Feature Matching of Multimodal Images Based on Nonlinear Diffusion and Progressive Filtering

Qiang Xiong, Shenghui Fang, Yi Peng, Yan Gong, and Xiaojuan Liu

***Abstract*—Traditional image feature matching methods cannot obtain satisfactory results for multimodal images in most cases because different imaging mechanisms bring significant nonlinear radiation distortion differences and geometric distortion. The key to multimodal image matching is trying to eliminate the nonlinear radiation distortion and extract more robust features. This article proposes a new robust feature matching method for multimodal images. Our method starts by detecting feature points on phase congruency maps in nonlinear scale space and then removing mismatches by progressive filtering. Specifically, the phase congruency maps are generated by the Log-Gabor filter (LGF). Then, the feature points on phase congruency maps are detected in nonlinear scale space constructed by the nonlinear diffusion filter. Subsequently, the structure descriptor is established by the LGF, and the initial correspondences are constructed by bilateral matching. Finally, an iterative strategy is used to remove mismatches by progressive filtering. We perform comparison experiments on our proposed method with the SIFT, RIFT, VFC, LLT, LPM, and mTopKPR methods using multimodal images. The algorithms of each method are comprehensively evaluated both qualitatively and quantitatively. Our experimental results indicate the superiority of our method over the other six matching methods.**

***Index Terms*—Feature matching, multimodal images, phase congruency, progressive filtering.**

I. Introduction

FEATURE matching is usually defined as extracting the correct feature correspondence from the overlapping regions of two or more images [1]. It has been wildly used in photogrammetry and remote sensing [2], computer vision [3], and artificial intelligence [4]. However, because of the imaging characteristics of the sensor itself and distortion of light and atmosphere, the nonlinear radiation distortions between images always exist [5]. In general, images with overlapping areas obtained by different sensors at different times and imaging angles are called multimodal images, which are more prone to nonlinear radiation distortion [6]. Traditional image matching methods cannot solve the nonlinear radiation distortion between multimodal images. Developing algorithms for generic models, capable of working across multimodal images, will be essential to employing remote sensing as a practical and high-through tool to assist in image matching.

In the past few decades, there have been many image matching methods. Existing matching methods can be broadly classified into feature based, area based, and learning based.

Feature-based matching methods extract salient features such as points, lines, and surfaces and then establish the corresponding reliable relationship according to the similarity descriptors. Point features are the most widely used in feature matching. The traditional point feature refers to the corner point because it contains the local gray feature. Moravec [7] first proposed the corner point extraction algorithm in 1977, Harris and Stephens [8] improved the Moravec operator and then proposed the Harris operator [9]. SIFT [10] is one of the most widespread and effective feature-based matching methods, which extracts feature points in the Gaussian scale space. In recent years, a series of optimized SIFT algorithms have been proposed, such as SURF [11], PCA-SIFT [12], ASIFT [13], UC-SIFT [14], SAR-SIFT [15], and AB-SIFT [16]. However, image matching based on feature is faced with two problems: 1) in the feature description stage, when the principal direction estimation (PDE) is used to resist rotation distortion, a lot of homonymous points will be removed due to incorrect PDE. 2) In the feature matching stage, due to the nonlinear radiation difference of multimodal images, the corresponding points are often unable to be recognized, resulting in numerous anomalies in the matching results.

Area-based matching, otherwise known as template matching, is achieved by searching the reference image through the predefined template window on the input image, which utilizes the similarity measure of the image without involving feature detection [17]. There are some common similarity measures, such as the sum of squares of gray difference (SSD) [18], correlation coefficient (CC) [19], and mutual information (MI) [20]. Although SSD is simple and efficient, it is very susceptible to gray differences. CC has been extensively used because of its linear intensity changes and high computational efficiency. Despite the fact that MI can resist the intensity difference between images well, the large amount of computation limits the wide application of image matching. However, these area-based matching methods usually achieve locally optimal solutions,

Manuscript received 10 June 2022; revised 27 July 2022; accepted 17 August 2022. Date of publication 22 August 2022; date of current version 5 September 2022. This work was supported in part by the National Key Research and Development Project under Grant 2016YFD0101105 and in part by National High-tech Research and Development Program under Grant 2013AA102401. *(Corresponding author: Shenghui Fang.)*

Qiang Xiong, Shenghui Fang, Yi Peng, and Yan Gong are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: xiongq2019@whu.edu.cn; shfang@whu.edu.cn; ypeng@whu.edu.cn; gongyan@whu.edu.cn).

Xiaojuan Liu is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: xiaojliu@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3200424

especially when there are nonlinear radiation distortions among multimodal images. At the same time, the local optimization process of image matching has high computational complexity.

The learning-based method needs to use the manual feature extractor to extract the key points on the image block as the matching candidate points, and then take the image block of a specific size as the center to input the convolutional neural network (CNN) for feature description. CNN has been applied to learn the similarity between image blocks from the reference image and feature matching image, respectively, and the center point corresponding to the image block meeting the similarity condition is regarded as the corresponding point. There are some representative learning-based matching methods, such as LIRT [21], NCN [22], and LF-NET [23]. However, studies of multimodal image matching by the learning-based method are relatively rare. The main reasons are as follows.

1) The nonlinear radiation distortions of multimodal images will lead to the inability of learning-based methods relying on low-level structure and texture information to extract the corresponding features.
2) It needs a lot of expert knowledge and manpower for training, but manually annotated multimodal image datasets are very rare.
3) It is difficult to use a model trained on one modal image to match another modal image.

Recently, Zhou et al. [24] used the deep learning method to refine the image structure features and put the multidirectional gradient features of the image pair into the pseudo-Siamese neural network. Although it can capture the finer common features between SAR images and optical images, it does not have rotation and scale invariance. Ye et al. [25] proposed a multiscale framework with unsupervised learning (MU-Net) for remote sensing image registration. MU-Net does not need ground truth labels, and directly learns the end-to-end mapping from image pairs to their transformation parameters. However, extracting the structural features of image pairs with large-scale differences and unclear textures is difficult.

Several recent kinds of research have reported that the structural information is more stable than the gradient or intensity information between multimodal images [26], [27], [28]. LNIFT [29] converts different modes into the same modal based on the local normalized filter, which is robust to severe nonlinear radiation distortion. Although it realizes rotation invariance, it does not have scale invariance. LAM [30] is a robust and effective mismatching elimination algorithm suitable for rigid and nonrigid image matching. The limitation of LAM is that it only considers geometric constraints and is not suitable for image matching that does not meet local affine constraints. Phase congruency (PC) image [31] has the structural information, and there are already lots of methods for feature descriptors based on PC, such as EOH [32], PIIFD [33], and DLSS [34]. However, image noise seriously affects the detection accuracy of PC [35]. Some researchers used Log-Gabor filter (LGF) to improve PC models, such as HOPC [36], PCSD [37], and RIFT [6]. HOPC and RIFT algorithms are typically matching methods of multimodal images. The HOPC algorithm extends the phase consistency algorithm and adds phase direction statistics to enhance the robustness of the description process. However, HOPC has three significant deficiencies as follows.

1) It needs to know the geographic information for the image. However, various multimodal images do not have geographic information.
2) It is sensitive to geometric distortions such as rotation and scale.
3) It is susceptible to nonlinear radiation distortions because it detects feature points by the Harris operator.

The RIFT algorithm uses the maximum index map for feature description, which takes into account the rotation invariance. However, RIFT has two disadvantages: 1) it uses the "convolution sequence ring" to deal with the rotation distortion of the image, which may lose some spatial information, resulting in the lack of rotation invariance and unfavorable feature matching; and 2) it does not consider the scale invariance and is easily affected by the image scale distortion.

Although numerous image matching methods have appeared in the past few decades, there is still no unified feature matching framework with rotation-radiation-scale-invariant to automatically match multimodal images. In this article, we propose an automatic feature matching method for multimodal images based on progressive filtering. The main advantages of the proposed method are as follows: 1) A new feature descriptor is established according to the orientation of the average oriented amplitude map based on the two-dimensional (2-D) LGF (2D-LGF), which is more robust in nonlinear radiation distortion differences. 2) It converts the initial correspondence to a convolution filtering and removes the false correspondences by progressive filtering, and then restores the structural consistency of the motion vectors.

The rest of this article is organized as follows. Section II introduces the methodology of the proposed method. Section III presents the experimental results and discussions. Finally, Section IV concludes this article.

II. Methodology

The implementation of our method is illustrated in Fig. 1. It consists of four steps as follows.

1) PC maps [see Fig. 1(b)] of multimodal image pair [see Fig. 1(a)] are constructed by LGF.
2) Feature points are detected in nonlinear scale space, which is invariant to scale [see Fig. 1(c)].
3) LGF is used to establish a structural descriptor, and initial correspondences are constructed by bilateral matching [see Fig. 1(d)].
4) Progressive filtering is used for the convolution operation to restore the real smooth field, and mismatches are removed through the structural congruence of the motion vector step-by-step [see Fig. 1(e)–(h)].

