Rectified Neighborhood Construction for Robust Feature Matching With Heavy Outliers

Yizhang Liu, Brian Nlong Zhao, and Shengjie Zhao, *Senior Member, IEEE*

***Abstract*— This letter is concerned with constructing reliable neighborhoods for the local consistency-based feature matching methods. To alleviate the impact of outliers on neighborhood construction, we propose a rectified neighborhood construction (RNC) strategy, which can effectively enlarge the distribution between inliers and outliers. Besides, we also integrate an adaptive parameter estimation into the aforementioned rectified strategy, and it can contribute to determining a reasonable parameter for the rectified strategy. Finally, the experimental results on two representative remote sensing image datasets show that the proposed method can achieve satisfactory feature matching results compared with some state of the arts.**

***Index Terms*— Feature matching, heavy outliers, motion coherence, rectified neighborhood construction (RNC).**

I. Introduction

FEATURE matching plays an important role in various vision-based tasks, such as information fusion [1], [2] and 3-D reconstruction [3]. Accurate feature matching can provide a good initialization for high-level applications. Feature matching mainly includes feature detection, feature description, feature matching, and outlier removal [4], [5]. In this letter, we focus on the outlier removal problem, namely, selecting inliers from the initial putative matches and removing the outliers. However, the matching problem is essentially an NP-hard problem, and it often suffers from noise, viewpoint changes, small overlap, and nonrigid deformation, resulting in many outliers existing in the putative matches. Therefore, it is significant to propose a robust method to filter out the outliers.

In order to construct initial putative matches, several methods have been proposed, such as classical scale-invariant feature transform (SIFT) [6], speeded up robust features (SURF) [7], oriented FAST and rotated BRIEF (ORB) [8], deep learning-based learned-invariant feature transform (LIFT) [9], and SuperPoint (self-supervised interest point detection and description) [10]. No matter what descriptors are used, the ambiguity of which is inevitable. Thus, the putative matches often consist of outliers. Researchers have proposed many methods to accomplish outlier removal. Random sample consensus (RANSAC) [11] and its variants are resampling-based methods, which aim to find a clean subset without outliers by resampling. They can achieve satisfactory results when the transformation between images is rigid, and the major component of the putative matches is inliers. Robust feature matching using spatial clustering with heavy outliers (RFMSCAN) [12] is an unsupervised method that integrates spatial clustering into the outlier removal. Grid-based motion statistics (GMS) [13] and locality preserving matching (LPM) [14] are local consistency-based methods, and their mathematical models are based on the motion consistency theory [15], saying inliers are consistent with their neighboring matches and outliers are randomly distributed among the images. Therefore, the motion statistics of matches in a small region can be used to separate inliers from outliers. It is noticed that neighborhood construction is crucial for the local consistency-based methods. For GMS, it uses grids with fixed sizes to represent the neighborhood of feature points, and for LPM, it uses k nearest neighbors. Meanwhile, they use a multiscale neighborhood construction strategy to maintain the stability of the methods. Nevertheless, the multiscale strategy cannot work well for high proportions of outliers.

Motivated by the above observation, this letter aims to enhance the reliability of the neighborhood construction of feature points and derive a solution to the motion consistency-based mathematical model. The main contributions of this letter can be summarized as follows: 1) different from [13], [14], a rectified neighborhood construction (RNC) strategy is developed in the letter, by which the neighborhood information statistics of the feature points are more reliable; 2) an adaptive parameter estimation strategy is proposed, which can contribute to determining a reasonable parameter for the RNC strategy; and 3) compared with some state of the arts, our method achieves comparable and even superior matching results.

