Multi-Modal Remote Sensing Image Matching Considering Co-Occurrence Filter

Yongxiang Yao, Yongjun Zhang, Yi Wan, Xinyi Liu, Xiaohu Yan, and Jiayuan Li

Abstract— **Traditional image feature matching methods cannot obtain satisfactory results for multi-modal remote sensing images (MRSIs) in most cases because different imaging mechanisms bring significant nonlinear radiation distortion differences (NRD) and complicated geometric distortion. The key to MRSI matching is trying to weakening or eliminating the NRD and extract more edge features. This paper introduces a new robust MRSI matching method based on co-occurrence filter (CoF) space matching (CoFSM). Our algorithm has three steps: (1) a new co-occurrence scale space based on CoF is constructed, and the feature points in the new scale space are extracted by the optimized image gradient; (2) the gradient location and orientation histogram algorithm is used to construct a 152-dimensional log-polar descriptor, which makes the multi-modal image description more robust; and (3) a position-optimized Euclidean distance function is established, which is used to calculate the displacement error of the feature points in the horizontal and vertical directions to optimize the matching distance function. The optimization results then are rematched, and the outliers are eliminated using a fast sample consensus algorithm. We performed comparison experiments on our CoFSM method with the scale-invariant feature transform (SIFT), upright-SIFT, PSO-SIFT, and radiation-variation insensitive feature transform (RIFT) methods using a multi-modal image dataset. The algorithms of each method were comprehensively evaluated both qualitatively and quantitatively. Our experimental results show that our proposed CoFSM method can obtain satisfactory results both in the number of corresponding points and the accuracy of its root mean square error. The average number of obtained matches is namely 489.52 of CoFSM, and 412.52 of RIFT. As mentioned earlier, the matching effect of the proposed method was significantly greater than the three state-of-art methods. Our proposed CoFSM method achieved good effectiveness and robustness. Executable programs of CoFSM and MRSI datasets are published: https://skyearth.org/publication/project/CoFSM/**

Manuscript received July 28, 2021; revised February 4, 2022; accepted February 27, 2022. Date of publication March 14, 2022; date of current version March 21, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42030102, Grant 42192583, Grant 42001406, and Grant 62102268; in part by the Fund for Innovative Research Groups of the Hubei Natural Science Foundation under Grant 2020CFA003; and in part by the China Postdoctoral Science Foundation under Grant 2020M672416. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chandra Sekhar Seelamantula. *(Corresponding authors: Yongjun Zhang; Yi Wan.)*

Yongxiang Yao, Yi Wan, Xinyi Liu, and Jiayuan Li are with the School of Remote Sensing Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: yaoyongxiang@whu.edu.cn; yi.wan@whu.edu.cn; liuxy0319@whu.edu.cn; ljy_whu_2012@whu.edu.cn).

Yongjun Zhang is with the School of Remote Sensing Information Engineering, Wuhan University, Wuhan 430079, China, and also with the Hubei Luojia Laboratory, Wuhan 430079, China (e-mail: zhangyj@whu.edu.cn).

Xiaohu Yan is with the School of Artificial Intelligence, Shenzhen Polytechnic, Shenzhen 518055, China (e-mail: yanxiaohu@szpt.edu.cn).

Digital Object Identifier 10.1109/TIP.2022.3157450

Index Terms— **Multi-modal remote sensing image, matching, co-occurrence filter, new image gradient, log-polar descriptor.**

I. Introduction

IMAGE matching is the process of aligning two or more images with overlapping ranges, which were taken by the same or different modals of sensors at the same or different shooting angles[1]. In the case of Multi-modal Remote Sensing Image (MRSIs), (e.g., infrared, multispectral, synthetic aperture radar (SAR), unmanned aerial vehicle (UAV) image, high resolution, and multispectral imagery), obtaining satisfactory matching results is a challenging problem [2]. Because their imaging mechanisms are different, MRSIs produce different images for the same target, resulting in significant nonlinear radiation distortions (NRD) and geometric differences, which makes it difficult to obtain a successful match. Given that MRSI data plays an important role in target detection, disaster assessment, illegal building detection, and monitoring of changes in land and resources, resolving this issue with MRSI matching method is of major importance.

Due to the NRDs and geometric differences in multi-modal images, obtaining accurate matching corresponding points is difficult. As an alternative, the image texture edge information is known to better preserve the image information, which could be helpful for extracting the similar features of multi-modal images. However, image texture edge processing mainly operates in image scale space. In the traditional scale space construction, the image is generally noisy, which is controlled by Gaussian blur. Although this operation reduces the image noise to a certain extent, the edge information between different textures in the image is weakened. Therefore, finding a way to better preserve or enhance the edge information between textures in the scale space construction to achieve the feature matching of multi-modal images would be very beneficial.

The co-occurrence filter (CoF), which is an edge-preserving filter by Jevnisek *et al.* [3] has been shown to distinguish the edge within the image texture effectively. The pixel values that appear frequently in the image will have a higher weight in the co-occurrence matrix, which can smooth the image texture without considering the intensity difference. The pixel values that rarely appear in the image at the same time will have a lower weight in the co-occurrence matrix, will not smooth across the texture boundary, and can better preserve the boundary within the image texture area. At the same time, CoF has no parameters and has good co-occurrence information collection capabilities for different images. It is helpful to apply CoF to the matching of MRSIs to reduce the NRD differences and to extract more effective edge information. Therefore, a CoF image scale space constructed has the advantages of image edge feature retention, which can increase the detection effect of the points, could lessen the difference between the descriptors and achieve a better multi-modal image matching effect.

In the study presented in this paper, a MRSI matching method based on CoF was created. This paper proposed two contributions.

(1) A new image scale space is constructed by the proposed co-occurrence filtering algorithm, which can increase the number of feature points extracted from most modalities.

(2) The proposed CoFSM algorithm involves the low-pass butterworth filter (LPBF) when generating the image gradient and optimizes the partition of the log-polar descriptor grids, which both enhance the robustness of the feature descriptor across multi-modal images.

II. Related Work

Image matching is the key to remote sensing image processing and is widely used in many fields such as change detection, image mosaic, aerial triangulation, 3D reconstruction, and medical image analysis.

Traditional image matching mainly focuses on intensity and features. The methods based on image intensity include shape context, mutual information, and pixel intensity [4], but they cannot achieve a good effect when there is a large difference in image intensity. The methods based on image features [5] include scale-invariant feature transform (SIFT), speeded up robust features (SURF), oriented fast and rotated brief (ORB), enlarged descriptor window, multi-directional assignment of key points, enhanced feature matching algorithm, fast sample consensus algorithm, pattern search [6] and progressive feature matching [7]. Among them, Bellavia [8] *et al.* proposed an improved sGLOH algorithm, which improved the robustness of the descriptor and achieved robust matching. Jin [9] *et al.* also proposed a wide baseline matching benchmark method, which analyzed the performance of different matching methods and provided a reference for the application of different algorithms. They explore image matching from the perspective of scale robustness, rotation invariance, binary description optimization, image nonlinear diffusion, etc. The feature matching methods are effective for images with linear distortion differences, among which the SIFT algorithm is the most robust. Therefore, feature matching is more widely used, and the advantages of feature matching methods has been demonstrated [10], but in the case of large changes in time and geometry, the algorithm does not perform well. At the same time, MRSI images with larger NRDs are more sensitive to grayscale and gradient. These factors reduce the correlation between the corresponding images; therefore, the advantages of the feature matching methods cannot be exploited in MRSI matching, which often leads to matching difficulties.

A great deal of researches has been conducted in this area. The SAR-SIFT algorithm was proposed by Dellinger *et al.* [11] to solve the matching problem of SAR images. This method has a good effect on remote sensing images to some extent; but when the RNDs of the image are large, the matching fails. A new similarity measurement for image matching based on shape attributes was proposed by Ye *et al.* [12]. Although their method matches the image with good effect by constructing a self-similarity shape descriptor within the image, the applicability of the algorithm is restricted because it relies heavily on the image contour or shape. Recently, the Histogram of Orientated Phase Congruency algorithm was proposed to optimize MRSI matching [13]. Good results can be obtained, but the method is limited by the accuracy of the geographic information. The Radiation-Invariant Feature Transform (RIFT) method, which was proposed by Li *et al.* [14], mainly uses phase congruency and builds a Maximum Index Map to weaken the NRD difference of the MRSI. However, the RIFT method does not support scale differences in images. The extended phase correlation algorithm based on log Gabor filtering was proposed by Xie *et al.* [15], which could better optimize the matching difficulty caused by NRD and large-scale difference, but this method relies on structural information from the imagery, which would otherwise be restricted. Wu *et al.* [16] proposed fast visual highlighting and descriptor rearrangement methods. This method can obtain the corresponding points faster, but it is not applicable to large scale change of images; Liu *et al.* [17], [18] proposed local frequency information multimodal matching. This method can effectively resist image noise and scene distortion, but the method does not have the image alignment function; Zhao *et al.* [19] proposed a multimodal SURF algorithm (MM-SURF), this method has high operating efficiency and matching accuracy. However, when few feature points are extracted, MM-SURF will not achieve the prospective result; and Zhao *et al.* [20] proposed a multimodal images method based on multimodality robust line segment descriptor (MRLSD). The MRLSD method has the advantages of high applicability and high accuracy, but the long running time is its biggest shortcoming. As mentioned earlier, although the above methods have achieved certain results, they still have a series of problems such as high time consumption, non-support of geometric transformation, limitation of image structure information, insufficient key point extraction, and high computational cost.

Deep learning methods also have been developing rapidly, such as convolutional neural network to generate feature descriptors [21], [22], multi-scale neural network [23], multi-relation attention network [24], similarity supervision [25], hybrid convolutional neural network [26] and recursive cascade network [27]. These methods use deep learning to supervise and match part of the multi-modal image matching and then gradually realize end-to-end unsupervised matching, constantly enriching the matching methods. However, the computational complexity and generalization capacity of these methods are still in need of study.