A. Phase Congruency Based on 2D-LGF

Sun et al. [38] first proposed the phase congruency (PC) theory, which points out that the perception of image features by the eyes mainly depends on phase rather than amplitude. Different from the feature detection method based on the gradient in the spatial domain, PC is a feature perception model

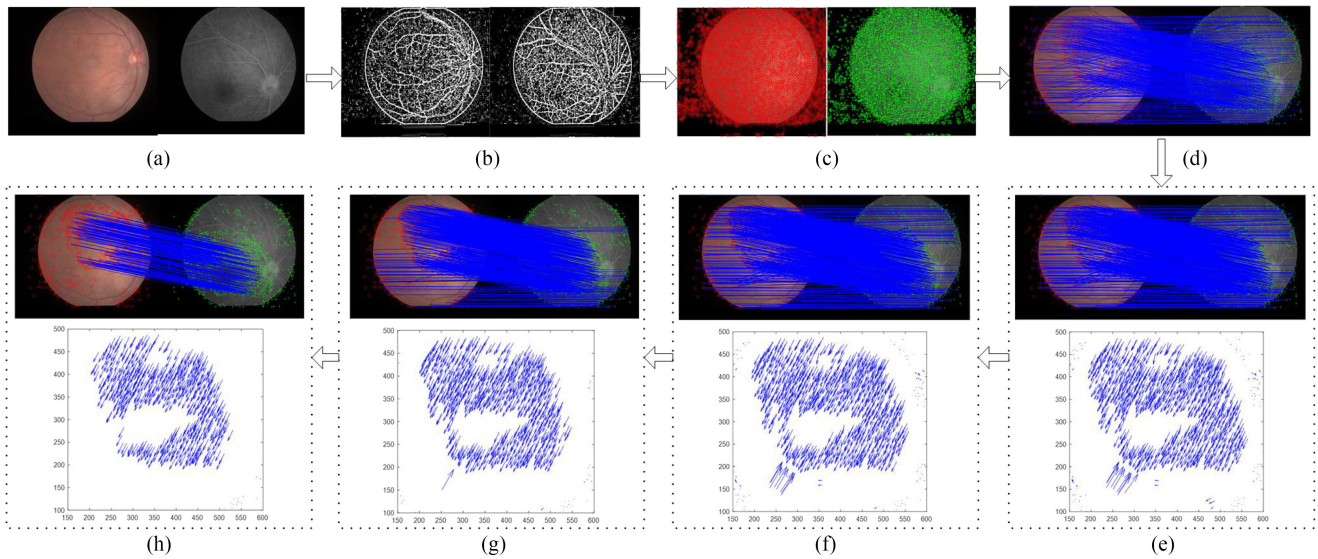

Fig. 1. Implementation of feature matching by using the proposed method. (a) Multimodal image pair. (b) PC maps. (c) Feature points (marked by red and green points). (d) Initial feature matching. (e)–(h) Matching results (top) and motion vectors (bottom) when the iteration threshold is τ_1–τ_4.

that detects local energy in the frequency domain. Although the PC model is robust to illumination and contrast differences, the noise and aliasing effects of the edge structures greatly affect the accuracy of feature detection. Do and Vetterli [39] extended PC to be robust to noise by the 2D-LGF. 2D-LGF is constructed by a Gaussian function in multiple directions. In the frequency domain, 2D-LGF is defined as

$$L(\rho,\theta)=\exp(-(\rho-\rho_s)/(2\sigma_\rho^2))\exp(-(\theta-\theta_{\text{so}})/(2\sigma_\theta^2)) \tag{1}$$

where ρ denotes radius and θ denotes angle in log-polar, ρ_s is the center frequency, θ_{so} is the center orientation of 2D-LGF at scale s and orientation o, σ_ρ is the width parameter of ρ, and σ_θ is the width parameter of θ.

The equation of 2D-LGF in the spatial domain is defined as follows:

$$L_{\text{so}}(x,y)=L_{\text{so}}^e(x,y)+iL_{\text{so}}^o(x,y) \tag{2}$$

where $L_{\text{so}}^e(x,\ y)$ represents the even-symmetric Log-Gabor wavelet and $L_{\text{so}}^o(x,y)$ represents the odd-symmetric Log-Gabor wavelet.

The input image $I(x,y)$ is convolved with 2D-LGF to construct the response components at different scales and orientations. The convolution is defined as follows:

$$[E_{\text{so}}(x,y),O_{\text{so}}(x,y)]=[I(x,y)*L_{\text{so}}^e(x,y)+I(x,y)*L_{\text{so}}^o(x,y)] \tag{3}$$

where $E_{\text{so}}(x,y)$ and $O_{\text{so}}(x,y)$ are the convolution response of $L_{\text{so}}^e(x,y)$ and $L_{\text{no}}^o(x,y)$ at scale s and orientation o.

The amplitude response component of $I(x,y)$ can be obtained by

$$A_{\text{so}}(x,y)=\sqrt{E_{\text{so}}(x,y)^2+O_{\text{so}}(x,y)^2}. \tag{4}$$

The phase response component of $I(x,y)$ can be obtained by

$$\varphi_{\text{so}}(x,y)=\arctan(E_{\text{so}}(x,y),O_{\text{so}}(x,y)). \tag{5}$$

Based on the $A_{\text{so}}(x,y)$ and the $\varphi_{\text{so}}(x,y)$, PC is calculated as the ratio of the sum of scale weighted and energy of noise compensation in all directions at point (x, y) to the sum of the average direction and amplitude on the filter response. The final PC model is as follows:

$$\text{PC}(x,y)=\frac{\sum_s\sum_o W_o(x,y)\lfloor A_{\text{so}}(x,y)\Delta\Phi_{\text{so}}(x,y)-T\rfloor}{\sum_s\sum_o A_{\text{so}}(x,y)+\delta} \tag{6}$$

where $W_o(x,\ y)$ is a weight function, δ is a minimum, $\Delta\Phi_{so}(x,y)$ is a phase deviation function, and the operator $\lfloor\cdot\rfloor$ prevents the enclosed quantity from being negative.

B. Feature Detection

Traditional feature detection algorithms (e.g., SIFT and SURF) construct linear Gaussian scale space. However, Gaussian scale space will result in the loss of detailed information about the image. Nonlinear scale space (NSS) is expected to solve this problem, but the traditional method based on forwarding the Euler scheme has a very short iterative convergence step [40]. In this article, the nonlinear diffusion filter is used to construct a stable NSS, it describes the illumination variation at different scales, which can be expressed as

$$\frac{\partial I}{\partial t}=div\left(\exp\left(-\frac{|\nabla I_\sigma|^2}{k^2}\right)\cdot\nabla I\right) \tag{7}$$

where *div* is the divergence operator, ∇ is the gradient operator, k is the contrast factor, and ∇I_σ represents the gradient after Gaussian filtering.

Then, the differential equation is approximated using dispersion analysis. Thus, the discretization of (7) can be expressed as

$$\frac{I^{i+1}-I^i}{\Delta t}=\sum_{l=1}^{m}A_l(I^i)I^{i+1} \tag{8}$$

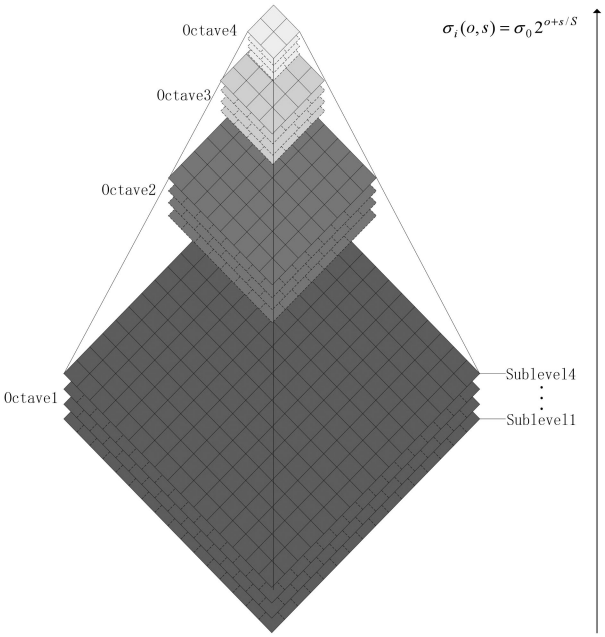

Fig. 2. Construction of the nonlinear scale space.

where *l* indicates the direction, A_l is the derivative along *l*, and Δt is the time step.

Scale levels increased logarithmically when a nonlinear scale space is constructed. Different from SIFT, the same resolution is used for all levels of the nonlinear scale space, and the scale between each level is described as

$$\sigma_i(o,s) \ = \sigma_0 \ 2^{o+s/S} \tag{9}$$

where o represents the index of octave *O*, s represents the index of sub-leave *S*, and σ_0 is the initial value of scale σ_i. As shown in Fig. 2, the image of the last sublevel in each octave is down sampled, and the down sampled image is used as the initial image for the next octave. After the creation of the NSS, attained multiscale images can preserve structural and detailed information. Therefore, it expects that our proposed method can detect more feature points.

The NSS is based on the time unit, so we need to convert scale unit σ_i to time unit t_i, and the relationship is as follows:

$$t_i = \sigma_i^2/2. \tag{10}$$

Since $\Delta t \ = \ t_{i+1} \ - \ t_i$, (8) can be described as

$$I^{i+1} = 2I^i/(2\mathbf{I} - (\sigma_{i+1}^2 - \sigma_i^2)\sum_{l=1}^{m} A_l(I^i)) \tag{11}$$

where I^{i+1} represents the solution of the nonlinear diffusion equation.

After constructing the NSS, feature points can be detected by searching local maximum points of the Hessian matrix. The Hessian matrix is calculated as follows:

$$I_H = \sigma^2 \left(I_{xx} I_{yy} - I_{xy}^2\right) \tag{12}$$

where I_{xx} is the second order derivative in *x* direction, I_{yy} is the second order derivative in *y* direction, and $I_{xy}$is the second order cross derivative in *x* and *y* directions.

C. Feature Description

1) Average Oriented Amplitude Map (AAM): Aguilera and Sappa [41] concluded that the distribution of the high-frequency amplitude components is robust to nonlinear radiation variations. The RIFT descriptor calculates the sum amplitudes of all scales to obtain a Log-Gabor layer. Different from RIFT, we use the average amplitude of the different scales to calculate the distribution. The average amplitude is calculated by adding the amplitudes of all scales for each orientation o and then dividing N_s, and the formula is as follows:

$$\bar{A}_o(x,y) = \left(\sum_{s=1}^{N_s} A_{so}(x,y)\right)/N_s \tag{13}$$

where $o \in [1, N_o]$ and $s \in [1, N_s]$, N_o represents the number of orientations. N_s represents the number of scales, $\bar{A}_o(x,y)$ is defined as the AAM of orientation o.