II. Method

Given two images, we can use the off-the-shelf feature detection and description method (e.g., SIFT) to construct initial putative matches $S = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i and y_i are the coordinates of two matched feature points and N is the number of putative matches. According to [14], the

Manuscript received 9 April 2022; revised 19 May 2022; accepted 6 June 2022. Date of publication 23 June 2022; date of current version 29 August 2022. This work was supported in part by the National Key Research and Development Project under Grant 2019YFB2102300, in part by the National Natural Science Foundation of China under Grant 61936014, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100, and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding author: Shengjie Zhao.)*

Yizhang Liu and Shengjie Zhao are with the School of Software Engineering, Tongji University, Shanghai 201804, China (e-mail: lyz8023lyp@gmail.com; shengjiezhao@tongji.edu.cn).

Brian Nlong Zhao is with the Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: briannlongzhao@gmail.com).

Digital Object Identifier 10.1109/LGRS.2022.3181754

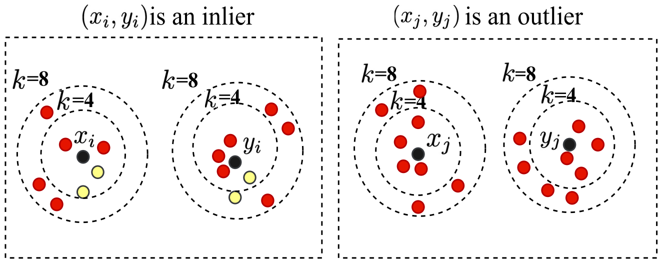

Fig. 1. Comparison of the neighborhood construction used in LPM and our method. Black: feature points to be matched. Red: feature points of outliers. Yellow: feature points of inliers.

outlier removal problem can be accomplished by optimizing the objective function

$$\mathcal{C}(\mathbf{p}; S, \lambda) = \sum_{i=1}^{N} \frac{p_i}{k} \sum_{j|x_j \in \mathcal{N}_{x_i}^{k}} d\left(y_i, y_j\right) + \lambda\left(N - \sum_{i=1}^{N} p_i\right) \quad (1)$$

with

$$d\left(y_i, y_j\right) = \begin{cases} 0, & y_j \in \mathcal{N}_{y_i}^{k} \\ 1, & y_j \notin \mathcal{N}_{y_i}^{k} \end{cases} \quad (2)$$

where $\mathcal{N}_{y_i}^{k}$ denotes the neighborhood of y_i consisting of its k nearest neighbors, $\mathbf{p} = [p_1, p_2, \ldots, p_N]$ is a binary vector representing the correctness of each putative match, and $\lambda > 0$ is a parameter that balances the first and second terms of this objective function.

For simplicity, we denote $k - \sum_{j|x_j \in \mathcal{N}_{x_i}} d(y_i, y_j)$ as n_i. Carefully observing (1) and (2), we can find that n_i is essentially the number of neighboring matches with regard to the putative match (x_i, y_i). It is noticed that $(k - n_i/k)$ changes slightly by using multiscale neighborhood construction. More specifically, with increased k, $(k - n_i/k)$ may remain the same or increase accordingly. Therefore, the ratio of $k - n_i$ to k changes slightly. The purpose of using a multiscale strategy is to enlarge the distribution between inliers and outliers. However, it does not seem to work well.

Since the feature detection and description on two images are performed independently, when the putative matches contain lots of outliers, the neighborhood information statistics are unreliable. For example, for an inlier (x_i, y_i), $\mathcal{N}_{x_i}^{k}$ and $\mathcal{N}_{y_i}^{k}$ should have the same elements without considering outliers and noise. However, in the case of heavy outliers, they have a small number of common elements, leading to a large value of $\sum_{j|x_j \in \mathcal{N}_{x_i}} d(y_i, y_j)$. Since the value of $\sum_{j|x_j \in \mathcal{N}_{x_i}} d(y_i, y_j)$ with regard to an outlier is large, it is difficult to use a predefined threshold to separate inliers from outliers.