In summary, good progress has been made in advancing multi-modal image matching from the classical feature methods to similarity matching, to phase correlation methods and deep learning methods, but MRSI matching methods still have challenges, which are mainly concentrated in three areas:

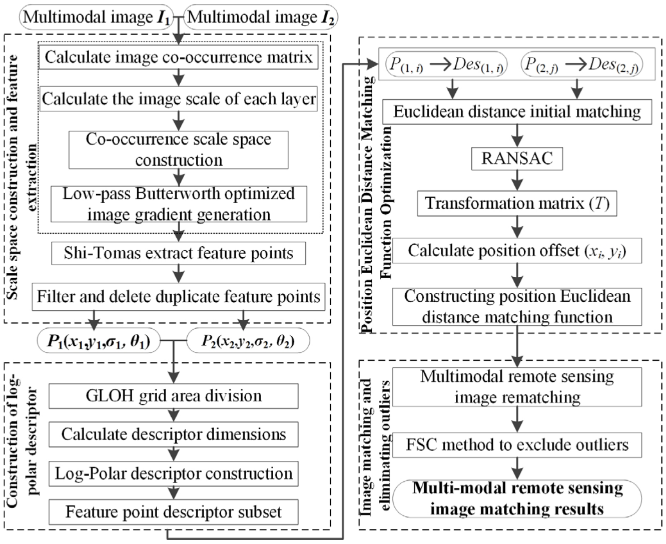

Multimodal image I_1
Multimodal image I_2
Scale space construction and feature extraction
Calculate image co-occurrence matrix
Calculate the image scale of each layer
Co-occurrence scale space construction
Low-pass Butterworth optimized image gradient generation
Shi-Tomas extract feature points
Filter and delete duplicate feature points
$P_1(x_1,y_1,\sigma_1,\theta_1)$
$P_2(x_2,y_2,\sigma_2,\theta_2)$
Construction of log-polar descriptor
GLOH grid area division
Calculate descriptor dimensions
Log-Polar descriptor construction
Feature point descriptor subset
Position Euclidean Distance Matching Function Optimization
$P_{(1,i)} \rightarrow Des_{(1,i)}$
$P_{(2,j)} \rightarrow Des_{(2,j)}$
Euclidean distance initial matching
RANSAC
Transformation matrix (T)
Calculate position offset (x_i, y_i)
Constructing position Euclidean distance matching function
Image matching and eliminating outliers
Multimodal remote sensing image rematching
FSC method to exclude outliers
Multi-modal remote sensing image matching results

Fig. 1. Multi-modal image matching process of CoFSM method.

(1) the current multi-modal image feature point extraction is fewer, and the edge information detection is not perfect; (2) the traditional descriptor cannot describe the feature points well, resulting in the sparse points of the same name obtained by the initial matching and even the matching failure; (3) the NRD differences between multi-modal images is significant. The existing distance matching method is more sensitive to MRSIs so its applicability is reduced or even unusable. To fill this knowledge gap, this paper introduces a MRSI matching method that uses CoF to construct the image scale space, and enhances gradient and feature description optimization to achieve effective MRSI matching.

III. Co-occurrence Filter Space Matching (CoFSM)

Image matching generally includes: image scale space construction, feature point extraction, descriptor construction, matching relationship construction, and mismatch elimination. In our study, the process of the proposed CoFSM method was organized into four sections as shown Fig.1: (1) image co-occurrence scale space construction and feature point extraction; (2) log-polar descriptor extraction suitable for multi-modal images; (3) multi-modal image matching optimized for position error; and (4) outlier removal of multimodal images.

A. Co-Occurrence Scale Space Construction and Feature Extraction

Feature point extraction in the CoF scale space is an important topic of this paper, which mainly includes three sections: (1) CoF image scale space construction, (2) feature point extraction, and (3) feature point optimization filtering.

1) Co-Occurrence Filter (CoF) Scale Space Construction: The purpose of the classic SIFT image scale space is mainly to convolve the image with Gaussian kernel functions of different sizes and Meanwhile down-sample between the hierarchical transformations to construct the scale space in the pyramid mode. The traditional image scale space is constructed with isotropic diffusion or anisotropic diffusion [28], [29], which smooth or blur the image to a certain extent. However, these methods are not conducive to the preservation of the image edge features and the removal of image noise. In particular, the NRDs of MRSIs are larger than those with the same sensor and the same modal, and their nonlinear luminance differences are obvious. Therefore, it is of great significance to effectively retain the image edge feature information and construct a new image scale space that can optimize the nonlinear distortion difference.

The co-occurrence matrix collects the point-like mutual information in the image to obtain the probability of the boundary in the image, which is used to measure the similarity between textures. The definition of CoF is shown in Equation (1):

$$J_p = \frac{\sum_{q \in N(p)} G_{\sigma_s}(p,q) \cdot M(I_p, I_q) \cdot I_q}{\sum_{q \in N(p)} G_{\sigma_s}(p,q) \cdot M(I_p, I_q)} \tag{1}$$

In Equation (1), J_p and I_q are the output and input pixel values, p and q are pixel indices; $G_{\sigma_s}(p,q) \cdot M(I_p, I_q)$ is the weight of the contribution of pixel q to the output of pixel p; $G_{\sigma_s}(p,q)$ is Gaussian filter; $M(I_p, I_q)$ is the calculation result of the co-occurrence matrix.

The weight of image co-occurrence filtering is mainly calculated by the co-occurrence matrix. M is a 256×256 matrix, the equation is shown in (2):

$$\begin{cases} M(a,b) = \dfrac{C(a,b)}{h(a)h(b)} \\ C(a,b) = \displaystyle\sum_{p,q} \exp(-\frac{d(p,q)^2}{2 \cdot \sigma^2})[I_p = a][I_q = b] \\ h(a) = \displaystyle\sum_{p} [I_p = a], \quad h(b) = \sum_{q} [I_q = b] \end{cases} \tag{2}$$

In Equation (2), $M(a,\ b)$ is based on the co-occurrence matrix $C(a,\ b)$ that counts the co-occurrence of values a and b divided by their frequencies(i.e., the histogram of pixel values), $h(a)$ and $h(b)$, in the image; σ^2 is a fixed parameter determined, $\sigma^2 = 2 \cdot \sqrt{5} + 1$.

The co-occurrence space of the current image layer is calculated by Equation 1 and Equation 2. To reduce the computational complexity, the image is not down sampled. Therefore, the resolution of the image in the co-occurrence scale space is the same. The scale space is divided into $(S + 1)$ layers (generally no more than 8 layers), then the scale definition of each layer image is shown in (3):

$$\sigma_n = \sigma_0 \cdot \sqrt[3]{2^n}, \quad (n = 0, 1, 2 \cdots S) \tag{3}$$

In Equation (3), σ_n represents the scale of the nth layer image in scale space; σ_0 represents the scale of the first layer image in scale space; S represents the number of scale space layers of the multimodal image. Considering the need to calculate the size of the statistical window when calculating co-occurrence information, the size of the co-occurrence window is obtained by combining the initial window size of the filter and the image scale. The definition is shown in (4):

$$\begin{cases} OC_n = \dfrac{\sigma_n^2 \cdot N_O}{2}, \quad (n = 0, 1, 2 \cdots S) \\ CoFSpace = \left\{ OC_n \cdot J_p^n \right\}_{n=0}^{N} \end{cases} \tag{4}$$

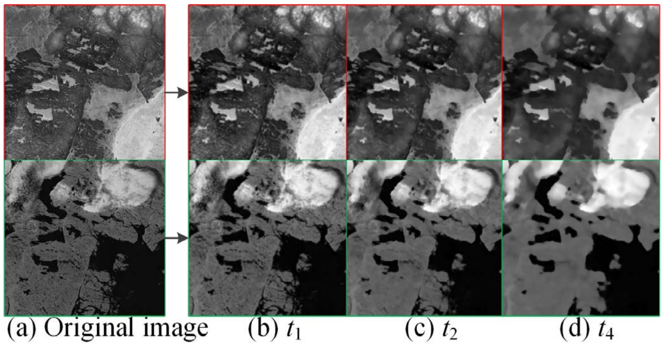


Fig. 2. Multi-modal remote sensing image of co-occurrence filtering (CoF). (a) is the original image; (b) is the first layer result of the CoF scale space; (c) is the second layer result of the CoF scale space; (d) is the third level result of the CoF scale space; (e) is the fourth level result of the CoF scale space.

In Equation (4), *COFSpace* represents the image collection of co-occurring scale space; OC_n represents the size of the nth co-occurrence statistical window; N_O represents the initial window size of the co-occurrence filter (this paper is set to 5, see Tables II); n represents the scale space layer; J_p^n represents the *nth* layer multimodal image after co-occurrence filtering. The results are shown in Fig. 2.

2) Shi-Tomasi Feature Point Extraction: After construction of the CoF scale space of the image, the corner detection operator can be used to extract the feature points. The feature points are extracted through the gradient image calculation of the MRSIs. Considering that the traditional image gradient is sensitive to NRDs, the Low-Pass Butterworth (LPB) filter is introduced as a filtering method to generate a new gradient to weaken the influence of NRD. The new gradient constructed by LPB filtering can increase the number of homonymy points for most MRSIs, which is beneficial to obtain more matching results.

The LPB filter was proposed by Kovesi *et al.* [30], [31], for image processing. Because of its maximum flatness in the pass band, it can reduce the energy of the high-frequency part of the image to achieve the effect of smoothing the image and reducing noise. Therefore, the LPB filter has a certain effect on the optimization of nonlinear distortion of the image. Its mathematical expression can be written as follows:

$$\mathbf{LPB} = \frac{1}{1.0 + (D(u, v)/cut_{off})^{2n}} \tag{5}$$

In Equation (5), **LPB** represents the convolution kernel of the Low-Pass Butterworth filter; $D(u,\ v)$ represents the distance from the pixel point to the reference points (u, v) [32]; (u, v) is a two-element vector specifying the size of filter to construct (In this paper, we set (u, v) to [3,3]); cut_{off} represents the cutoff frequency of the filter, with a value between 0 and 0.5; n represents the order of the filter.

The LPB filter is integrated into the gradient calculation of MRSI to obtain new first-order and second-order gradient amplitude diagrams, the definition is shown in (6):

$$\mathbf{G}_\sigma^1 = \sqrt{(\mathbf{L}_{(x,\sigma)} \otimes \mathbf{LPB})^2 + (\mathbf{L}_{(y,\sigma)} \otimes \mathbf{LPB})^2} \tag{6}$$

In Equation (6), $\mathbf{G}_\sigma^1$ represents the first-order gradient amplitude diagram of MRSI; $\otimes$ represents the convolution operator. σ represents the scale of a CoF scale space image; $\mathbf{L}_{(x,\sigma)}$ and $\mathbf{L}_{(y,\sigma)}$ represent the difference of the CoF scale space image L along the horizontal and vertical results of the scale σ, respectively.