2) Direction of Phase Congruency: The odd-symmetric wavelet of 2D-LGF is a smooth derivative filter whose convolution result represents the energy change. Therefore, the direction of PC can be constructed by using the odd-symmetric wavelet of 2D-LGF. The odd-symmetric wavelet can obtain o convolution results according to different directions, which are projected to the *x*- and *y*-axes, respectively. The direction of the PC is defined as

$$\varphi_{\mathrm{so}} = \arctan\left(\sum_{\theta} \mathrm{PC}_{\mathrm{so}}(\theta)\sin(\theta)\right), \sum_{\theta} \mathrm{PC}_{\mathrm{so}}(\theta)\cos(\theta)+\delta)) \tag{14}$$

where $\mathrm{PC}_{\mathrm{so}}(\theta)$ is the convolution result of odd-symmetric wavelet at the direction θ and δ is a minimum in case the denominator is zero.

For multimodal images, gradient inversion may occur when the PC direction exceeds π. In this article, the direction of PC is restricted to 0 to π as

$$\varphi'_{so} = \begin{cases} \varphi_{so} \varphi_{so} \in [0, \pi] \\ \varphi_{so} - \pi \varphi_{so} \in (\pi, 2\pi]. \end{cases} \tag{15}$$

3) Feature Descriptor: For each feature point, a local area with P×P pixels is divided into 6×6 patches, and the block distribution histograms are established by the average oriented amplitude maps and amplitude maps. The local feature descriptors are generated by merging the histograms of each patch. The histogram is divided into 36 equal parts at intervals of 10°, and the phase consistency gradients of each equal part are counted. The peak direction of the histogram is selected as the main direction of the features. The specific construction process is shown in Fig. 3, and the main steps of descriptor construction are as follows.

1) The odd and even convolution at scale s and direction o is calculated using 2D-LGF.
2) The average oriented amplitude map is calculated by (13).
3) The PC orientation map is calculated by (14) and (15).

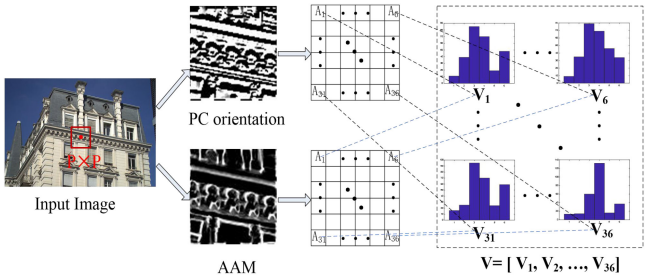

Fig. 3. Construction of proposed feature descriptor.

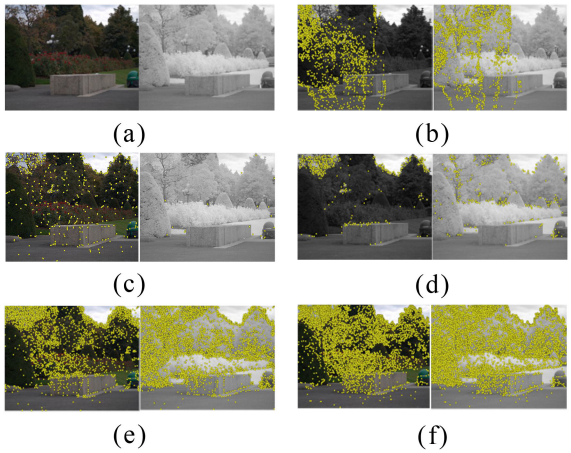

Fig. 4. Feature points are detected by five methods (a) visible image (left) and near-infrared image (right). (b) SIFT. (c) OS-SIFT. (d) SURF. (e) RIFT. (f) FMPF. (Yellow points represent feature points.).

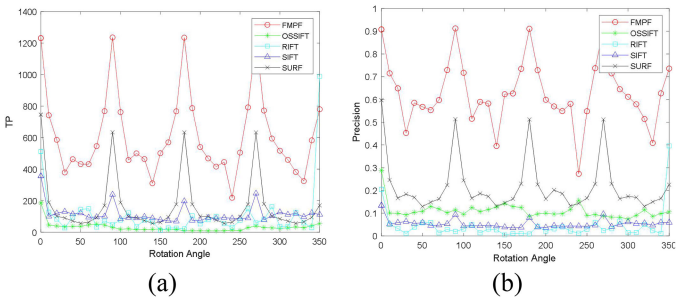

Fig. 5. TP and precision of the rotation angle from 0° to 350°. (a) TP. (b) Precision.

4) The PC orientation maps and AAM are divided into 6×6 patches.
5) The feature vector of each patch is calculated by using six bins of PC orientations. The sample adds to the histogram is the element of the corresponding position on the AAM. To interpolate the peak position for better accuracy, a parabola is fitted to the three histogram values closest to each peak.
6) The descriptor of each feature point is composed of feature vectors of 36 subareas calculated by step 5. The feature vector of patch A_1 corresponds to V_1, so the feature vector is as follows: $\mathbf{V} = [\mathbf{V_1}, \mathbf{V_2}, \ldots, \mathbf{V_{36}}]$.

D. Feature Matching

Jiang et al. [42] used linear adaptive filtering (LAF) to eliminate mismatches, which takes the local structural consistency as the constraint condition and transforms the local topology space into domain sequence space for consistency measurement to improve the matching accuracy. However, the LAF algorithm uses SIFT algorithm to establish the initial matching set, which largely depends on the local consistency between the potential real inliers. Due to the large nonlinear radiation difference between multimodal images, the initial matching obtained by SIFT algorithm contains a large number of mismatches, and the assumption of local structural consistency cannot satisfy. Therefore, LAF is not suitable for multimodal image matching.

In this article, we use progressive filtering to eliminate mismatches by the structural consistency of the motion vector. The motion vector is defined as the difference between the coordinates of two points with a matching relationship in the sensed image and reference image. The motion vectors formed by correct matches of adjacent pixels have strong structural consistency. The motion vectors formed by mismatches are different from the correct matches and can be regarded as outliers in the smooth field. Therefore, for multimodal image matching, we focus on how to preserve or recover the correctly matched smooth field with a large number of outliers. In this article, we transform the initial matching set into a matrix and removed outliers by kernel convolution. The elements in the matrix can represent the spatial properties of the initial matches, so we can solve the matching problem with kernel convolution filtering.

For two images I and I' with overlapping areas, the initial matching set $D = \{(I_i, I_i')\}_{i=1}^{N}$ is extracted by using the proposed method, where $I_i = (x_i, y_i)^{\mathrm{T}}$ and $I_i' = (x_i', y_i')^{\mathrm{T}}$ are the coordinates of the ith feature matching point on images I and I'. The difference between I_i' and I_i represents motion vector, which is defined as $\mathbf{m}_i = I_i' - I_i$. The initial matching set D can be transformed into $D' = \{(I_i, m_i)\}_{i=1}^{N}$, then the correct matches are found from the initial correspondences according to the structure congruence of motion vectors.

It is practicable to calculate the average motion vectors of initial correspondences and eliminate mismatches by checking the consistency between each image pair. Therefore, we divide D' into $c \times c$ nonoverlapping areas and define the average motion vector in the (m, n)th area as

$$\bar{\mathbf{m}}_{m,n} = \begin{cases} \frac{1}{A_{m,n}} \sum \mathbf{m_i}, & A_{m,n} > 0 \\ 0, & A_{m,n} = 0 \end{cases} \tag{16}$$

where $A_{m,n}$ is the number of motion vectors in the (m, n)th area.

The initial set of matches is converted into the estimation of average motion matrix $\bar{\mathbf{m}}_{m,n}$. We define the deviation between the initial motion vectors and the average motion vectors as $\varepsilon = \{e_i = \mathbf{m}_i - \bar{\mathbf{m}}_{m,n}\}_{i=1}^{A_{m,n}}$. Due to the random distribution of initial motion vectors, we assume that the inlier set obeys the normal distribution $e_{\text{inlier}} \sim N(0, \sigma^2\mathbf{I})$ and the outlier set obeys the uniform distribution $e_{\text{outlier}} \sim U(-b\mathbf{I}, b\mathbf{I})$, where $\mathbf{0}$ is the 2-D 0 vector, $\mathbf{I}$ is a 2×2 identity matrix.

Algorithm 1: The FMPF Algorithm.

Input: Multi-modal images
Output: Correct matching set I^*
Calculate PC maps by using (1) - (6);
Obtain the initial matching set $\mathbf{D} = \{(x_i, y_i)\}_{i=1}^{N}$ by using (7)–(12);
Initialize c, k and τ;
Convert D to D' and divide it into $c \times c$ grids;
Iteration:
Calculate $\bar{\mathbf{m}}$ by using (16);
Calculate $\tilde{\mathbf{m}}$ by using (17);
Calculate d_i by using (19);
Determine correct matching set I^* by using (20);
Update τ;
Until Convergence;
Return I^*

To link the initial correspondences in the surrounding units and improve the robustness and filtering performance, we calculate the typical motion vector of the current unit by convolution theory. The typical motion vector $\tilde{\mathbf{m}}_{m,n}$ is defined as follows:

$$\tilde{\mathbf{m}}_{m,n} = (\mathbf{w} \cdot \bar{\mathbf{m}}_{m,n}) * \mathbf{k} / (\mathbf{w} * \mathbf{k} + \delta) \tag{17}$$

where $\mathbf{w}$ is a counting matrix, $|\mathbf{w}_{m,n}| = A_{m,n}$, δ is a minimum in case the denominator is zero. $\mathbf{k}$ is a $k \times k$ Gaussian kernel distance matrix, which is defined as

$$\mathbf{k}_{i,j} = \exp\{-\mathbf{n}_{i,j}\} / \left(\sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \exp\{-\mathbf{n}_{i,j}\} \right), \mathbf{n}_{i,j} = \|c_{i,j} - c^*\|_2 \tag{18}$$

where $c_{i,j} = (i,\ j)^{\mathrm{T}}$ and $c^* = (k/2, k/2)^{\mathrm{T}}$.