In order to solve the aforementioned problem, we propose an RNC strategy, which is based on the assumption that inliers are consistent with each other in a small region and outliers are distributed randomly among the images. Specifically, given a fixed k, we choose to construct $k + \varepsilon$ nearest neighbors for each feature point, and the objective function in (1) turns out to be

$$\mathcal{C}(\mathbf{p}; S, \lambda, \varepsilon) = \sum_{i=1}^{N} \frac{p_i}{k} \min\left\{ \sum_{j|x_j \in \mathcal{N}_{x_i}^{k+\varepsilon}} d\left(y_i, y_j\right), k \right\} + \lambda\left(N - \sum_{i=1}^{N} p_i\right). \quad (3)$$

From (3), we can find that the neighborhood information statistics $\sum_{j|x_j \in \mathcal{N}_{x_i}^{k+\varepsilon}} d(y_i, y_j)$ of feature points are based on the $k + \varepsilon$ nearest neighbors, but the normalization of the cost value is based on k instead of $k + \varepsilon$. To better illustrate the effectiveness of the RNC, we show an example in Fig. 1. Here, (x_i, y_i) is an inlier and the cost value calculated by LPM is 0.75 both for $k = 4$ and 8. By contrast, the cost value calculated by the RNC with $\varepsilon = 4$ is 0.5. For the outlier (x_j, y_j), the cost value calculated by LPM and the RNC is 1. Therefore, we can find that for inliers, the RNC will help to find those consistent matches that were missed due to the existence of heavy outliers by decreasing the cost value. Conversely, for outliers, the number of their neighboring matches n_i will not be greatly affected by the increased neighborhood size. Thus, the margin between inliers and outliers will be enlarged to some extent.

The parameter ε means the degree of the increased neighborhood size. A fixed value of ε may not adapt to complex neighborhood situations. In order to solve this problem, we propose an adaptive parameter estimation method. Here, we denote the ranking list $\sigma_{x_i}^{k}$ and $\sigma_{y_i}^{k}$ consisting of k nearest neighbors of feature point x_i and y_i, respectively, as

$$\sigma_{x_i}^{k} = \left[x_1^i, x_2^i, \ldots, x_k^i\right], \quad \sigma_{y_i}^{k} = \left[y_1^i, y_2^i, \ldots, y_k^i\right]. \quad (4)$$

Let $d_i = \hat{d}(x_i, x_k^i) - \hat{d}(y_i, y_k^i)$, where $\hat{d}(\cdot)$ returns the Euclidean distance between the elements. If $d_i \geq 0$, we search $k+\varepsilon$ nearest neighbors for y_i such that $\hat{d}(y_i, y_{k+\varepsilon}^i) \leq \hat{d}(x_i, x_k^i)$ and $\hat{d}(y_i, y_{k+\varepsilon+1}^i) > \hat{d}(x_i, x_k^i)$. Similarly, if $d_i < 0$, we search $k + \varepsilon$ nearest neighbors for x_i such that $\hat{d}(x_i, x_{k+\varepsilon}^i) \leq \hat{d}(y_i, y_k^i)$ and $\hat{d}(x_i, x_{k+\varepsilon+1}^i) > \hat{d}(y_i, y_k^i)$. As described above, the value of ε is adaptively determined based on the range covered by the neighborhoods.

Besides, considering that the aforementioned minimization problem ignores the topological structure of the putative match and its neighboring matches, we also use the consensus of neighborhood topology proposed in [14]

$$s\left(v_i, v_j\right) = \frac{\min\left\{|v_i|, |v_j|\right\}}{\max\left\{|v_i|, |v_j|\right\}} \cdot \frac{\left(v_i, v_j\right)}{|v_i| \cdot |v_j|} \quad (5)$$

where v_i and v_j are vectors associated with putative match (x_i, y_i) and one of its neighboring matches, respectively. Similarly, the quantized distance between v_i and v_j is as follows:

$$d\left(v_i, v_j\right) = \begin{cases} 0, & s\left(v_i, v_j\right) \geq \tau \\ 1, & s\left(v_i, v_j\right) < \tau \end{cases} \quad (6)$$

where τ is a predefined threshold. Following [14], considering the multiscale neighborhood representation, the objective function turns out to be