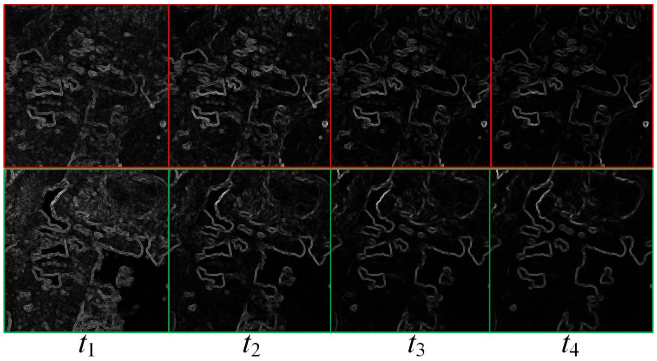


Fig. 3. Results of new gradient magnitude. The (t_1), (t_2), (t_3) and (t_4) respectively represent new gradient magnitude layers generated based on LPB filtering optimization.

Ma *et al.* [33] used a Sobel operator to eliminate the non-linear brightness difference of an image and achieved good results in remote sensing images. Therefore, to highlight the feature information, the LPB filter was combined with the Sobel operator to obtain second and third order image gradients, as shown in the following:

$$\begin{cases} \mathbf{G}_\sigma^2 = \sqrt{(\mathbf{G}_{x,\sigma}^1 \otimes \mathbf{\Gamma}_x)^2 + (\mathbf{G}_{y,\sigma}^1 \otimes \mathbf{\Gamma}_y)^2} \\ \mathbf{Angle}_\sigma^2 = \arctan(\dfrac{\mathbf{G}_{y,\sigma}^1 \otimes \mathbf{\Gamma}_y}{\mathbf{G}_{x,\sigma}^1 \otimes \mathbf{\Gamma}_x}) \end{cases} \tag{7}$$

$$\begin{cases} \mathbf{G}_\sigma^3 = \sqrt{(\mathbf{G}_{x,\sigma}^2 \otimes \mathbf{\Gamma}_x)^2 + (\mathbf{G}_{y,\sigma}^2 \otimes \mathbf{\Gamma}_y)^2} \\ \mathbf{Angle}_\sigma^3 = \arctan(\dfrac{\mathbf{G}_{y,\sigma}^2 \otimes \mathbf{\Gamma}_y}{\mathbf{G}_{x,\sigma}^2 \otimes \mathbf{\Gamma}_x}) \end{cases} \tag{8}$$

In Equation (7) and (8), $\mathbf{G}_\sigma^2$ represents the new second-order gradient amplitude; $\mathbf{Angle}_\sigma^2$ represents the new second-order gradient direction; $\mathbf{G}_\sigma^3$ represents the new third-order gradient amplitude; $\mathbf{Angle}_\sigma^3$ represents the new third-order gradient direction; $\mathbf{\Gamma}_x$represents the Sobel operator template in the X direction; $\mathbf{\Gamma}_y$ represents the Sobel operator template in the Y direction.

In our study, the LPB and Sobel filters were combined to generate a new image gradient, and more influential gradient information was extracted. To highlight the new gradient effect, the gradient amplitude and gradient direction were visualized, and the results are shown in Fig. 3 and Fig.4.

After the multimodal image calculation was completed, the Shi-Tomasi operator was used to extract the feature points. The Shi-Tomasi operator was proposed by Shi *et al.* [34]. It is an improvement of the Harris operator and has a strong anti-noise ability for MRSI images and can extract sufficient feature points. Therefore, the Shi-Tomasi operator was chosen for extracting the feature points in our study. Meanwhile, to avoid duplication of feature points, all feature points need to be filtered. In the paper, all feature points are sorted from the highest to the lowest by response value. Only points with duplicate coordinate positions (x-coordinates and

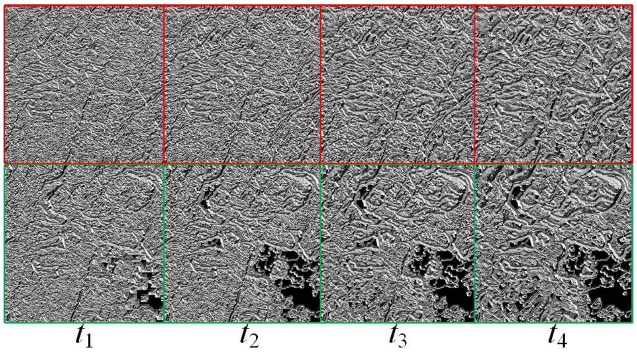
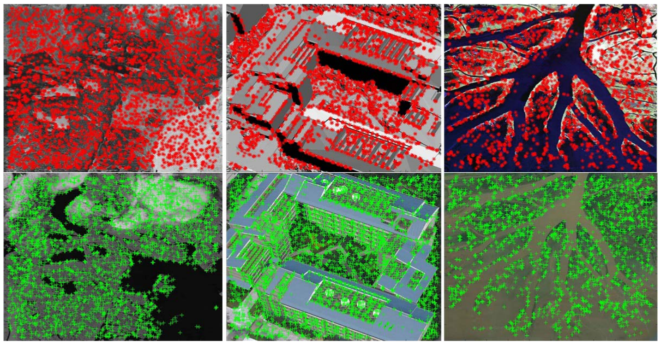


Fig. 4. Results of new gradient angle. The (t_1), (t_2), (t_3) and (t_4) respectively represent new gradient angle layers generated based on LPB filtering optimization.



(a) infrared-optical (b) depth-optical (c) SAR-optical

Fig. 5. The results of feature points detected and optimized based on Shi-Tomasi algorithm. (a) is the final extraction result of feature points of infrared-optical images; (b) is the final extraction result of the feature points of the depth-optical images; (c) is the final extraction result of feature points of SAR-optical images.

y-coordinates) with higher response values are retained. The filtered feature points are shown in Fig. 5. Fig. 5(a) through 5(c) are the feature points extracted from an infrared-optical-image-pair, a depth-optical-image-pair, and a SAR-optical-image-pair respectively.

B. Improved Log-Polar Coordinate Descriptor

After the feature points are extracted, describing them is an important step for subsequent successful matching. A great deal of research has been conducted on the generation of descriptors. The more commonly used descriptors in image feature matching include SIFT descriptors and log-polar descriptors. However, due to the difference in the RNDs of MRSIs, the SIFT descriptor cannot be accurately described, and an incorrect matching result usually is obtained. The log-polar description method using the gradient location and orientation histogram (GLOH) algorithm has obvious advantages and is relatively stable [35], [36]. However, the log-polar description method is not the only descriptor. It depends heavily on the division of the polar coordinate grid, and different division methods can generate different descriptors. Meanwhile, it is also sensitive to the gradient amplitude and gradient direction of the image. Of course, the advantage of this method is that the description method is flexible and therefore has better applicability than the SIFT descriptor.

The new gradient amplitude and the new gradient direction calculated in Section III.(A) were used to construct a log-polar descriptor. Therefore, our study chose the GLOH algorithm to generate the log-polar descriptors. However, the division of the sub-region grids in the neighborhood of the feature points is the key to constructing log-polar descriptors. The dimensions of the descriptor are determined by different grid division methods, and the descriptors of different dimensions have different stability levels in describing the feature points. The neighborhood grids constructed by the GLOH algorithm commonly are used in four and eight division grids. Although these divisions have achieved significant results in optical matching, the description of the MRSIs may not be robust enough, such as SAR images and optical images.

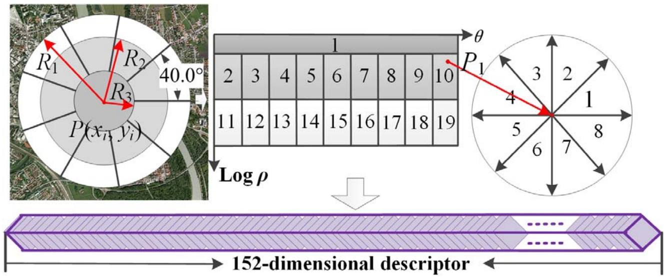


Fig. 6. Log-Polar descriptors of grid-optimized for Multi-modal image.

Therefore, considering the stability and robustness of the descriptor, we divided the neighborhood grid from the right end at zero degrees into a fan-shaped neighborhood every 40° and finally divided the entire circular neighborhood into nine equal parts. A new pair of log-polar coordinate grids of $(9 \times 2 + 1)$ sub-region grids was generated. This grid division method not only compensated for the instability of the descriptor caused by less division of the grid (such as four equal divisions and eight equal divisions), but also avoided the redundant calculation caused by dividing too many grids (e.g., 10 equal divisions).

The area of each sub-region was approximately the same. The horizontal direction of each grid represented the polar angle of the pixel location in the circular neighborhood, and the whole circular neighborhood was divided into eight fan-shaped intervals by 45°. Therefore, the pixels of each sub-region had a gradient amplitude and direction histogram of eight dimensions. Finally, the number of logarithmic polar coordinate sub-region grids (19) and the direction histogram (eight dimensions) were multiplied to generate a new 152-dimensional log-polar descriptor. The log-polar descriptor is shown in Fig. 6. Each purple square represents a vector value of one dimension, and there is a total of 152 vector values.

The improved Log-polar descriptor is quantified as: the mathematical equation of the vector component of the descriptor characteristic is as follows:

$$LPD = [D_1, D_2, \cdots, D_N]^T \tag{9}$$

In Equation (9), LPD represents the descriptor set of all feature points; D_i^T represents the descriptor vector of a feature point; T represents the matrix transposition character; N represents the number of feature points.

The components of the feature vector of the 152-dimensional descriptor can be expressed as $D_i^T = [V_1, V_2, V_3, \cdots, V_{152}]$.The dimension of each descriptor can be expressed as $(2^*n + 1)^*d$, which n

represents the number of grids divided by each circular neighborhood; d represents the direction dimension of each feature point.

C. Optimization of MRSI Position

After the descriptor is obtained, the initial match can be performed. However, it is often difficult for the existing distance ratio matching methods (e.g., Euclidean distance, hamming distance) to describe an accurate relationship between the MRSI feature points, which makes it difficult to match the correct pair of points with the homonymy points. In the case of correct matching, the horizontal and vertical displacements of the feature points usually have common characteristics. Therefore, if the offset of the feature point in the horizontal and vertical direction can be roughly obtained, and the offset is optimized to describe the distance function, more matching results can be obtained. Our study used a similar transformation to calculate the initial conversion model of the MRSI. The Euclidean distance extended would be named as the Position Euclidean Distance (PED). The processing consists of three steps: 1) image transformation parameters initialization; 2) position offset calculation; 3) descriptor matching distance optimization. The detailed process is as follows.