The deviation between $\mathbf{m}_i$ and $\tilde{\mathbf{m}}_{m,n}$ is defined as

$$d_i = 1 - \exp\left\{-\|\mathbf{m}_i - \tilde{\mathbf{m}}_{m,n}\|^2 / 0.08\right\}. \tag{19}$$

Then by comparing the deviation d_i with the threshold τ, the correct matching set I^* can be obtained as

$$I^* = \{(I_i, I_i') : d_i \leq \tau, i \in N\}. \tag{20}$$

As shown in Fig. 1(d), the initial matching set contains many outliers, which makes it difficult to separate the outliers from the initial matches. As shown in Fig. 1(e), only a litter of mismatches can be filtered out by the given threshold τ. To solve this problem, an iterative method is used to anneal the threshold τ to remove outliers until convergence.

As shown in Fig. 1(f)–(h), mismatches are filtered out gradually as the iterations proceed until the correct matching set is obtained. Meanwhile, the motion vectors corresponding to the correct matching set have structural consistency. Since the feature matching method is based on progressive filtering, the algorithm is named FMPF and the whole algorithm is summarized in Algorithm 1.

III. Experimental Settings and Results

In this section, we perform extensive experiments to test the performance of the proposed method and compare it with six advanced feature matching methods such as SIFT [10], RIFT [6], VFC [43], LLT [44], LPM [45], and mTopKRP [46]. All experiments are implemented on a laptop with 32 GB RAM, 2.6 GHz Intel Core i7-10750h CPU, and MATLAB R2021a compiler.

A. Settings

1) Evaluation Criteria: Repetition rate (RR) is used to illustrate the robustness of the feature detection of FMPF. RR describes the percentage of repeatable features detected in the image pair and is defined as

$$\mathrm{RR} = \frac{2 \cdot n}{n_1 + n_2} \tag{21}$$

where *n* is the number of homonymous points, n_1 is the number of feature points on the reference image, and n_2 is the number of feature points on the target image.

Precision, recall, and F-score are used to evaluate the feature matching performance with the following definitions:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{22}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{23}$$

$$\mathrm{F-score} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{24}$$

where TP, FP, and FN represent true positive, false positive, and false negative, respectively.

In addition, root-mean-square error (RMSE), MEAN, and standard deviation (Std) are used to measure the registration accuracy with the following definitions:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i^s - x_i^r)^2 + (y_i^s - y_i^r)^2} \tag{25}$$

$$\mathrm{MEAN} = \left(\sum_{i=1}^{N} \sqrt{(x_i^s - x_i^r)^2 + (y_i^s - y_i^r)^2} \right) / N \tag{26}$$

$$\mathrm{Std} = \sqrt{\frac{\sum_{i=1}^{N} (\sqrt{(x_i^s - x_i^r)^2 + (y_i^s - y_i^r)^2} - \mathrm{MEAN})}{N}} \tag{27}$$

where N represents the number of correct matches and (x_i^r, y_i^r) and (x_i^s, y_i^s) are the coordinate of the *i*th feature matching point on the reference and registration image.

2) Parameter Settings: When using progressive filtering to eliminate mismatches, parameters *c*, *k,* and τ severely influence the filtering results. We restrict *c* to an odd number between 15 and 30, and set k not greater than *c*/3, as follows:

$$\begin{cases} c = \min\{\max\{[\sqrt{N}], 15\}, 30\} \\ k = \mathrm{odd}(c/3) \end{cases} \tag{28}$$

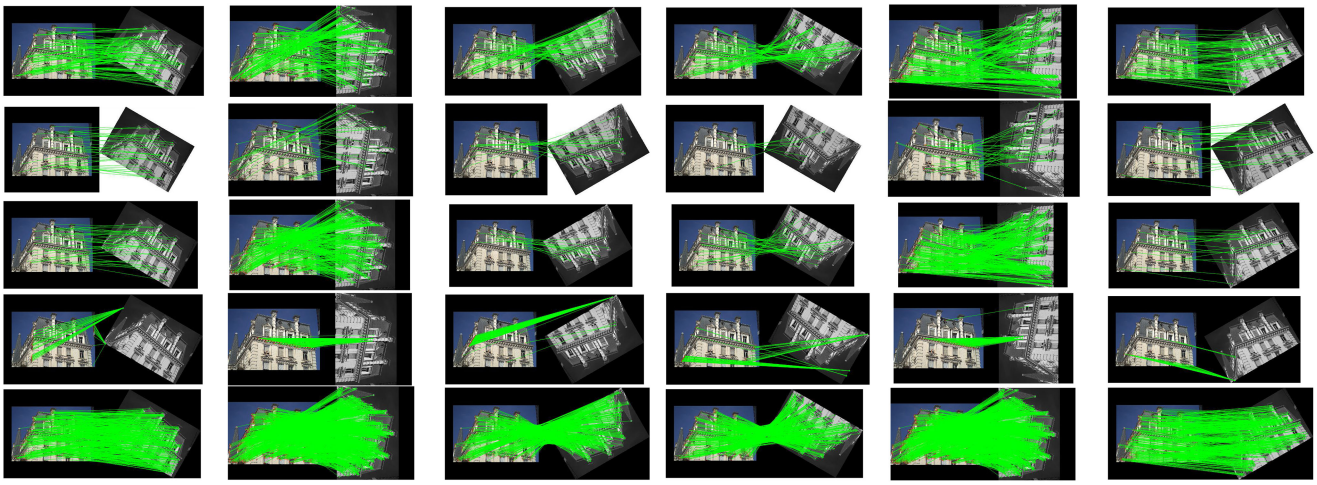

Fig. 6. Feature matching results of five methods with six rotation angles. (From left to right) 30°, 90°, 150°, 210°, 270°, and 330°. (From top to bottom) SIFT, OS-SIFT, SURF, RIFT, and FMPF.

where N is the number of initial correspondences, [] rounds the elements, and odd(c/3) denotes the odd number not greater than c/3.

As shown in Fig. 1(d), there are many mismatches in the initial matches. A progressive iteration method is used to gradually separate outliers and outliers, which is similar to the simulated annealing strategy. We set a larger threshold at the beginning of the iteration, and gradually lower the threshold as the number of iterations increases. It eliminates outliers from coarse to fine and optimizes the threshold τ. The inlier set is approximated with the result of each iteration until convergence. In our experiments, reliable matching performance is obtained in four iterations, and the threshold for each iteration can be set to 0.8, 0.2, 0.1, and 0.01.

B. Detector Evaluation

1) Repetition Rate: A pair of visible-near-infrared images are used as test images [see Fig. 4(a)], and the feature points are detected by five methods (i.e., SIFT, OS-SIFT, SURF, RIFT, and FMPF). The feature detection results are shown in Fig. 4(b)–(f). In Fig. 4(b), the feature points detected by SIFT are unevenly distributed. In Fig. 4(c) and (d), the number of feature points detected on the near-infrared image is very small, resulting in a low repetition rate. In Fig. 4(e) and (f), both RIFT and FMPF can detect more feature points, but the feature points detected by FMPF are more evenly distributed. Repetition rates of five methods are computed by using (21), which are 0.1260, 0.0147, 0.0996, 0.1308, and 0.2512, respectively. The repetition rate of FMPF is higher than that of the other four methods, which indicates that the feature detection performance of FMPF is more robust to multimodal images.

2) Rotation Invariance: To verify the rotation invariance of FMPF, a pair of visible-NIR images are chosen for testing. Thirty-six NIR images are obtained by rotating the NIR images from 0–350° with an interval of 10°. These 36 near-infrared images and the visible image consist of 36 image pairs. SIFT, OS-SIFT, SURF, RIFT, and FMPF are used to match those image pairs. Fig. 5 shows the results of the TP and precision on all image pairs. It can be found that the TP and precision of FMPF are higher than that of the other four methods. Specifically, although TP and precision of FMPF are different at different rotation angles, all TPs are greater than 200 and all precisions are greater than 0.2. The results show that the FMPF has rotation invariance in the whole 360° range. In order to qualitatively analyze the matching performance, we selected the matching results at six rotation angles (i.e., 30°, 90°, 150°, 210°, 270°, and 330°) of each method, as shown in Fig. 6. We can find that the matching performance of RIFT is the worst, and there are several-for-one correspondences among the six matching results. Compared with SIFT, OS-SIFT, SURF, and RIFT, FMPF shows the best matching performance.

3) Scale Invariance: To study the impact of scale change, a group of SAR-optical images with different scales are chosen for testing. The scales are 0.5, 1, 1.5, 2, and 2.5, respectively. The matching results of FMPF are shown in Fig. 7, it can be seen that the matching is successful when the scales are 0.5, 1, 1.5, 2, and 2.5, respectively. Although the number of correct matching decreases with the increase of scale, the distribution of feature matching points is relatively uniform. It indicates that FMPF has scale invariance.

C. Experimental Results on MIDs

Four multimodal image datasets (MIDs) are used to evaluate the performance of FMPF, including 45 pairs of visible-near-infrared (VIS-NIR) [47] images, 44 pairs of visible-thermal infrared (VIS-TIR) [41] images, 40 pairs of optical–optical (OPT–OPT) [48] images and 24 pairs of depth map-RGB (DEPTH-RGB) [49] images.

1) Feature Matching: We first give the result of feature matching and motion vectors of our FMPF on eight multimodal image pairs in Fig. 8. These image pairs are chosen from the aforementioned datasets, and each dataset contains two pairs. It can be found that our PMFP method can remove most of the mismatches are removed and realize structural congruence from

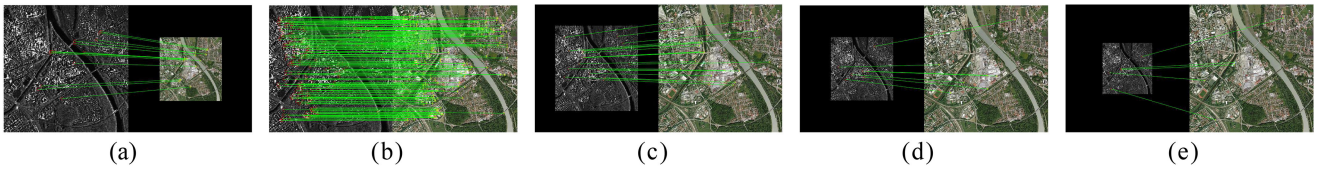

Fig. 7. Image matching results. (a) s = 0.5. (b) s = 1. (c) s = 1.5. (d) s = 2. (e) s = 2.5.