$$\begin{aligned}&\mathcal{C}(\mathbf{p}; S, \lambda, \tau, \varepsilon)\\&= \sum_{i=1}^{N} \frac{p_i}{M} \sum_{m=1}^{M} \frac{1}{k_m} \left(\min\left\{ \sum_{j | x_j \in \mathcal{N}_{x_i}^{k_m+\varepsilon}} d\left(y_i, y_j\right), k_m \right\} \right.\\&\qquad\left. + \min\left\{ \sum_{j | x_j \in \mathcal{N}_{x_i}^{k_m+\varepsilon}, y_j \in \mathcal{N}_{y_i}^{k_m+\varepsilon}} d\left(v_i, v_j\right), k_m \right\} \right)\\&\quad + \lambda\left(N - \sum_{i=1}^{N} p_i\right) \end{aligned} \tag{7}$$

where $1/M$ is used to normalize the contribution of each scale neighborhood. We can reorganize the objective function by merging the terms related to p_i

$$\mathcal{C}(\mathbf{p}; S, \lambda, \tau, \varepsilon) = \sum_{i=1}^{N} p_i(c_i - \lambda) + \lambda N \tag{8}$$

where

$$\begin{aligned}c_i &= \sum_{m=1}^{M} \frac{1}{M k_m} \left(\min\left\{ \sum_{j | x_j \in \mathcal{N}_{x_i}^{k_m+\varepsilon}} d\left(y_i, y_j\right), k_m \right\} \right.\\&\qquad\left. + \min\left\{ \sum_{j | x_j \in \mathcal{N}_{x_i}^{k_m+\varepsilon}, y_j \in \mathcal{N}_{y_i}^{k_m+\varepsilon}} d\left(v_i, v_j\right), k_m \right\} \right).\end{aligned} \tag{9}$$

It is obvious that any match with a cost smaller than λ will result in a negative term and decrease the objective function. Conversely, any match with a cost larger than λ will lead to a positive term and increase the objective function. Thus, the correctness of each putative match can be determined by the following criterion:

$$p_i = \begin{cases} 1, & c_i \le \lambda \\ 0, & c_i > \lambda, \end{cases} \quad i = 1, \ldots, N. \tag{10}$$

Actually, if the neighborhood can be constructed based on a subset with high ratio inliers, the neighborhood information statistics can also be more accurate. Therefore, similar to [16], [17], we also use an iterative strategy to refine the neighborhood construction. Specifically, by calculating p_i for each putative match, we can determine a subset with high ratio inliers, denoted as $\mathbf{I}_1$ according to (10). Then, the neighborhood construction for each feature point is based on $\mathbf{I}_1$ instead of the initial putative set. In this iteration, the calculation of p_i is more reliable and the inlier set $\mathbf{I}^*$ can be determined according to (10).

III. Experimental Results

To validate the effectiveness of the proposed method, we compare it with some state of the arts on the feature matching task. RANSAC [11] is a resampling-based method, and robust feature matching using spatial clustering with heavy outliers (RFMSCAN) [12] is an unsupervised clustering-based method. LPM [14] and GMS [13] are the motion consistency-based methods. The parameter settings are the same as the original papers. We use the publicly available dataset [18] and [19] to test the performance of the methods, and the