1) Image Transformation Parameters Initialization: In our study, Euclidean distance was used for initial matching; and the Random Sample Consensus (RANSAC) algorithm [37] then was used for fast outlier removal. The elimination threshold was set harsh to quickly obtain more accurate model transformation parameters. To convert the success rate of the model calculation, the threshold error of the least squares iterative calculation was set to six pixels, and the matching point pair was used for the least squares iterative calculation to obtain the model transformation parameters.

2) Position Offset Calculation: By using the above determined model transformation parameters and combining the position, the position offset error between the feature points was calculated as shown.

Suppose the set of feature points extracted from the reference MRSI is defined as P_{left}. The set of feature points extracted from the MRSI to be matched is defined as P_{right}. Combined with the transformation parameters, the position transformation error can be defined following [6]:

$$\Delta\mathrm{XY}(P_{left}^{i}, P_{right}^{j}) = \left\| P_{left}^{i} - T(P_{right}^{j}, \mu) \right\| \tag{10}$$

In Equation (10), $\Delta\mathrm{XY}(P_{left}^{i}, P_{right}^{j})$ represents the position transformation error; T represents the transformation model; μ represents the parameters of the transformation model.

3) Descriptor Matching Distance Optimization: The results of Equation (10) are used as a constraint to optimize the Euclidean distance. These operations finally acquire a new descriptive distance for MRSI matching. The mathematical expression is shown in Equation 11.

$$\begin{aligned} PED(P_{left}^{i}, P_{right}^{j}) &= ED(P_{left}^{i}, P_{right}^{j}) \\ &\quad \cdot (1 + \Delta\mathrm{XY}(P_{left}^{i}, P_{right}^{j})) \end{aligned} \tag{11}$$

In Equation (11), $PED(P_{left}^{i}, P_{right}^{j})$ represents the Euclidean distance after the error of position offset transformation is optimized; $ED(P_{left}^{i}, P_{right}^{j})$ represents the Euclidean distance between the descriptors of the feature points P_{left}^{i} and P_{right}^{j}. When the value is minimum, the feature points P_{left}^{i} and P_{right}^{j} are considered as a correct corresponding. Then, the point pair with the shortest PED is used as a candidate matching pair. The above steps are calculated iteratively, and more matching point pairs are retained.

D. MRSI Matching and Outlier Removal

The optimized descriptor distance was rematched to obtain a new matching result. However, there were still a lot number of outliers that did do not correspond to each other, and these outliers still needed to be removed. The outlier removal threshold is set to 3 pixels. The use of logical filtering has been shown to remove position offset errors [38]. Therefore, we used this method for reference to deal with logical filtering. Then the fast sample consensus (FSC) method [39] was used to eliminate the small outliers.

IV. Experimental Results and Analysis

A. Datasets and Parameter Settings

To verify the effectiveness of our CoFSM method, sixty pairs of multi-modal images were selected for qualitative and quantitative evaluation. To further verify the performance of the CoFSM algorithm, the proposed CoFSM algorithm was compared its performance to four of the latest matching algorithms such as SIFT, upright-SIFT(i.e. for SIFT with no dominant orientation estimation for generating patches), PSO-SIFT and RIFT. For fair comparison, all the implementation details of SIFT, PSO-SIFT and RIFT were obtained from their authors' websites.

First, feature matching was performed on the six image-pairs and the correspondences with less-than-three-pixel residuals were regarded as the correct matches [14]. The number of correct matches (NCM), the ratio of corrected number (NCR) root mean square error (RMSE), and success rate (SR) were used as the evaluation metrics. The SR was calculated with the following definition of success matching: (1) the NCM should be sufficient to obtain a solution of the selected geometric transformation model and have at least one redundant observation; and (2) the NCR should be larger than 20% because the literature shows that the latest robust solvers can deal with over 80% of the outliers[40]. The definition of SR is as follows:

$$I(p_i) = \begin{cases} 1, & NCM(p_i) \geq N_{\min} \& \dfrac{NGT(p_i)\big|_{\geq\tau}}{NCM(p_i)} \geq \xi \\ 0, & else \end{cases} \tag{12}$$

$$SR = \frac{\sum_{i} I(p_i)}{M} \times 100\% \tag{13}$$

In Equation (12), $I(p_i)$ represents a logical value, 1 represents a success matching trial and 0 represents a failed matching trial. $NCM(p_i)$ represent the number of matched points of the i-th image pair. $N_{\min}$ represents the minimum number of matches sufficient to obtain a solution of the

TABLE I
PARAMETER SETTING OF OUR COFSM

Experiments	Variable	Fixed Parameters
Parameter N_O	N_O= [3,4,5,6,7]	N_L = 4, N_T= 1300
Parameter N_L	N_L = [1,2,4,6,8]	N_O = 5, N_T = 1300
Parameter N_T	N_T=[600,1000,1300,1600,2000]	N_O = 5, N_L = 4

selected geometric transformation model and has at least one redundant observation. In this paper, we used an affine model for transformation, which needs at least three matches a solution; thus, the $N_{\min}$ was set as 4. $NGT(p_i)$ represents the number of correct matches, which were judged by the ground truth affine model with an τ-pixel distance threshold. The ground truth model was obtained by calculation of the corresponding points collected manually. In this paper, the ground truth affine model was calculated with the manually collected corresponding points, and the threshold τ was set as 3. ξ represents the correct-match-ratio threshold, which was set at 20% in this paper. In Equation (13), *SR* represents the matching success rate; M represents the total number of image pairs of a multimodal image sets.

1) MRSI Datasets: Six types of multi-modal remote sensing image data sets were selected as the experimental sets. The image size ranges from 400 × 400 pixels to 650 × 650 pixels. These multi-modal remote sensing image datasets include multi-temporal-optical images, infrared-optical images, depth-optical images, map-optical images, SAR-optical images, and night-day images, which considers almost all the application scenarios of multi-modal image matching like multi-source data interpretation, multi-structure data registration, and multi-spectral data fusion. Each dataset contains 10 image pairs for a total of 60 MSRI pairs. Most image pairs have significant nonlinear radiation distortions which is very representative and can fully verify and compare the performances of the MRSI matching algorithms. The first image pair of each of the six MRSI datasets is shown in Fig. 7.

2) Parameter Settings: Our study can be mainly divided into four parts, i.e., co-occurrence filtering scale space construction, feature point extraction, log-polar descriptor construction, and position Euclidean distance matching. Table I shows the three parameters which are needed to be analyzed, while other parameters are defined with author’s published papers.

Parameter N_O represents the initial co-occurrence filter window size. In general, the larger the initial CoF window value is set, the larger the co-occurrence matrix range and the less feature point extraction; in the opposite case, the more feature points will be extracted. Parameter N_L represents the number of scale space layer settings. In general, the higher the number of layers is set, the more the feature points will be extracted and the greater the time-cost will be needed. Parameter N_T represents the threshold of feature point extraction. In general, the smaller the threshold is set, the more feature points will be extracted, but more noise will be extracted. This section describes the parameter study and sensitivity analysis conducted based on the night-day sensing images dataset. Three independent experiments were designed to determine parameters N_O, N_L and N_T, where each experiment had only one parameter as a variable and the other parameters were fixed. For each parameter, NCM and SR were used as the evaluation metrics. The experimental results are shown in Tables II. In Tables II, the first row represents three parameter variables, one of which represents a variable, and the other two remain unchanged. The second row is the change value of the variable.

TABLE II
THE RESULTS OF PARAMETER N_O, N_L *and* N_T

	N_O,N_L,N_T	3,1,600	4,2,1000	5,4,1300	6,6,1600	7,8,2000
N_O	NCM	282.7	271.6	**351.6**	336.9	296.9
	SR/%	90	90	**100**	90	80
N_L	NCM	129.8	166.9	351.6	**368.5**	308.9
	SR/%	60	80	**100**	80	70
N_T	NCM	291.1	374.4	351.6	**414.4**	366.7
	SR/%	**90**	**100**	**100**	**100**	**100**

From the experimental results, the following conclusions can be inferred:

(1) in our study, the parameter N_O was set to 5. Not only average NCM was very high, but also when the N_O value was set to 5, the SR was 100%, and the SRs with other’ N_O (3, 4, 6, and 7) did not reach 100%. Therefore, considering the results of NCM and SR, and the time efficiency, it was better to set the N_O value to 5.

(2) Parameter N_L of the image generally does not exceed eight; but in the actual matching task, it needs to be chosen according to the characteristics of the image. As the results in Table II, when the value of N_L was smaller, the results of NCM and SR were poor. When the value of N_L was larger, the result of NCM was better, but the result of SR was worse. When the value of N_L was equal to 4, NCM and SR obtained a better result. At the same time, the larger the value of N_L was set, the more time that was necessary. Therefore, considering the results of NCM and SR, and the time efficiency, it was better to set N_L to four layers.

(3) Parameter N_T of the image generally is important in feature point extraction. As the results show in Table II, when the value of N_T was set smaller, more noise was extracted, affecting the results of NCM and SR. When $N_T = 600$, the SR of the image was 90%, and it did not reach 100%. However, when the value of $N_T \geq 1000$, both NCM and SR obtained a better result. Setting the value of N_T between 1000 and 2000 can achieve a better matching effect. Considering the calculation efficiency and time, when the value of $N_T = 1300$, the result was the best. Therefore, considering the results of NCM and SR and the time efficiency achieved, the threshold was set at N_T to 1300.

The parameters for the comparison methods of SIFT, upright-SIFT, PSO-SIFT and RIFT were set according to their original authors. For all the matching methods, the outlier elimination threshold was set to 3 pixels. To improve the efficiency of matching, a multi-thread parallel processing strategy was used for matching both the descriptor calculation and the matching stage.