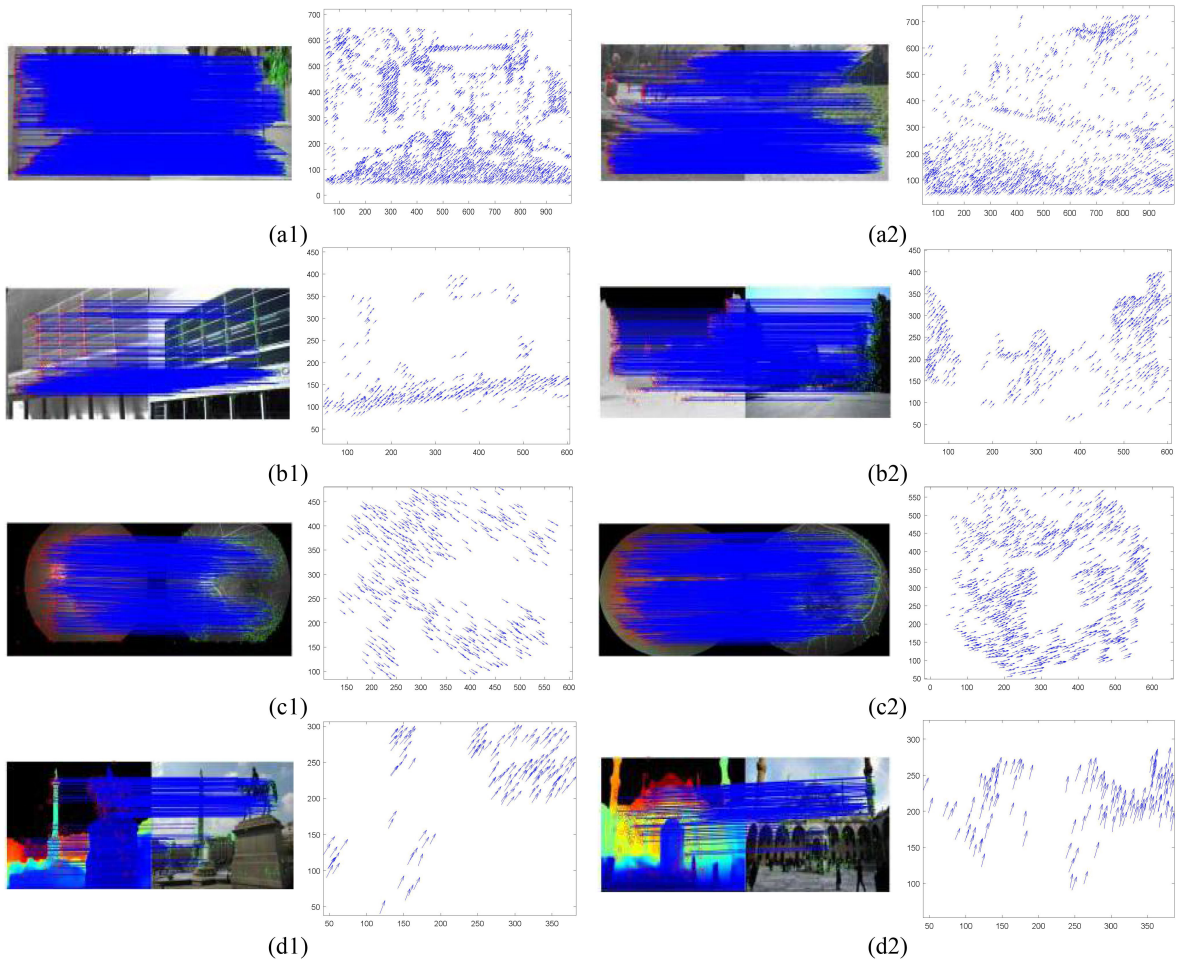

Fig. 8. Feature matching results (left) and motion vectors (right) on eight multimodal image pairs. (a1) VIS-NIR1. (a2) VIS-NIR2. (b1) VIS-TIR1. (b2) VIS-TIR2. (c1) OPT-OPT1. (c2) OPT–OPT2. (d1) DEPTH-RGB1. (d2) DEPTH-RGB2.

the motion vectors. By using (21)–(23), the precision, recall, and F-score on eight image pairs are (83.98%, 95.72%, 0.9847), (82.43%, 94.02%, 0.8785), (65.16%, 74.47%, 0.6951), (63.89%, 98.64%, 0.7755), (51.78%, 95.83%, 0.6723), (63.02%, 98.82%, 0.7697), (47.52%, 85.82%, 0.6117), and (41.70%, 88.89%, 0.5678). Although the precision of the last two image pairs is lower than 50%, the correctly matched motion vectors have structural congruence.

The feature matching performance of FMPF is quantitatively compared with four advanced feature matching methods, including VFC, LLT, LPM, and mTopKPR. The local features of the four comparison methods are detected in the nonlinear scale space, which is the same as FMPF. We calculate the precision, recall, and F-score on all MIDs, and then plot precision-recall and cumulative distribution curves of F-score, as shown in Fig. 9. The cumulative distribution curve indicates that 100×x% of the image pairs have performance values not greater than y. The larger the precision, recall, and F-score, the better the matching performance. The average precision, recall and F-score of FMPF are (73.55%, 95.50%, 0.8136), (58.13%, 76.83%, 0.6508), (54.97%, 88.96%, 0.6627), and (43.21%, 78.12%, 0.5286). The feature matching performances are characterized by the cumulative distribution of the F-score, and it can be noted that our FMPF method is superior to the other four advanced feature matching methods.

2) Image Registration: The key point of image registration is whether the transformed image can maximize the alignment of overlapping regions. We use the affine transformation model to transform the sensed image after feature matching. Fig. 10 shows the eight multimodal image pairs (left) and the visual

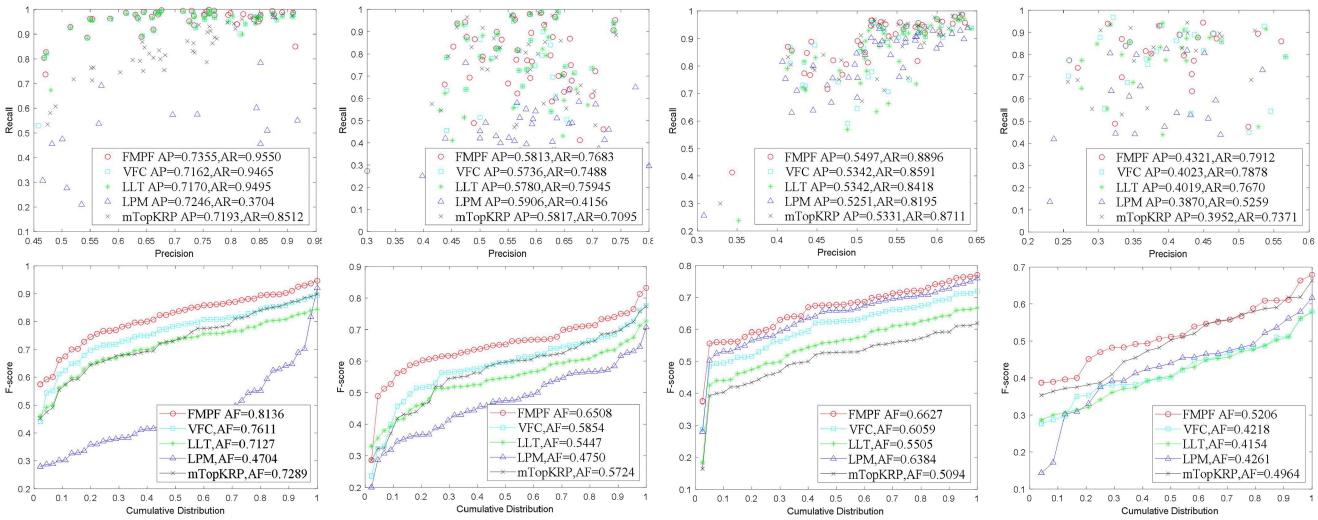

Fig. 9. Quantitative comparison of feature matching on four MIDs. (From left to right) VIS-NIR, VIS-TIR, OPT–OPT, and DEPTH-RGB. The average value of each comparison method is displayed in the legend (AP denotes average Precision, AR denotes average Recall, and AF denotes average F-score).