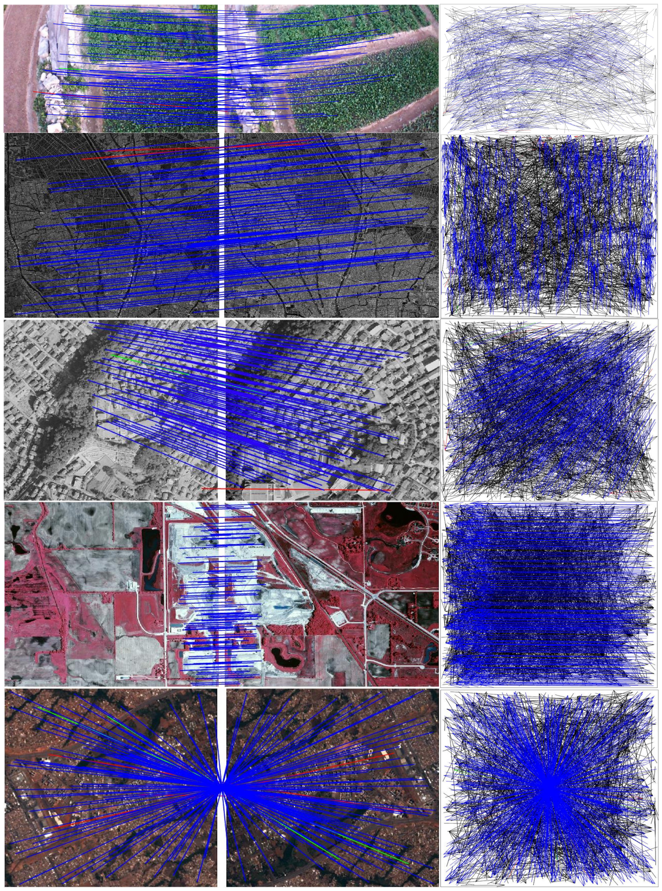

Fig. 2. Feature matching results on five representative image pairs. For each group, (left) the matching results and (right) the corresponding vector field (blue: true positive, black: true negative, green: false negative, and red: false positive). For visibility, at most 100 correspondences are shown randomly.

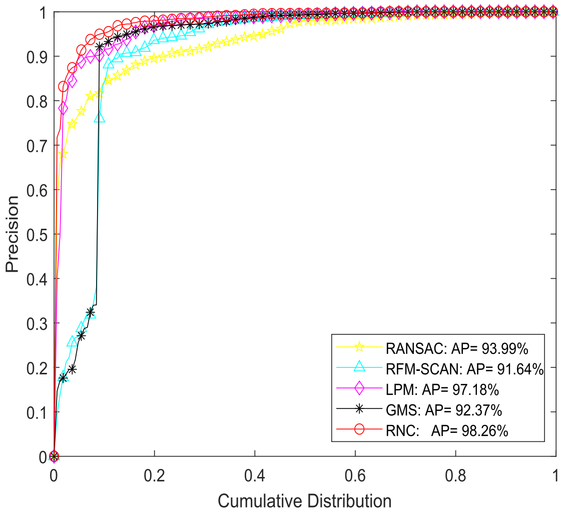

Fig. 3. Precision statistics with respect to the cumulative distribution. A point (x, y) on the curve denotes that there are $(100 \times x)\%$ of image pairs whose precision does not exceed y.

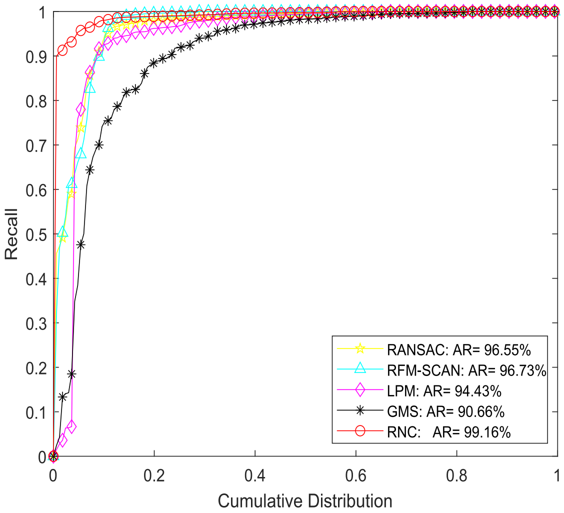

Fig. 4. Recall statistics with respect to the cumulative distribution. A point (x, y) on the curve denotes that there are $(100 \times x)\%$ of image pairs whose recall does not exceed y.