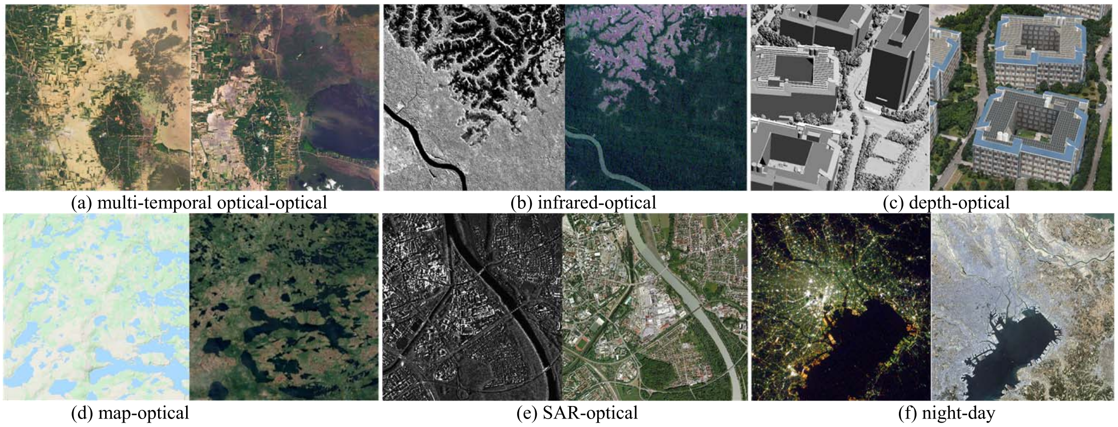


Fig. 7. Part of multi-modal images. The six groups of image pairs are composed of the first group of image pairs of six types of multi-modal images.

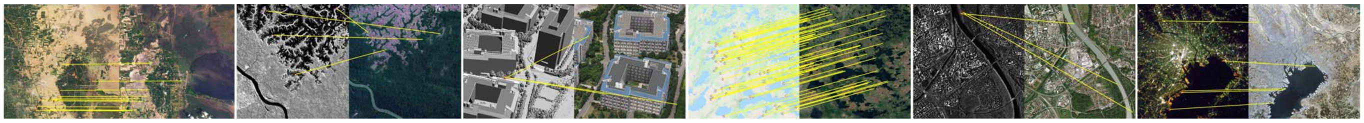


Fig. 8. Matching results of the images in Fig. 7 with SIFT.

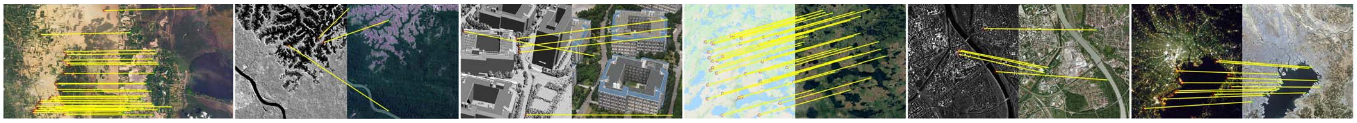


Fig. 9. Matching results of the images in Fig. 7 with upright-SIFT.

3) Experimental Conditions: All the algorithms were implemented under MatlabR2018a; and the experiments were performed on a Window10 X64 laptop computer with Intel(R) Core(TM) i7-9750H CPU at 2.59 GHz, and 16 GB RAM.

B. Qualitative Evaluation of Matching Results

The matching results of SIFT, upright-SIFT, PSO-SIFT, RIFT and CoFSM are shown in Fig.(8) through Fig.(12), respectively.

As is shown in Fig. (8) through Fig. (12), the results of CoFSM was significantly better than those of SIFT, upright-SIFT, and PSO-SIFT algorithms.

SIFT failed to match the image pairs in Fig. 8(b), 8(c), and 8(e) while successfully matching in Fig.8(a), 8(d), and 8(f). SIFT's SR accuracy therefore was 50%; however, even if its matching had been successful in the other figures, its NCMs were also small (i.e., 12, 62, and 5, respectively). SIFT uses the Gaussian image pyramid to construct the image scale space, which performs global blur processing on the image, resulting in the image texture edge feature being weakened and making it difficult to extract the feature information of the image contour edge. The above treatment led to poor matching results due to the huge differences in the computed gradient histogram. In summary, the traditional image scale space was not good for MRSI matching.

Fig.9 shows that the upright-SIFT algorithm can obtain better matching results than the SIFT algorithm. The upright-SIFT failed to match the image pairs in Fig. 9(b), 9(c), and 9(e) while successfully match the image pairs in Fig.9(a), 9(d), and 9(f). However, the NCMs were also low (i.e., 31, 36, and 11, respectively), which shows that the upright-SIFT still cannot obtain robust results in MRSI matching. The reason may be that the SIFT-descriptor is very sensitive to the NRD of the MRSIs.

PSO-SIFT failed to match on the image pairs in Fig. 10(c), and successfully matched the image pairs of Fig. 10(a), 10(b), 10(d), 10(e), and10(f). Its SR accuracy was 83.33%; however, except for the larger value of NCM in Fig. 10(b), the other results were still low (i.e., 43, 201, 18, 38 and 10) because PSO-SIFT mainly utilizes the Sobel algorithm to redefine the image gradient to optimize the SIFT algorithm. Therefore, it was concluded from Fig.10 that when the modal difference of an image is small, the matching can be successful; and conversely, when the modal difference is large, the matching effect can be poor or may even fail.

Fig. 11 shows that RIFT successfully matched all six image pairs, and its matching SR accuracy was 100%. However, the NCMs in Fig. 11(c) and 11(f) were still low, 197 and 142, respectively. Because the RIFT algorithm matches by the phase consistency information, the image therefore was converted from the spatial domain to the frequency domain. Although it had good adaptability to multi-modal image matching, it also caused the matching accuracy to be less robust than required.

As shown in Fig. 12, CoFSM successfully matched all six image pairs, and its SR accuracy was 100%. At the same time, the NCM of all its results (i.e., 677, 835, 322, 674, 1014, and 370) better than RIFT and much higher than SIFT and PSO-SIFT.

C. Quantitative Evaluation of Matching Results

Fig.13 shows the quantitative results of the NCM metric for the four comparison methods on the six MRSI datasets.

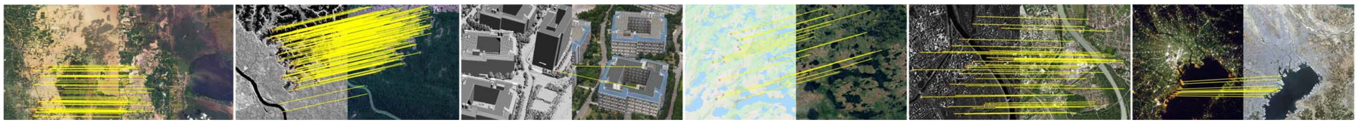


Fig. 10. Matching results of the images in Fig. 7 with PSO-SIFT.

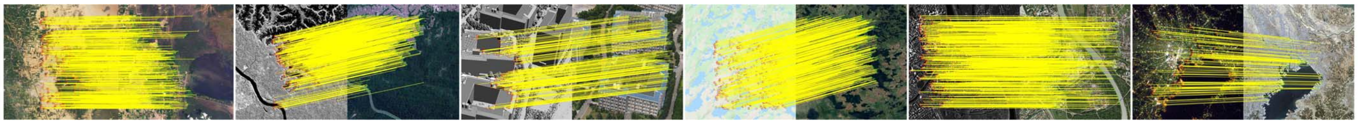


Fig. 11. Matching results of the images in Fig. 7 with RIFT.

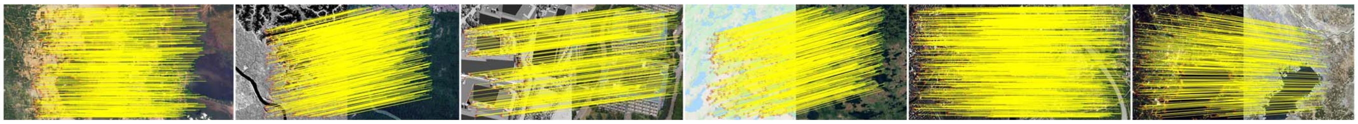
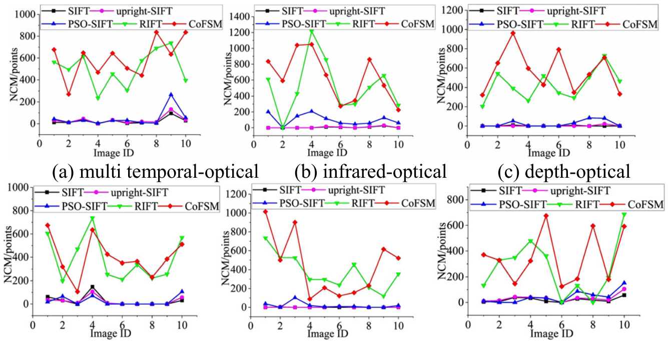


Fig. 12. Matching results of the images in Fig. 7 with CoFSM.



(a) multi temporal-optical (b) infrared-optical (c) depth-optical

(d) map-optical (e) SAR-optical (f) night-day

Fig. 13. Comparisions on NCM metric.

TABLE III
COMPARISIONS ON SR METRIC

Method	SR/%					
	1[1]	2	3	4	5	6
SIFT	90	40	20	50	10	90
upright-SIFT	90	40	10	60	20	90
PSO-SIFT	100	90	40	40	70	70
RIFT	100	90	100	100	100	80
CoFSM	**100**	**100**	**100**	**100**	**100**	**100**

[1] MULTI-MODAL TYPE 1~6 OF CoFSM .1. MULTI TEMPORAL-OPTICAL;2. INFRARED-OPTICAL;3. DEPTH-OPTICAL;4. MAP-OPTICAL;5. SAR-OPTICAL;6. NIGHT-DAY

As can be seen, SIFT performed better on the multi-temporal-optical dataset and the day-night dataset than on the other four datasets because of its resistance to illumination changes. In the multi-temporal-optical dataset, the difference in modal between the images was smaller than that of the other four data sets, and matching became relatively easy. Also note that the night-day dataset essentially also was a multi-temporal-optical dataset, but, their light differences made the matching more complex. SIFT produced its worst performance on the SAR-optical dataset and the depth-optical dataset, and their SR accuracy was 10% and 20%, respectively; therefore, few correct matches were obtained. SIFT may have performed poorly for the following reasons. (1) The gradient constructed by SIFT was more sensitive to MRSI and the feature information description was insufficient. For example, the depth-optical dataset and the SAR-optical dataset have huge modal differences. (2) SIFT detects feature points directly based on intensity; the number of extracted feature points was few and the distribution was poor. In most of the successfully matched image pairs, most of the NCMs of SIFT were very small (smaller than 50 points). In some images, there were only a few correct matching points.