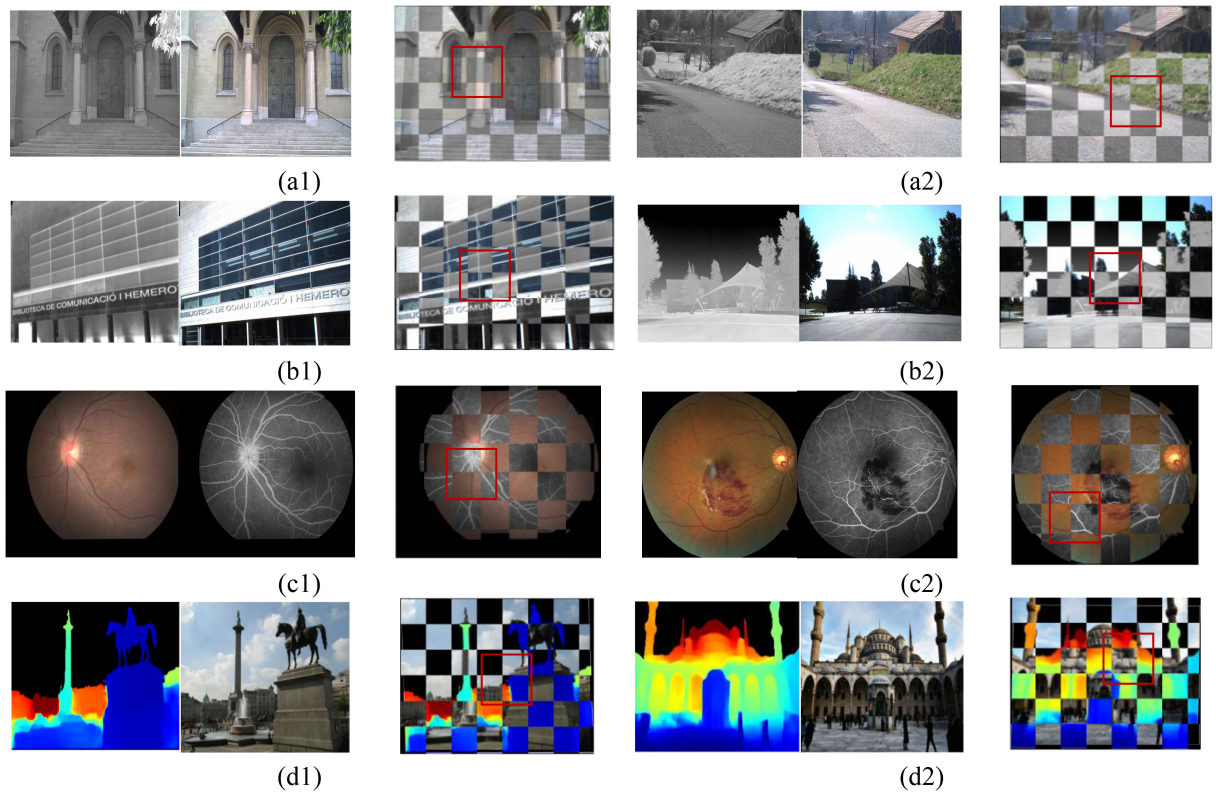

Fig. 10 Image pairs (left) and registration results (right) on eight multimodal image pairs. (a1) VIS-NIR1. (a2) VIS-NIR2. (b1) VIS-TIR1. (b2) VIS-TIR2. (c1) OPT–OPT1. (c2) OPT–OPT2. (d1) DEPTH-RGB1. (d2) DEPTH-RGB2.

registration results (right). It can be found that FMPF achieves satisfactory performance and obtain high registration accuracy.

Then, the registration accuracies of VFC, LLT, LPM, and mTopKPR are quantitatively compared and the cumulative distribution curves are plotted. The cumulative distribution curve indicates that 100×x% of the image pairs have performance values not greater than y (i.e., RMSE, MEAN, and Std). The less the RMSE, MEAN, and Std, the better registration performance. As shown in Fig. 11, the average RMSE, MEAN, and Std of mTopKPR are the largest, while the average RMSE, MEAN, and Std of FMPF are the smallest on four multimodal image datasets. It can be concluded that the accuracy of FMPF on RMSE, MEAN, and Std is higher than that of the other four algorithms.

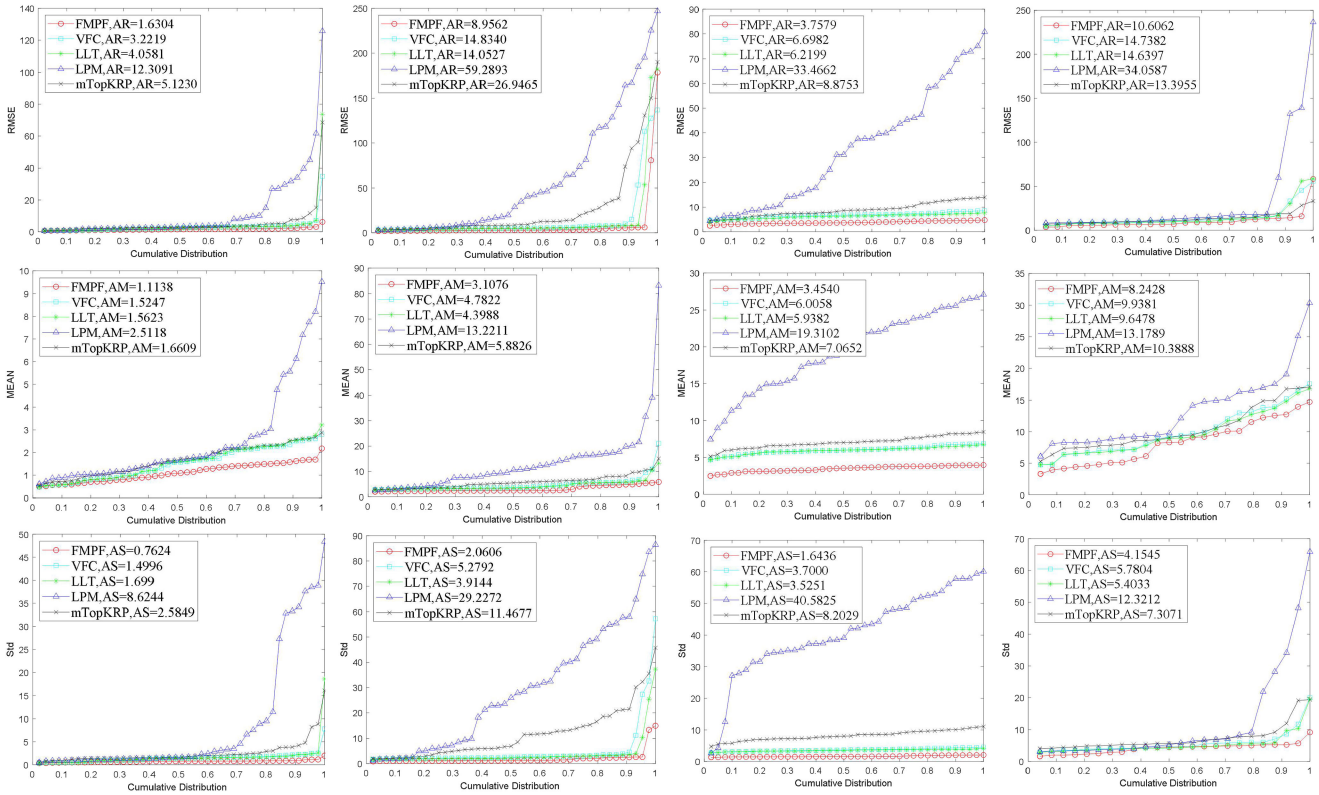

Fig. 11. Quantitative comparison of image registration on four MIDs. (From left to right) VIS-NIR, VIS-TIR, OPT–OPT, and DEPTH-RGB. (From top to bottom) RMSE, MEAN, and Std to the cumulative distribution. The average value of each comparison method is displayed in the legend (AM denotes average RMSE, AM denotes average MEAN, and AS denotes average Std).

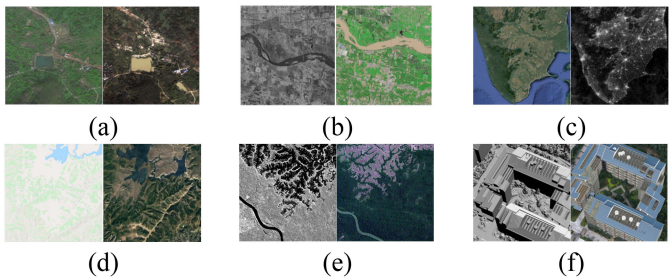

Fig. 12. MRSIPs. (a) OPT–OPT. (b) SAR-optical. (c) Day-night. (d) Map-optical. (e) Infrared-optical. (f) Depth-optical.

D. Experimental Results on MRSIPs

In Section III-C, all MIDs have no coordinate information. In this part, six types of multimodal remote sensing image pairs (MRSIPs) with geographic coordinate information are selected for testing. These MRSIPs include OPT–OPT images, SAR-optical images, day-night images, map-optical images, infrared-optical images, and depth-optical images, which consider almost all applications of multimodal remote sensing images (MRSI) matching, such as image fusion, image interpretation, and image registration. The six MRSIPs are displayed in Fig. 12. The size of MRSI ranges from 400×400 pixels to 1000×1000 pixels. We can find that the geometric distortions and radiation distortions of the MRSIPs are much more serious than those of MIDs. Therefore, it is essential to test our proposed method's robustness on multimodal remote sensing image matching. Qualitative and quantitative experiments are conducted to evaluate the accuracy of feature matching and image registration on these MRSIPs.

1) Feature Matching: Two descriptors (SIFT and RIFT) and four mismatch removal methods (VFC, LLT, LPM, and mTopKPR) are used to compare with the proposed FMPF method.

a) The qualitative results of feature matching are shown in Fig. 13. SIFT failed to match the MRSIPs in Fig. 13(a2), (a5), and (a6) while successfully matching in Fig. 13(a1), (a3), and (a4). SIFT performs blur processing on the MRSI using the Gaussian pyramid to construct the image scale space, which leads to the weakening of image texture edge features and the difficulty of extracting contour edge features. In summary, the traditional Gaussian pyramid scale space is not conducive to MRSI matching.

Although the LPM algorithm matches successfully, there are a small number of FPs in Fig. 13(e2), (e3), and (e4) (i.e., 2, 2, and 1, respectively). If we use FPs to transform the image, it will increase the error of image matching. Moreover, the average TP of the LPM algorithm is 38.83, which is less than that of RIFT, VFC, and FMPF. In summary, the LPM algorithm is not good for MRSI matching.