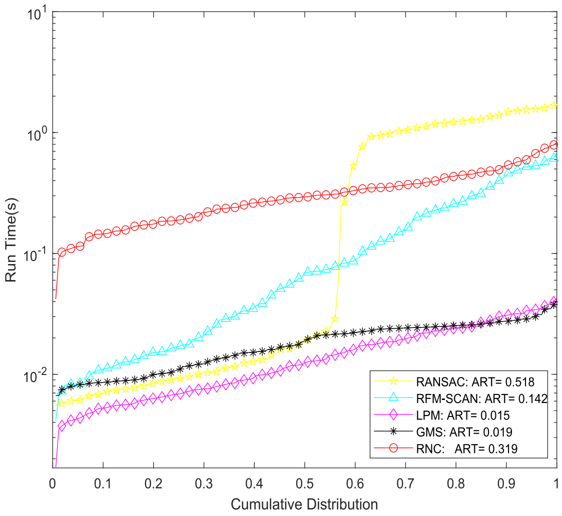

Fig. 6. Run time statistics with respect to the cumulative distribution. A point (x, y) on the curve denotes that there are $(100 \times x)\%$ of image pairs whose run time does not exceed y.

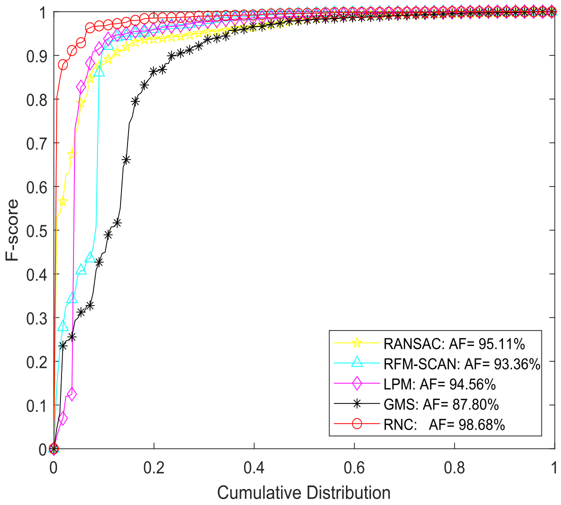

Fig. 5. F-score statistics with respect to the cumulative distribution. A point (x, y) on the curve denotes that there are $(100 \times x)\%$ of image pairs whose F-score does not exceed y.

precision, recall, and F-score are used as the performance metrics. The number of image pairs is 167, and the average match number is 869.4. Besides, the SIFT feature provided by VLFeat toolbox [20] is used to construct initial putative matches for the urban change detection for multispectral earth observation using convolutional neural networks (UCD) dataset [19]. To ensure objectivity, we manually check the correctness of each match. The parameter setting of the proposed method is that in the first iteration, $k = [8, 10, 12]$, $\lambda = 0.9$, and $\tau = 0.2$. In the second iteration, $k = [6, 8, 10]$, $\lambda = 0.5$, and $\tau = 0.2$.

A. Qualitative Results

We show the feature matching results on five representative image pairs with projective distortion, heavy noise, repetitive structure, small overlap, and large rotation changes. The numbers of putative matches with regard to the five image pairs are 724, 1814, 1537, 2115, and 1219 and the corresponding initial inlier ratios are 49.7%, 40.5%, 28.9%, 17.0%, and 32.4%. The performance statistics (precision, recall, and F-score) of the proposed method on the five image pairs are (99.16%, 98.89%, and 99.03%), (99.23%, 99.36%, and 99.29%), (99.55%, 98.65%, and 99.10%), (100.0%, 100.0%, and 100.0%), and (96.78%, 98.23%, and 97.50%). From Fig. 2, we can find that the proposed method can identify most of the inliers and only some matches are misjudged even if the putative matches contain heavy outliers.