Fig.13 shows that the upright-SIFT algorithm improves the matching performance of the SIFT algorithm by about 15%∼60%. However, it is still difficult to obtain robust results in the depth-optical and SAR-optical types, where the matching SRs are only about 20% and 10%, and the average NCMs are only 1.8 and 0.4. Only in the multi-temporal optical and the night-day datasets, the upright-SIFT achieved relatively better results, where the matching SRs are both 90% and the average NCMs are about 33.6 and 33.3 respectively. In conclusion, even using a fixed orientation angle, the upright-SIFT cannot achieve robust matching with all the MRSI types.The performance of PSO-SIFT was better than SIFT. Its performance on the multi-temporal-optical dataset and the infrared-optical dataset were better than for the other four datasets, but the NCMs of PSO-SIFT were still very few. The performance of RIFT was significantly better than PSO-SIFT and SIFT, but the NCMs of the matching depth-optical dataset and the night-day dataset were not stable enough. In contrast, CoFSM successfully matched all the image pairs of the six datasets, and the NCMs were much greater than 100 on most of the image pairs. The matching performance of CoFSM was very stable and robust because the CoFSM algorithm better weakened the NRDs and geometric differences, enhanced the edge information of the texture features of the MRSIs, and reduce the influence of the differences between the image modalities to a certain extent.

Table III summarizes the matching SRs of the four comparison methods on each dataset. SIFT had the highest SRs on the multi-temporal-optical dataset and night-day dataset, both of which were 90%. The matching performance of the upright-SIFT method is better than that of the SIFT algorithm in both map-optical and SAR-optical modality types. While for the other four modality types, no method is superior to

TABLE IV
COMPARISIONS ON NCM AND NCR OF RIFT AND COFSM

	Metric	1	2	3	4	5	6
RIFT	NCM	506.9	515.2	425	386.2	375.5	266.3
	NCR	16.84	17.44	12.25	11.27	14.43	11.07
CoFSM	NCM	**595.9**	**641.3**	**566.9**	**400**	**435.4**	**351.6**
	NCR	**27.49**	**31.48**	**24.57**	**19.05**	**21.03**	**16.62**

the others. PSO-SIFT had the highest SRs on the multi-temporal-optical dataset and the infrared-optical dataset, which were 100% and 90%, respectively. The SRs of RIFT were 100%, except for the infrared-optical dataset and the night-day dataset, which were 90% and 80%, respectively. In contrast, the SRs of CoFSM were 100% on all the datasets. The average SRs of SIFT, upright-SIFT, PSO-SIFT, RIFT, and CoFSM on the six datasets were 50%, 51.67%, 68.33%, 95%, and 100%, respectively. Compared with SIFT, upright-SIFT, PSO-SIFT, and RIFT, CoFSM improved by 50, 48.33, 31.67, and 5 percentage points, respectively. Fig. 13 plots the average NCM of in four algorithms on each image pair.

In Table IV compares the average NCR and NCM of the CoFSM and RIFT. The results show that the CoFSM had better average NCRs and NCMs than RIFT in all types of MRSI. Table IV shows the NCMs of RIFT and CoFSM on each dataset. As RIFT method was used, most image pairs had the NCM greater than 200. The average NCM of all image pairs was 412.52, with an average NCR of 13.88%. In contrast, CoFSM method brings the NCM of each image pair to over 300, with the average NCM of 498.52, the average NCR of 23.37%. Sufficient corresponding points would be obtained for MRSIs. The NCMs of CoFSM were more stable and robust than RIFT. As is shown, the average NCRs are all less than 50% whether using CoFSM or RIFT, which weakens the registration accuracy of MRSIs (see section IV (D)). This is mainly because: (1) The imaging mechanisms have significant differences, which causes non-linear radiation distortion; (2) some types of MRSI have relatively low signal-to-noise ratio; (3) change happens between the acquisition time of the multi-temporal images. In conclusion, obtaining more corresponding points is important in MRSI matching and registration.

Overall, the proposed CoFSM has been demonstrated good matching ability in MSRI as it achieved good NCMs on all six datasets. The CoFSM was shown to be suitable for MRSI matching, and its performance was better than any of the current three feature matching methods to which we compared it.

However, this question remains unanswered: was the uniformity of the distribution of feature points also excellent? To evaluate the uniformity of the distribution of feature points, the standard deviation of box-counting is used to evaluate the uniformity of the distribution of corresponding points. First, a square box with a radius of 20 pixels (that is, a diameter of 40 pixels) is established for each corresponding point, then all the corresponding points in the box are counted, and finally their standard deviations are counted. Considering comprehensively, calculating the standard deviation of box-counting can better represent the uniform distribution of the corresponding points. The following are the standard deviation results in the box of RIFT and the proposed CoFSM, as shown in Fig.14.

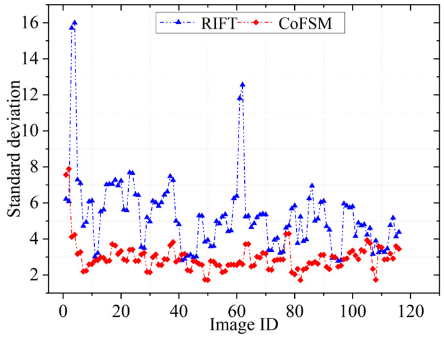
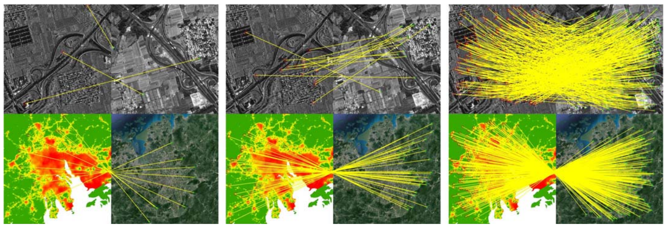


Fig. 14. The distribution of the standard deviation of box counting results of the two methods.



(a) Matching results of the SIFT. (b) Matching results of the PSO-SIFT. (c) Matching results of our CoFSM.

Fig. 15. Matching results of partial MRSI datasets.

The standard deviation of the proposed CoFSM is significantly lower than the standard deviation of RIFT. The average standard deviation of CoFSM is 2.956, and the average standard deviation of RIFT is 5.294. In summary, the corresponding points matched by CoFSM have better distribution uniformity than RIFT.

To evaluate the rotation-invariance, the rotated MRSIs were used to test the SIFT, PSO-SIFT, and CoFSM. The upright-SIFT removes the principal orientation of the SIFT method, which loses rotation-invariance. Since the open-sourced RIFT codes from the authors' website do not support rotation-invariant matching, upright-SIFT and RIFT were not involved in this evaluation. The matching results of six MRSI pairs are illustrated in Fig.15, which show that CoFSM had better results than SIFT or PSO-SIFT. However, considering space issues, only part of the matching results is displayed. The SIFT algorithm can successfully match some images in the multi temporal-optical, map-optical, and night-day data sets. The PSO-SIFT algorithm can obtain better matching results in the infrared-optical and night-day multi-modal data sets, but in other datasets, the results of matching are not robust. The proposed CoFSM is successfully matched in six types of MRSIs with large rotation, and enough corresponding points can be obtained.

D. Accuracy Assessment

In this section, the geometric accuracy of the RIFT matches and the CoFSM matches are compared. In each of the MRSI

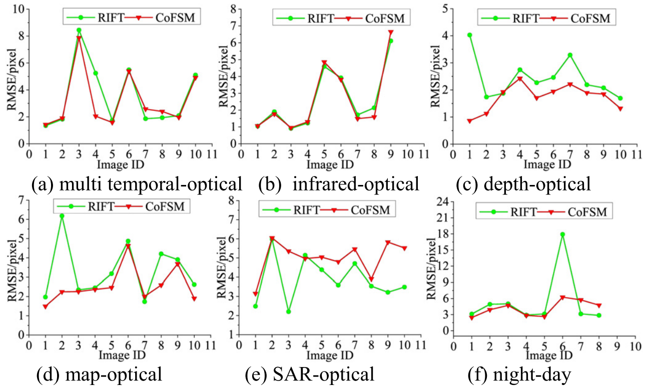


Fig. 16. Comparisions on RMSE metric.

TABLE V

AVERAGE RMSE RESULTS OF RIFT ALGORITHM AND COFSM ALGORITHM IN SIX TYPES OF MRSIs

Method	Type	1	2	3	4	5	6
RIFT	RMSE	3.51	2.62	2.44	3.35	**3.87**	5.38
CoFSM	RMSE	**3.21**	**2.61**	**1.73**	**2.57**	5.01	**4.17**

image pairs, no-less-than 20 true matches were manually collected and carefully checked and then, the RMSE of the transformation residual error of the true matches was computed with the homography model solved by the RIFT matches or CoFSM matches. The less the RMSE of true-matches, the better the matching accuracy of corresponding points.

In Fig.16, the true-match-RMSE of each MRSI pair is illustrated. In multi-temporal-optical image pairs, the results of CoFSM are slightly better than that of RIFT. In the infrared-optical image pairs, the results of the two methods are very close. In depth-optical, map-optical, and night-day image pairs, the results of CoFSM are obviously better than that of RIFT.

The average RMSE of each type of MRSI is in Table V. In multi temporal-optical image pairs, the average RMSE of CoFSM 8% lower than that of RIFT. In the infrared-optical image pairs, the results of the two methods are very close. In depth-optical, map-optical, and night-day image pairs, the average RMSE of CoFSM all are at least 22% lower than that of RIFT. But, the average RMSE of RIFT is 22% lower than that of CoFSM. In summary, the RMSE of the CoFSM algorithm is between 1 and 6 pixels. The average RMSE of the proposed CoFSM algorithm is 3.19 pixels, and the average RMSE of RIFT is 3.48 pixels. The RMSE of CoFSM is 8.3% lower than that of RIFT.

The possible reasons are: (1) the difference in imaging mechanism between MRSIs causes their matching accuracy to be lower than that of traditional matching. As a result, the upper limit of its accuracy is affected; (2) there are also differences between different MRSIs. To improve the robustness and universality of the algorithm, the advantages of some algorithms will inevitably be sacrificed, which lead to differences in the matching robustness between images of different modalities. In summary, MRSI matching is a challenging work. The proposed CoFSM can obtain robust matching performance in most image pairs.