The matching results of VFC, LLT, mTopKPR, and FMPF are all successful, and the average TPs are 47.83, 32.67, 29.83, and 220, respectively. LLT has the least TPs on the OPT–OPT dataset, SAR-optical dataset, day-night dataset, and map-optical

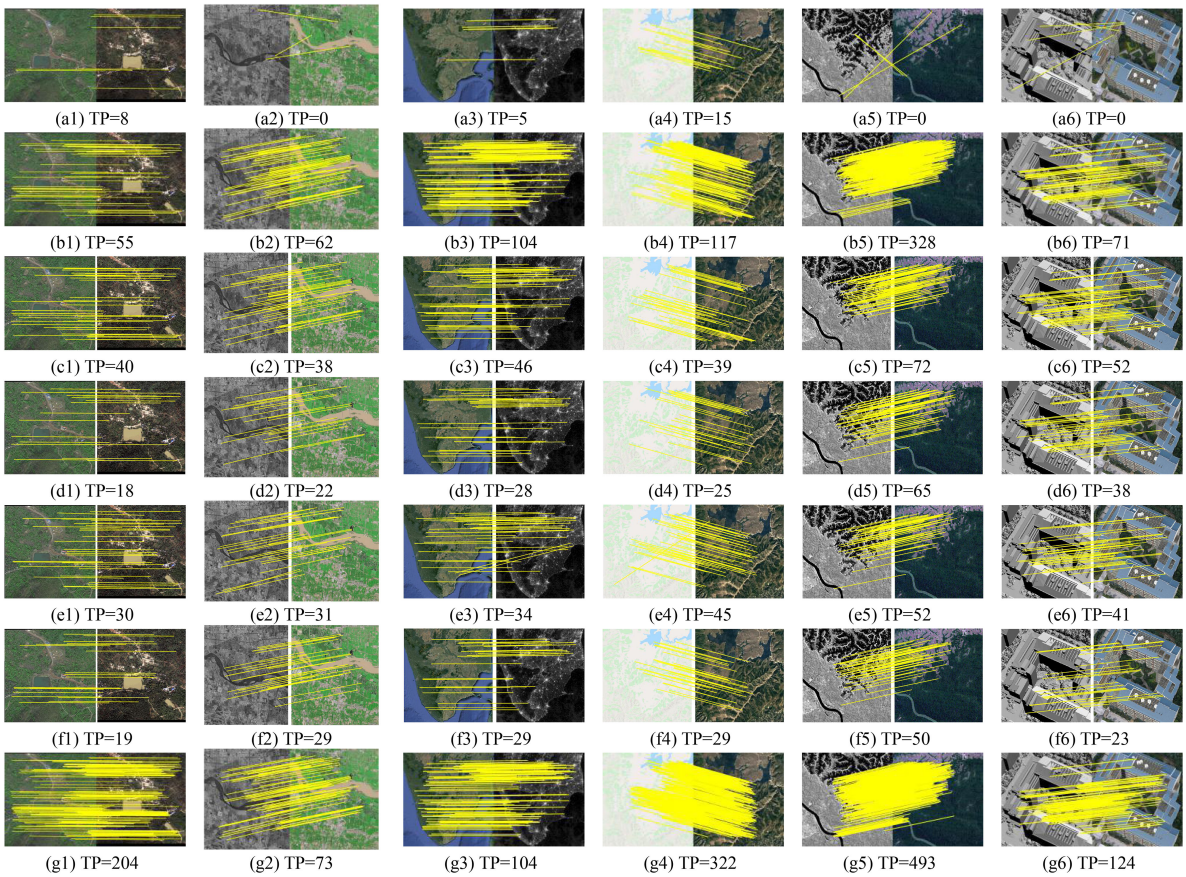

Fig. 13. Feature matching results on MRSIPs. (From left to right) OPT–OPT, SAR-optical, day-night, map-optical, infrared-optical, and depth-optical. (From top to bottom) SIFT, RIFT, VFC, LLT, LPM, mTopKPR, and FMPF.

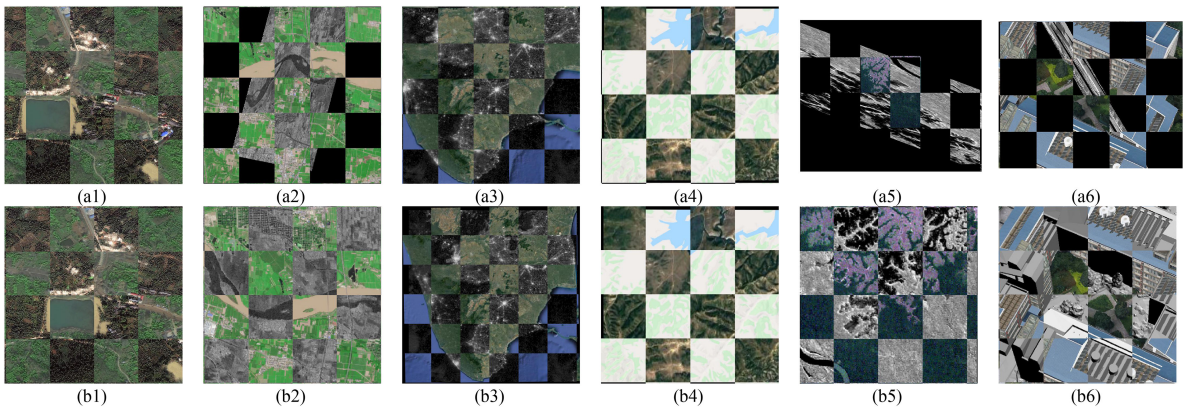

Fig. 14. Registration results on MRSIPs. (From left to right) OPT–OPT, SAR-optical, day-night, map-optical, infrared-optical, and depth-optical. (From top to bottom) SIFT and FMPF.

dataset, which are 18, 22, 28, and 25, respectively. The mTopKPR has the least TPs on the infrared-optical dataset and depth-optical dataset, which are 50 and 23. FMPF produces the best performance on the six MRSIPs, and the TPs are 204, 73, 104, 322, 439, and 124, respectively.

b) We use the F-score to quantitatively evaluate the matching performance, and the quantitative comparison results are shown in Tables I and II. SIFT performs better on the OPT–OPT dataset, the day-night dataset, and the map-optical dataset than on the other three datasets. In three successfully matched image pairs, all F-scores are very small (smaller than 0.02). It proves that SIFT is not suitable for matching multimodal remote sensing images.

The F-score of LPM on six data sets is much lower than that of the other four mismatch removal algorithms. The main reason may be that TP contains a small amount of NP, resulting in a

TABLE I
F-SCORE RESULTS OF SIFT, RIFT, AND FMPF ON SIX MRSIPS
("-" MEANS NO RESULT)

Method	Optical-Optical	SAR-Optical	Day-Night	Map-Optical	Infrared-Optical	Depth-Optical
SIFT	0.0039	-	0.0153	0.0099	-	-
RIFT	0.4828	0.6061	0.6289	0.6083	0.8073	0.6264
FMPF	**0.7626**	**0.7969**	**0.6667**	**0.8421**	**0.9265**	**0.8400**

TABLE II
F-SCORE RESULTS OF VFC, LLT, LPM, MTOPKPR, AND FMPF ON SIX MRSIPS

Method	Optical-Optical	SAR-Optical	Day-Night	Map-Optical	Infrared-Optical	Depth-Optical
VFC	0.4655	0.5432	0.5743	0.5864	0.6127	0.6000
LLT	0.4727	0.5449	0.5709	0.5831	0.6577	0.5991
LPM	0.0933	0.1899	0.1634	0.2063	0.2637	0.1897
mTopKRP	0.4662	0.5350	0.5497	0.5701	0.6546	0.5845
FMPF	**0.7626**	**0.7969**	**0.6667**	**0.8421**	**0.9265**	**0.8400**

TABLE III
RMSE RESULTS OF SIFT, RIFT, AND FMPF ON SIX MRSIPS
("-" MEANS NO RESULT)

Method	Optical-Optical	SAR-Optical	Day-Night	Map-Optical	Infrared-Optical	Depth-Optical
SIFT	8.9297	-	8.8726	6.1279	-	-
RIFT	5.4772	2.0371	2.3142	2.6212	1.8322	2.1994
FMPF	**2.1828**	**1.9624**	**2.0549**	**1.9959**	**1.7394**	**1.3532**

TABLE IV
RMSE RESULTS OF VFC, LLT, LPM, MTOPKPR, AND FMPF ON SIX MRSIPS

Method	Optical-Optical	SAR-Optical	Day-Night	Map-Optical	Infrared-Optical	Depth-Optical
VFC	6.1142	5.7167	5.4114	5.2388	3.1222	6.2768
LLT	6.6045	5.7167	5.2341	4.9859	3.0689	5.1857
LPM	8.4447	2.1213	3.2787	3.7920	1.7993	3.1623
mTopKRP	5.5297	2.2434	2.4138	2.9468	1.8632	2.5768
FMPF	**2.1828**	**1.9624**	**2.0549**	**1.9959**	**1.7394**	**1.3532**

significant decrease in the F-score. We can clearly observe that the F-score of FMPF reaches the maximum, which demonstrates its robustness and effectiveness on MRSIPs.

2) Image Registration: The affine transformation model is utilized to transform the sensed image. We compare the registration results of SIFT and FMPF, as shown in Fig.14. Since the TP of Fig. 13(a2), (a5), and (a6) is 0, there are no true matches to compute the homography model, resulting in the failure of image registration of Fig. 14(a2), (a5), and (a6). In Fig. 14(a1), (a3), and (a4), SIFT performed better registration performance on the OPT–OPT dataset, the day-night dataset, and the map-optical dataset because of its resistance to illumination changes. In Fig. 14(b1)–(b6), all the registered image has achieved a good alignment effect with our FMPF method, and the registration performance is better than SIFT.

We use the affine transformation model to transform the sensed image after feature matching. We select 20 pairs of landmarks manually as the truth values, they are uniformly distributed around the region of interest of each multimodal image pair. The RMSE of the transformation residual error is computed by the homography model. The less the RMSE, the better the registration performance of corresponding points. The RMSE of each method of MRSIPs is shown in Tables III and IV. As can be seen, the proposed FMPF method achieves the best performance followed by RIFT. SIFT obtains the worst RMSE performance, which proves that it is not suitable for multimodal image registration. The average RMSE of VFC, LLT, LPM, and mTopKPR is 122.83, 47.83, 32.67, 29.83, and 220, respectively. LPM and mTopKPR have competitive performance because they can maintain reliable matches, which can estimate the transformation correctly. As for VFC and LLT, some obvious false correspondences may be preserved, resulting in relatively poor registration performance.