B. Quantitative Results

We compare the proposed method with four state of the arts, namely, RANSAC [11], RFMSAN [12], LPM [14], and GMS [13]. Technically, LPM and GMS are the methods that are most relevant to the proposed method. We present the precision, recall, and F-score statistics with respect to the cumulative distribution in Figs. 3–5. Carefully observing Fig. 4, we can find that the proposed method can identify almost inliers and the RNC strategy makes the inliers more detectable in the case of heavy outliers. The performance of RANSAC extremely depends on a specific parameter model and it often does not work for the multiple consistencies. RFM-SCAN can cluster the matches with similar motion consistency, while the relatively slack constraint makes it misjudge several matches. The formulation of LPM is only based on motion consistency, and thus, it can adapt to image pairs with complex transformations. However, its neighborhood construction is easily affected by heavy outliers, which makes the neighborhood information statistics unreliable. GMS divides the image pairs into disjoint grids and according to the motion statistics of each grid pair, and the correspondences between grids can be determined. Nevertheless, the motion statistics just consider the interaction between neighborhood elements and ignore their topological structure. In addition, a fixed size of grid may separate one structure into several parts, resulting in missing some inliers. In short, the proposed method is relatively robust to outliers and it can also fit in image pairs with complex degradations.

We also present the run time statistics of the competitors on the whole dataset in Fig. 6. We find that LPM and GMS achieve approximately real-time performance on the outlier removal task. The computational complexity of RANSAC depends on the inlier ratio of the initial putative set. With higher inlier ratio, RANSAC is easier to converge and obtain a satisfactory result. Conversely, with lower inlier ratio, it needs more iterations to find a solution. For the proposed RNC method, since it needs to adaptively rectify the neighborhood construction of the feature points, the parameter ε estimation is somewhat time-consuming. In short, the RNC can also accomplish outlier removal from thousands of putative matches in a few milliseconds.

IV. Conclusion

This letter aims to conquer the problem that the neighborhood information statistics are unreliable when a large number of outliers exist in the putative matches. Therefore, an RNC strategy is proposed, which can enlarge the margin between inliers and outliers to a certain degree. In addition, in order to adapt to complex neighborhood situations, we propose an adaptive parameter estimation strategy. The experimental results show the effectiveness of the proposed method. In the future, we will consider reducing the computational complexity of the method from the perspective of parallel computing.

References

[1] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2020.

[2] Y. Xing, Y. Zhang, S. Yang, and Y. Zhang, "Hyperspectral and multispectral image fusion via variational tensor subspace decomposition," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[3] Y. Peng, M. Yang, G. Zhao, and G. Cao, "Binocular-vision-based structure from motion for 3-D reconstruction of plants," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[4] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, pp. 23–79, Aug. 2021.

[5] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Inf. Fusion*, vol. 73, pp. 22–71, 2021.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[7] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. 9th Eur. Conf. Comput. Vis.*, vol. 3951, May 2006, pp. 404–417.

[8] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[9] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 467–483.

[10] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 224–236.

[11] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[12] X. Jiang, J. Ma, J. Jiang, and X. Guo, "Robust feature matching using spatial clustering with heavy outliers," *IEEE Trans. Image Process.*, vol. 29, pp. 736–746, 2020.

[13] J.-W. Bian *et al.*, "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1580–1593, Jun. 2020.

[14] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *Int. J. Comput. Vis.*, vol. 127, no. 5, pp. 512–531, 2019.

[15] A. L. Yuille and N. M. Grzywacz, "A mathematical analysis of the motion coherence theory," *Int. J. Comput. Vis.*, vol. 3, pp. 155–175, Jun. 1989.

[16] Y. Liu *et al.*, "Robust feature matching via advanced neighborhood topology consensus," *Neurocomputing*, vol. 421, pp. 273–284, Jan. 2021.

[17] Y. Liu *et al.*, "Motion consistency-based correspondence growing for remote sensing image matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[18] J. Ma, J. Jiang, H. Zhou, J. Zhao, and X. Guo, "Guided locality preserving feature matching for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4435–4447, Aug. 2018.

[19] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *Proc. Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 2115–2118.

[20] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1469–1472.