TABLE VI

THE INFLUENCE OF LPB FILTERING ON THE MATCHING RESULTS OF SIX DATA SETS

Method	1	2	3	4	5	6
CoFSM(No-LPBF)	448.4	528.5	443.8	268.4	239.2	257.5
CoFSM(LPBF)	595.9	641.3	566.9	400	435.4	351.6

V. DISCUSSION

The CoFSM method achieved better results for two main reasons. (1) CoFSM constructs a multi-modal image scale space using the CoF, which allows the scale space to perform noise reduction and blur processing on the image at different scales and makes it convenient to extract the image feature information at different levels for better obtaining the texture edge similarity feature of MRSI. (2) A log-polar descriptor for optimizing grid division was established in CoFSM, which obtains the feature vectors by considering more feature directions and thus, makes the description of the feature points more robust.

To have a better understanding of CoFSM's matching performance the method is analyzed in terms of key-parameter influence, registration accuracy, and rotation invariance.

A. Analysis of Key Parameters

To fully demonstrate the influence of the presence or absence of LPB filtering on the matching results of the six multi-modal data types, this paper separately counts the average NCMs of the six multi-modal data of the two methods, and the results are shown in the following Table VI, where six data types are represented by Arabic numerals 1∼6.

Table VI shows that the matching results obtained with the LPB filter are significantly better than those without the LPB filter, where the NCMs increased about 20∼80% after the LPB filter involved. When the LPB filter is not used, the NCMs of the map-optical (#4), SAR-optical (#5) and night-day (#6) datasets (which are usually treated as more challenging MRSI matching tasks) were all less than 300, meanwhile, the matching SR of the SAR-optical dataset is only 90%. In summary, the new gradient features generated by LPB filtering is more suitable for MRSI matching, which not only improves the matching SR, but also obtains larger NCMs.

Since grid division will affect the results of multi-modal matching, this paper analyzes five grid division methods (6-sector, 8-sector, 9-sector, 10-sector and 11-sector). The details are shown Table VII.

Table VII shows the 9-sector grid generated has a higher matching SR and NCM than others grid. Their overall SRs are 93.33%, 95%, 100%, 95% and 96.67%, of which 6-sector has a lower SR in the night-day multi-modal data sets. Among them, in the night-day dataset, the SR results of the other four sector grids are all lower than the 9-sector grid. As the same time, comparing the results of NCM, it is found that the matching corresponding points obtained by the 9-sector method are significantly more than that of the other four sector grids methods. The average NCMs of the 9-sector method is 498.52, and the average NCMs for the 6-sector,

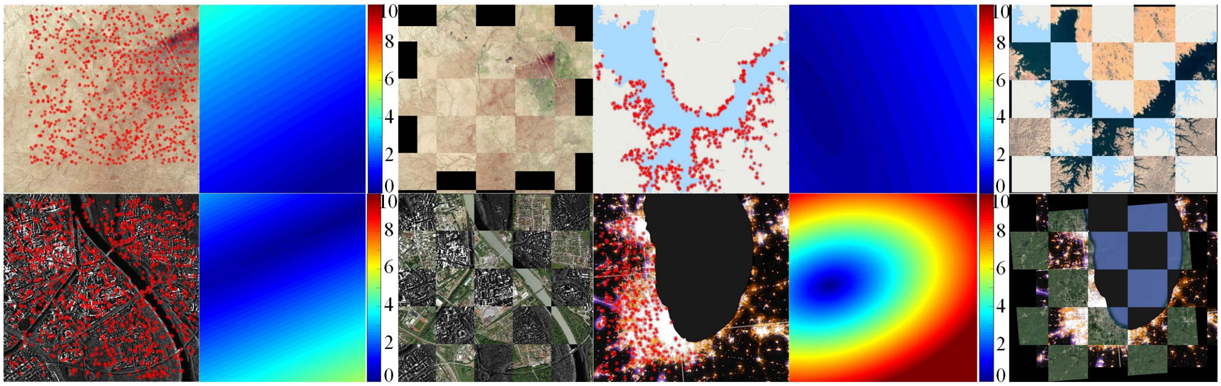


Fig. 17. Evaluation of image registration results. The first column is the distribution of feature points. The second column is the error distribution result of each pixel. The third column is the checkerboard registration result of the image.

TABLE VII
COMPARISON OF THE RESULTS OF FIVE SECTOR GRID METHODS

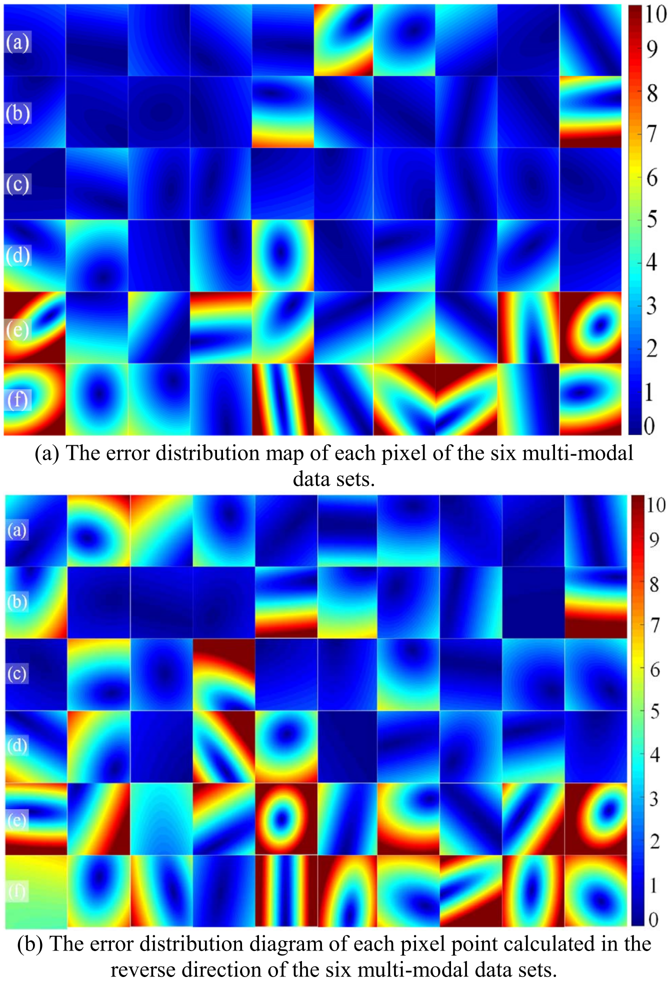
	Type	1	2	3	4	5	6
6	NCM	539.5	575.9	379.3	378.1	387.6	257.6
	SR/%	100	100	100	100	90	70
8	NCM	556.9	489.6	418.2	378.3	351.4	287.0
	SR/%	100	100	100	100	100	70
9	NCM	595.9	641.3	567.0	400.0	435.4	351.6
	SR/%	100	100	100	100	100	100
10	NCM	542.3	625.8	435.4	405.0	335.9	244.1
	SR/%	100	100	100	100	100	70
11	NCM	639.3	624.0	408.3	385.1	304.0	240.2
	SR/%	100	100	100	100	100	80

8-sector, 10-sector, and 11-sector methods are 419.67, 413.57, 431.42, and 433.48, respectively. Compared with the other four sector division methods (6-, 8-, 10- and 11-sector), the average NCM results obtained by 9-sector have increased by 15.82%, 17.04%, 13.46% and 13.05%, respectively.

B. Analysis of Registration Accuracy

To better evaluate the registration accuracy between images, the image registration is completed according to the affine matrix of the CoFSM model. At the same time, the error distribution map of each pixel of the image is calculated according to the homography of the ground truth and the homography of CoFSM. The purpose is to reflect the overall registration accuracy of MRSI and display them in a checkerboard form. The registration result of the MRSI is shown in Fig.17. At the same time, the distribution of error distances for all pairs is calculated and displayed, as shown in Fig.18.

As shown in Fig.17, the color on the map gradually changes from blue to red, which means that the error is gradually increasing. The error distribution situation corresponds to the actual registration situation, and the error in the overlapping area of the image is relatively low (mainly the blue area). The MRSI matching effect is better, and the checkerboard edges can be matched well without obvious dislocation. The homography matrix of the proposed CoFSM algorithm can better complete the registration of MRSIs.



(a) The error distribution map of each pixel of the six multi-modal data sets.

(b) The error distribution diagram of each pixel point calculated in the reverse direction of the six multi-modal data sets.

Fig. 18. The error distribution diagram of the six multi-modal data sets.

C. Robustness Analysis of the CoFSM Algorithm

To further examine whether the division of the feature neighborhood into nine sectors has a positive effect on the MRSI matching, an orientation shift on the grid by $0.5 \times 2\pi/9$ was added every time before the examination. A part of matching results was shown in Fig.19. The test results were shown in Fig.20. The proposed CoFSM algorithm can obtain a robust matching effect in the [0∼180] angle rotation transformation. In the case of different rotation angles, the matching performance of the CoFSM algorithm would have

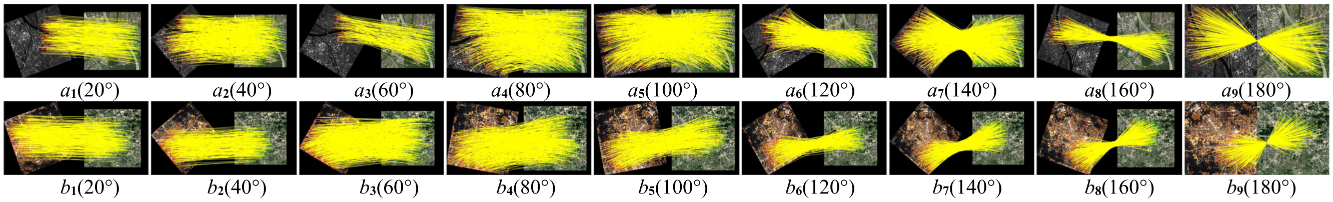


Fig. 19. Results of matching images with different rotation angles.

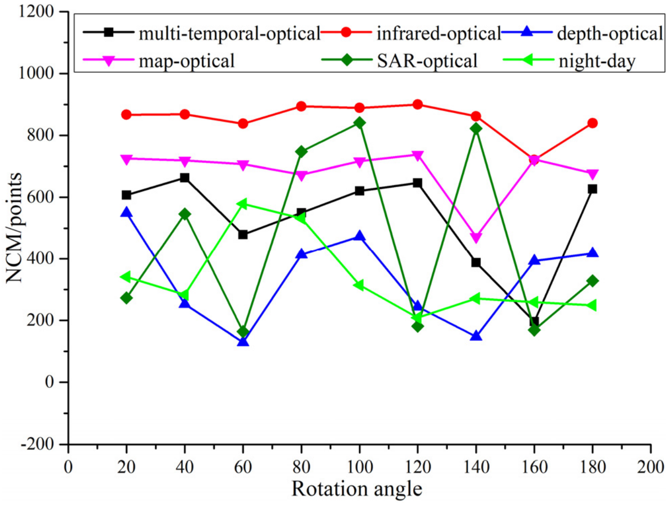


Fig. 20. NCM results of matching images with different rotation angles.

some differences. The CoFSM algorithm can obtain enough NCMs at different angles, and the proposed CoFSM in this paper is rotation invariant.