IV. CONCLUSION

We propose a feature matching method for multimodal images based on progressive filtering in this article. We conduct a series of experiments on MIDs and MRSIPs, the results show that our proposed FMPF method outperforms the other six advanced feature matching methods. However, the proposed FMPF method is tested only on rigid multimodal images. For non-rigid multimodal images, there may be only a small number of correct matches. The locations of the inliers can be very scattered, so the initial correspondences may not have structural consistency. We plan to use different feature detection and description methods to create more effective correspondences on nonrigid multimodal images in the feature work. In the feature matching stage, there are four empirical thresholds τ_1–τ_4. We use an iterative strategy to remove the mismatches progressively, which is similar to deterministic annealing. We use the same threshold for the MDIs and MRSIPs. We plan to use the dynamic adaptive threshold to remove the mismatches for different datasets in the feature work.

REFERENCES

[1] X. Chang, S. Du, Y. Li, and S. Fang, "A coarse-to-fine geometric scale-invariant feature transform for large size high resolution satellite image registration," *Sensors*, vol. 18, pp. 1–16, Apr. 2018.
[2] X. Lu, H. Ma, and B. Zhang, "A non-rigid medical image registration method based on improved linear elastic model," *Optik*, vol. 123, pp. 1867–1873, Mar. 2012.
[3] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "CoSpace: Common subspace learning from hyperspectral-multispectral correspondences," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4349–4359, Jul. 2019.
[4] L. Li et al., "Artificial intelligence test: A case study of intelligent vehicles," *Artif. Int. Rev.,* vol. 50, no. 3, pp. 441–465, Apr. 2018.
[5] S. Cao, X. Zhu, Y. Pan, and Q. Yu, "A stable land cover patches method for automatic registration of multitemporal remote sensing images," *IEEE J. Sel. Appl. Earth Observ. Remote Sens.*, vol. 7, no. 8, pp. 3502–3512, Aug. 2014.
[6] 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, no. 12, pp. 3296–3310, Dec. 2020.
[7] C. Han, W. Luo, H. Guo, and Y. Ding, "An image matching method for SAR orthophotos from adjacent orbits in large area based on SAR-moravec," *Remote Sens,* vol. 12, Sep. 2020, Art. no. 2892.
[8] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988, pp. 10–5244.
[9] L. Rasmy, I. Sebari, and M. Ettarid, "Automatic sub-pixel co-registration of remote sensing images using phase correlation and harris detector," *Remote Sens,* vol. 13, Jun. 2021, Art. no. 2314.
[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Jan. 2004.
[11] Q. Zeng, J. Adu, J. Liu, J. Yang, Y. Xu, and M. Gong, "Real-time adaptive visible and infrared image registration based on morphological gradient and C_SIFT," *J. Real-Time Image Process.*, vol. 17, pp. 1103–1115, Mar. 2020.

[12] M. Li et al., "Electronic image stabilization algorithm based on PCA-SIFT feature matching and self-adaptive high-pass filtering," in *Proc. Image Process. Conf.*, 2014, pp. 1–7.
[13] H. Fu et al., "An improved ASIFT algorithm for indoor panorama image matching," in *Proc. Image Process. Conf.*, 2017, pp. 1–9.
[14] Z. Ghassabi, J. Shanbehzadeh, A. Sedaghat, and E. Fatemizadeh, "An efficient approach for robust multimodal retinal image registration based on UR-SIFT features and PIIFD descriptors," *EURASIP J. Imag. Video Process.*, vol. 25, no. 2013, pp. 1–16, Jan. 2013.
[15] C. Dubois, A. Nascetti, A. Thiele, M. Crespi, and S. Hinz, "SAR-SIFT for matching multiple SAR images and radargrammetry," *PFG – J. Photogrammetry, Remote Sens. Geoinf. Sci.*, vol. 85, pp. 149–158, Jul. 2017.
[16] A. Sedaghat and H. Ebadi, "Remote sensing image matching based on adaptive binning sift descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5283–5293, Oct. 2015.
[17] Y. Ye, J. Shan, S. Hao, L. Bruzzone, and Y. Qin, "A local phase based invariant feature for remote sensing image matching," *ISPRS J. Photogrammetry Remote Sens.*, vol. 142, pp. 205–221, Aug. 2018.
[18] S. Kao and C. Ho, "Monitoring a process of exponentially distributed characteristics through minimizing the sum of the squared differences," *Qual. Quantity*, vol. 41, pp. 137–149, Feb. 2007.
[19] B. Wu and C. F. Hung, "Innovative correlation coefficient measurement with fuzzy data," *Mathematic Problems Eng.*, vol. 2016, pp. 1–11, Apr. 2016.
[20] Y. Hel-Or, H. Hel-Or, and E. David, "Matching by tone mapping: Photometric invariant template matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 317–330, Feb. 2014.
[21] 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.
[22] S. Li, K. Han, T. W. Costain, H. Howard-Jenkins, and V. Prisacariu, "Correspondence networks with adaptive neighborhood consensus," in *Proc. IEEE/CVF Eur. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10193–10202.
[23] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, "LF-Net: Learning local features from images," in *Proc. Conf. Neural Inf. Process. Syst.*, 2018, pp. 6237–6247.
[24] L. Zhou, Y. Ye, T. Tang, and Y. Qin, "Robust matching for SAR and optical images using multiscale convolutional gradient features," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4017605, doi: 10.1109/LGRS.2021.3105567.
[25] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li, and S. Hao, "A multiscale framework with unsupervised learning for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622215, doi: 10.1109/TGRS.2022.3167644.
[26] X. Liu, J. Li, and J. Pan, "Feature point matching based on distinct wavelength phase congruency and log-gabor filters in infrared and visible images," *Sensors*, vol. 19, Sep. 2019, Art. no. 4244.
[27] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, May 2017.
[28] Y. Ye, "Fast and robust registration of multimodal remote sensing images via dense orientated gradient feature," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 5141–5144.
[29] J. Li, W. Xu, P. Shi, Y. Zhang, and Q. Hu, "LNIFT: Locally normalized image for rotation invariant multimodal feature matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art. no. 5621314, doi: 10.1109/TGRS.2022.3165940.
[30] J. Li, Q. Hu, and M. Ai, "LAM: Locality affine-invariant feature matching," *ISPRS J. Photogrammetry Remote Sens.*, vol. 154, pp. 28–49, Jun. 2019.
[31] M. L. Uss, B. Vozel, S. K. Abramov, and K. Chehdi, "Selection of a similarity measure combination for a wide range of multimodal image registration cases," *IEEE Trans. Geosci. Remote Sens.,* vol. 59, no. 1, pp. 60–75, Jan. 2021.
[32] C. Aguilera, F. Barrera, F. Lumbreras, A. D. Sappa, and R. Toledo, "Multispectral image feature points," *Sensors,* vol. 12, pp. 12661–12672, Sep. 2012.
[33] J. Chen, J. Tian, N. Lee, J. Zheng, R. T. Smith, and A. F. Laine, "A partial intensity invariant feature descriptor for multimodal retinal image registration," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 6, pp. 1707–1718, Jun. 2010.
[34] 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.
[35] Y. Xiang, F. Wang, L. Wan, and H. You, "SAR-PC: Edge detection in SAR images via an advanced phase congruency model," *Remote Sens.*, vol. 9, Feb. 2017, Art. no. 209.
[36] Y. Ye, J. Shan, S. Hao, L. Bruzzone, and Y. Qin, "A local phase-based invariant feature for remote sensing image matching," *ISPRS J. Photogrammetry Remote Sens.*, vol. 142, pp. 205–221, Aug. 2018.
[37] J. Fan, Y. Wu, M. Li, W. Liang, and Y. Cao, "SAR and optical image registration using nonlinear diffusion and phase congruency structural descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5368–5379, Sep. 2018.
[38] M. Sun, T. Ma, Y. Song, and J. Peng, "Automatic registration of optical and SAR remote sensing image based on phase feature," *Opt. Precis. Eng.*, vol. 29, pp. 616–627, Sep. 2021.
[39] M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2091–2106, Dec. 2005.
[40] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE features," *Lecture Notes Comput. Sci.*, vol. 7577, pp. 214–227, 2012.
[41] C. A. Aguilera and A. D. Sappa, "LGHD: A feature descriptor for matching across non-linear intensity variations," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 178–181.
[42] X. Jiang et al., "Robust feature matching for remote sensing image registration via linear adaptive filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1577–1591, Feb. 2021.
[43] M. C. Morrone and R. A. Owens, "Feature detection from local energy," *Pattern Recognit. Lett.*, vol. 6, pp. 303–313, Dec. 1987.
[44] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015.
[45] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *Int. J. Comput. Vis.*, vol. 127, pp. 512–531, May 2019.
[46] X. Jiang, J. Jiang, A. Fan, Z. Wang, and J. Ma, "Multiscale locality and rank preservation for robust feature matching of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6462–6472, Sep. 2019.
[47] D. Firmenichy and M. Brown, "Multispectral interest points for RGB-NIR image registration," in *Proc. IEEE Int. Conf. Image Process*., 2011, pp. 181–184.
[48] A. Hoover and M. Goldbaum, "Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels," *IEEE Trans. Med. Imag.*, vol. 22, no. 7, pp. 951–958, Aug. 2003.
[49] W. Jing, S. Huo, Q. Miao, and X. Chen, "A model of parallel mosaicking for massive remote sensing images based on spark," *IEEE Access,* vol. 5, pp. 18229–18237, 2017.

Qiang Xiong received the B.S. degree in surveying and mapping engineering from Henan Polytechnic University, Jiaozuo, China, in 2019. He is currently working toward the Ph.D. degree in photogrammetry and remote sensing with Wuhan University, Wuhan, China.

His current research interests include image matching and registration.

Shenghui Fang received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2007.

His current research interests include remote sensing image processing and quantitative remote sensing.

Yi Peng received the Ph.D. degree in natural resource sciences from the School of Natural Resources and Environment, University of Nebraska, Lincoln, NE, USA, in 2012.

Her current research interest includes quantitative remote sensing.

Yan Gong received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2007.

His current research interests include hyperspectral remote sensing and agricultural remote sensing.

Xiaojuan Liu received the B.S. degree in surveying and mapping engineering in 2021 from Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree in resource and environmental sciences from Wuhan University, Wuhan, China.

Her current research interest includes quantitative remote sensing.