higher time cost than SGM. In the DIM of stereo pairs (pair 2 - 6) with the image size as 1500x1500 pixels, the average additional filtering time is only 8.20 s under only one single CPU configuration. If a parallel framework can be applied, the filtering time can be further reduced by several times. The more global SGM method achieved the highest time cost, since it optimized matching results through more global and more complicated paths. In overall consideration of the Figure 4 and Figure 5, the proposed method is able to achieved the best matching accuracy and the second best time cost. Therefore, it is a good compromise between the matching accuracy and the time cost.

For more comprehensive comparisons between SGM and the proposed method, we selected the matching results of SGM and the proposed method in some typical regions (i.e. grounds and buildings), as shown in Figure 6.
The proposed method

**Figure 6.** Visual comparisons between SGM and the proposed method. The matching results in (a), (b), and (c) are selected from the third stereo pair in Figure 3.

Figure 6 shows that the surfaces in the SGM matching results are noisy, especially in the grounds and the gable roofs. It is because the multi-directional propagation results contained ambiguities, due to their local disparity smoothness constraints along the 1D paths. Our proposed method locally optimized the matching aggregation results to reduce the ambiguity, thus achieving much smoother surfaces in the matching results.

However, the proposed method is essentially a pixelwise matching algorithm, which can get high matching accuracy in textured regions. However, in weak textured regions, the matching window of a single pixel may contain high matching uncertainties, which will decrease the final matching accuracy. Therefore, the proposed method may perform badly in some weak textured regions, e.g. rivers, lakes, and snow.
4. Conclusion
This paper mainly focuses on improving the matching aggregation step of SGM, and proposes a locally optimized matching aggregation constraints to improve the final matching accuracy. It is the first time to optimize the matching aggregation step of SGM in a local manner. Experiments on aerial image dataset show that the proposed method outperformed the traditional SGM and the more global SGM. Compared with the traditional SGM, it can improved the average matching accuracy by at most 11.36%. In the future work, we plan to further improve the matching accuracy by formulating the matching aggregation problem of SGM as a classification problem, and selecting the best matching propagation result in the eight propagation directions as the final matching result.

Acknowledgments
This work is supported by the National Natural Science Foundation (Award No. 41701540, 41871368). We would like to thank the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) for providing the Vaihingen dataset.

Reference
[1] Wang H, Sun Y, Liu M, et al. 2019 Self-Supervised Drivable Area and Road Anomaly Segmentation Using RGB-D Data For Robotic Wheelchairs IEEE Robotics and Automation Letters 4(4) 4386-4393
[2] Sun Y, Liu M, Meng MQ, et al. 2017 Improving RGB-D SLAM in Dynamic Environments: A Motion Removal Approach Robotics and Autonomous Systems 89 110-122
[3] Sun Y, Liu M, Meng MQ, et al. 2018 Motion removal for reliable RGB-D SLAM in dynamic environments Robotics and Autonomous Systems 108 115-128
[4] Sun Y, Liu M, Meng MQ, et al. 2019 Active Perception for Foreground Segmentation: An RGB-D Data-Based Background Modeling Method IEEE Transactions on Automation Science and Engineering 16(4) 1596-1609
[5] Hirschmuller H 2007 Stereo Processing by Semiglobal Matching and Mutual Information IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2) 328-341
[6] Taniai T, Matsushita Y, Sato Y, et al. 2018 Continuous 3D label stereo matching using local expansion moves IEEE Transactions on Pattern Analysis and Machine Intelligence 40(11) 2725-2739
[7] Li L, Zhang S, Yu X, et al. 2018 PMSC: PatchMatch-Based Superpixel Cut for Accurate Stereo Matching IEEE Transactions on Circuits and Systems for Video Technology 28(3) 679-692
[8] Park M, and Yoon K 2019 As-planar-as-possible depth map estimation Computer Vision and Image Understanding 181 50-59
[9] Han Y, Qin R, Huang X 2020 Assessment of dense image matchers for digital surface model generation using airborne and spaceborne images – an update The Photogrammetric Record 35(169) 58-80
[10] Žbontar J, Lecun Y 2016 Stereo matching by training a convolutional neural network to compare image patches Journal of Machine Learning Research 17(1) 2287-2318
[11] Huang X, Liu W, and Qin R 2020 A Window Size Selection Network for Stereo Dense Image Matching International Journal of Remote Sensing 41(12) 4836-4846
[12] Mei X, Sun X, Zhou M, et al. 2011 On building an accurate stereo matching system on graphics hardware International Conference on Computer Vision 467-474
[13] Taniai T, Matsushita Y, Sato Y, et al. 2018 Continuous 3D label stereo matching using local expansion moves. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(11) 2725-2739
[14] Lin Y, Lu N, Lou X, Zou F, Yao Y, & Du Z 2013 Matching Cost Filtering for Dense Stereo Correspondence Mathematical Problems in Engineering 2013 1-11
[15] Yang Q, Li D, Wang L, & Zhang M 2013 Full-Image Guided Filtering for Fast Stereo Matching IEEE Signal Processing Letters 20(3) 237-240
[16] Tomasi C, Manduchi R 1998 Bilateral filtering for gray and color images *International Conference on Computer Vision* 839-846
[17] Facciolo G, De Franchis C, & Meinhardt E 2015. MGM: A Significantly More Global Matching for Stereovision *British machine vision conference*
[18] Huang X, Zhang Y, Yue Z, et al. 2016 Image-guided Non-local Dense Matching with Three-Steps Optimization *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 67-74
[19] Seki A, Pollefeys M 2017 SGM-Nets: Semi-Global Matching with Neural Networks *Computer Vision and Pattern Recognition* 6640-6649
[20] Schonberger J L, Sinha S N, Pollefeys M, et al. 2018 Learning to Fuse Proposals from Multiple Scanline Optimizations in Semi-Global Matching *European Conference on Computer Vision* 758-775
[21] Huang X, Zhang Y, Zhou G, et al. 2017 Global Depth Refinement Based on Patches. *International conference on machine learning* 423-433
[22] Zhang Y, Zhang Y, Zhang Y, Zhao Z 2017 A Two-Step Semiglobal Filtering Approach to Extract DTM from Middle Resolution DSM *IEEE Geoscience and Remote Sensing Letters* 14(9) 1599-1603
[23] Qin R, Chen M, Huang X, Hu K 2019 Disparity refinement in depth discontinuity using robustly matched straight lines for digital surface model generation *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12(1) 174-185
[24] Zhihui X 2009 Dense Matching Algorithm Based on Region Boundary Restriction via Graph Cuts Optimization *Computer Science* 2009
[25] Zabih R, and Woodfill J 1994 Non-parametric local transforms for computing visual correspondence *European conference on computer vision* 151-58
[26] Mozerov M and van de Weijer J 2015 Accurate stereo matching by two-step energy minimization *IEEE Transactions on Image Processing* 24(3) 1153-1163
[27] Yang Q 2012 Recursive bilateral filtering *European Conference on Computer Vision* 399-413.
[28] Huang X, Qin R, Xiao C, and Lu X 2018 Super resolution of laser range data based on image-guided fusion and dense matching *ISPRS Journal of Photogrammetry and Remote Sensing* 144 105-118