VI. Conclusion

In this paper, a new matching method of CoF and feature displacement optimization, called CoFSM, was proposed, and the NRD difference problem of MRSI was transformed into the optimization of image feature similarity information. By constructing a new gradient, the feature information of the image can be optimized to the greatest extent. The similar information of MRSI can be found, which reduces the impact of NRD differences, and more feature points can be extracted. Furthermore, more corresponding points were successfully matched by optimizing the feature offset. From the results of experiments on MRSI datasets, which contained sixty pairs of images, we made the following conclusions. (1) Better matching results were obtained in different MRSI scenes, such as multi-temporal optical and optical, infrared and optical, depth and optical, map and optical, SAR and optical, and night and day, which proved the CoFSM method. (2) In MRSI matching, the results with CoFSM were obviously better than that of SIFT, upright-SIFT, PSO-SIFT, and RIFT. (3) Our CoFSM method was designed on the basis of feature matching, which can make full use of the advantages of existing feature matching methods and has stronger scalability. (4) The results of the matched points by our CoFSM method indicated reliable distribution uniformity, thereby providing a good basis for subsequent processing. The main deficiency in our CoFSM method is the computational overhead, reducing which will be our research focus.

References

[1] J. Senthilnath, S. N. Omkar, V. Mani, and T. Karthikeyan, "Multiobjective discrete particle swarm optimization for multisensor image alignment," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 5, pp. 1095–1099, Sep. 2013.

[2] Y. Ye, J. Shan, S. Hao, L. Bruzzone, and Y. Qin, "A local phase based invariant feature for remote sensing image matching," *ISPRS J. Photogramm. Remote Sens.*, vol. 142, pp. 205–221, Aug. 2018.

[3] R. J. Jevnisek and S. Avidan, "Co-occurrence filter," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3184–3192.

[4] M. Gong, S. Zhao, L. Jiao, D. Tian, and S. Wang, "A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 4328–4338, Jul. 2014.

[5] E. Karami, S. Prasad, and M. Shehata, "Image matching using SIFT, SURF, BRIEF and ORB: Performance comparison for distorted images," 2017, *arXiv:1710.02726*.

[6] Q. Li, G. Wang, J. Liu, and S. Chen, "Robust scale-invariant feature matching for remote sensing image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 287–291, Apr. 2009.

[7] S. Lee, J. Lim, and I. H. Suh, "Progressive feature matching: Incremental graph construction and optimization," *IEEE Trans. Image Process.*, vol. 29, pp. 6992–7005, 2020.

[8] F. Bellavia and C. Colombo, "Rethinking the sGLOH descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 931–944, Apr. 2018.

[9] Y. Jin *et al.*, "Image matching across wide baselines: From paper to practice," *Int. J. Comput. Vis.*, vol. 129, pp. 517–547, Oct. 2021.

[10] C. Zhao, Z. Cao, J. Yang, K. Xian, and X. Li, "Image feature correspondence selection: A comparative study and a new contribution," *IEEE Trans. Image Process.*, vol. 29, pp. 3506–3519, 2020.

[11] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: A SIFT-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 453–466, Jan. 2015.

[12] Y. Ye, L. Shen, M. Hao, J. Wang, and Z. Xu, "Robust optical-to-SAR image matching based on shape properties," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 564–568, Apr. 2017.

[13] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, "Fast and robust matching for multimodal remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9059–9070, Nov. 2019.

[14] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2020, doi: 10.1109/TIP.2019.2959244.

[15] X. Xie, Y. Zhang, X. Ling, and X. Wang, "A novel extended phase correlation algorithm based on log-Gabor filtering for multimodal remote sensing image registration," *Int. J. Remote Sens.*, vol. 40, no. 14, pp. 5429–5453, Feb. 2019.

[16] F. Wu *et al.*, "Visible and infrared image registration based on visual salient features," *J. Electron. Imag.*, vol. 24, no. 5, Sep. 2015, Art. no. 053017.

[17] X. Liu, Z. Lei, Q. Yu, X. Zhang, Y. Shang, and W. Hou, "Multi-modal image matching based on local frequency information," *EURASIP J. Adv. Signal Process.*, vol. 2013, no. 1, p. 3, Jan. 2013.

[18] X. Liu, Y. Ai, B. Tian, and D. Cao, "Robust and fast registration of infrared and visible images for electro-optical pod," *IEEE Trans. Ind. Electron.*, vol. 66, no. 2, pp. 1335–1344, Feb. 2019.

[19] D. Zhao, Y. Yang, Z. Ji, and X. Hu, "Rapid multimodality registration based on MM-SURF," *Neurocomputing*, vol. 131, pp. 87–97, May 2014.

[20] C. Zhao, H. Zhao, J. Lv, S. Sun, and B. Li, "Multimodal image matching based on multimodality robust line segment descriptor," *Neurocomputing*, vol. 177, pp. 290–303, Feb. 2016.

[21] Z. Yang, T. Dan, and Y. Yang, "Multi-temporal remote sensing image registration using deep convolutional features," *IEEE Access*, vol. 6, pp. 38544–38555, 2018.

[22] A. Zampieri, G. Charpiat, N. Girard, and Y. Tarabalka, "Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 657–673.

[23] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: A survey," *Mach. Vis. Appl.*, vol. 31, nos. 1–2, p. 8, Jan. 2020.

[24] D. Quan *et al.*, "Multi-relation attention network for image patch matching," *IEEE Trans. Image Process.*, vol. 30, pp. 7127–7142, 2021.

[25] X. Cao, J. Yang, L. Wang, Z. Xue, Q. Wang, and D. Shen, "Deep learning based inter-modality image registration supervised by intra-modality similarity," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, Sep. 2018, pp. 55–63.

[26] E. Ben Baruch and Y. Keller, "Joint detection and matching of feature points in multimodal images," 2018, *arXiv:1810.12941*.

[27] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1788–1800, Aug. 2019.

[28] J. Weickert, B. M. T. H. Romeny, and M. A. Viergever, "Efficient and reliable schemes for nonlinear diffusion filtering," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 398–410, Mar. 1998.

[29] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629–639, Jul. 1990.

[30] B. Yu, D. Gabriel, L. Noble, and K.-N. An, "Estimate of the optimum cutoff frequency for the Butterworth low-pass digital filter," *J. Appl. Biomech.*, vol. 15, no. 3, pp. 318–329, Aug. 1999.

[31] P. Kovesi, "Image features from phase congruency," *J. Comput. Vis. Res.*, vol. 1, no. 3, pp. 1–26, 1999.

[32] A. Makandar and B. Halalli, "Image enhancement techniques using highpass and lowpass filters," *Int. J. Comput. Appl.*, vol. 109, no. 14, pp. 12–15, Jan. 2015.

[33] W. Ma *et al.*, "Remote sensing image registration with modified SIFT and enhanced feature matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 3–7, Jan. 2017.

[34] J. Shi, "Good features to track," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 1994, pp. 593–600.

[35] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[36] Q. Li, S. Qi, Y. Shen, D. Ni, H. Zhang, and T. Wang, "Multispectral image alignment with nonlinear scale-invariant keypoint and enhanced local feature matrix," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 7, pp. 1551–1555, Jul. 2015.

[37] 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.

[38] B. Kupfer, N. S. Netanyahu, and I. Shimshoni, "An efficient SIFT-based mode-seeking algorithm for sub-pixel registration of remotely sensed images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 379–383, Feb. 2015.

[39] Y. Wu, W. Ma, M. Gong, L. Su, and L. Jiao, "A novel point-matching algorithm based on fast sample consensus for image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 43–47, Jan. 2015.

[40] J. Li, Q. Hu, M. Ai, and S. Wang, "A geometric estimation technique based on adaptive M-estimators: Algorithm and applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5613–5626, 2021.

Yongxiang Yao received the M.S. degree in cartography and geographic information system in 2016. He is currently pursuing the Ph.D. degree with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China. He has published more than eight research articles. His research interests include multi-modal image matching and combined block adjustment with multisource data set.

Yongjun Zhang received the B.S. degree in geodesy, the M.S. degree in geodesy and surveying engineering, and the Ph.D. degree in geodesy and photography from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively. He is currently the Dean of the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 180 research articles and one book. He holds 30 Chinese patents and 32 copyright registered computer software. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource data sets, object information extraction and modeling with artificial intelligence, integration of LiDAR point clouds and images, and 3D city model reconstruction. He is the Co-Editor-in-Chief of *The Photogrammetric Record*.

Yi Wan received the B.S. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2013 and 2018, respectively. He is currently an Associate Professor of photogrammetry and remote sensing with the School of Remote Sensing and Information Engineering, WHU. His research interests include photogrammetry, computer vision, 3D reconstruction, satellite image interpretation, and change detection.

Xinyi Liu received the B.S. and Ph.D. degrees from the School of Remote Sensing and Information Engineering, Wuhan University, China, in 2014 and 2020, respectively. She is currently a Postdoctoral Researcher with Wuhan University. Her research interests include 3D reconstruction, LiDAR and image integration, and texture mapping.

Xiaohu Yan received the B.S. degree from Huazhong Agricultural University in 2008, the M.S. degree from North China Electric Power University in 2010, and the Ph.D. degree from Wuhan University in 2017. He is currently a Lecturer at Shenzhen Polytechnic. His research interests include image registration, image matching, and optimization algorithm.

Jiayuan Li received the B.Eng., M.Eng., and Ph.D. degrees from the School of Remote Sensing Information Engineering, Wuhan University, Wuhan, China. He is currently an Associate Researcher with Wuhan University. He has authored more than 30 peer-reviewed articles in international journals. His research is mainly focused on SLAM, image matching, and point cloud registration. He was awarded the Best Youth Author Award by ISPRS in 2021 and the Talbert Abrams Award by ASPRS in 2018.