RIFT: Multi-modal Image Matching Based on Radiation-variation Insensitive Feature Transform

Jiayuan Li, Qingwu Hu, Mingyao Ai

***Abstract*—Traditional feature matching methods, such as scale-invariant feature transform (SIFT), usually use image intensity or gradient information to detect and describe feature points; however, both intensity and gradient are sensitive to nonlinear radiation distortions (NRD). To solve this problem, this paper proposes a novel feature matching algorithm that is robust to large NRD. The proposed method is called radiation-variation insensitive feature transform (RIFT). There are three main contributions in RIFT. First, RIFT uses phase congruency (PC) instead of image intensity for feature point detection. RIFT considers both the number and repeatability of feature points and detects both corner points and edge points on the PC map. Second, RIFT originally proposes a maximum index map (MIM) for feature description. The MIM is constructed from the log-Gabor convolution sequence and is much more robust to NRD than traditional gradient map. Thus, RIFT not only largely improves the stability of feature detection but also overcomes the limitation of gradient information for feature description. Third, RIFT analyses the inherent influence of rotations on the values of the MIM and realises rotation invariance. We use six different types of multi-modal image datasets to evaluate RIFT, including optical-optical, infrared-optical, synthetic aperture radar (SAR)-optical, depth-optical, map-optical, and day-night datasets. Experimental results show that RIFT is superior to SIFT and SAR-SIFT on multi-modal images. To the best of our knowledge, RIFT is the first feature matching algorithm that can achieve good performance on all the abovementioned types of multi-modal images. The source code of RIFT and the multi-modal image datasets are publicly available[1].**

I. Introduction

Image feature matching is a fundamental and crucial issue in photogrammetry and remote sensing, whose goal is to extract reliable feature correspondences from two or more images with overlapping regions [1]. It is also widely used in the fields of computer vision [2, 3], robot vision [4, 5], medical image analysis [6] and so on. Image matching has always been a hot research issue and has made great progress in the past decades. However, image matching, especially remote sensing image matching, is still an ill-posed problem that suffers from many uncertainties. In general, a remote sensing image pair may contain scale, rotation, radiance, noise, blur, or temporal changes [7]. These huge differences in geometry and radiation pose a daunting challenge to the current image matching algorithms, which result in a substantial reduction in matching performance and make it difficult to meet the requirements of ever-changing practical applications. Therefore, it is very important to study more effective, universal, and robust image matching algorithms. To achieve universal robust image matching, three difficult problems need to be solved: (1) the robustness to various geometric and radiation distortions; (2) more efficient and robust outlier detection models; (3) the non-rigid deformation image matching problem. In our past work, we have studied geometric distortions [8], non-rigid matching [7, 9], and outlier removal [10-12]. In this paper, we will focus on radiation distortions, especially nonlinear radiation distortions (NRD).

Radiation distortions refer to the phenomenon in which the spectral emissivity of the ground objects is different from the real spectral emissivity in the process of the sensor imaging [13]. The factors that cause radiation distortions are various and can be summarised as two aspects [14]: (1) the imaging property of the sensor itself. This type of error can be regarded as systematic error. Images acquired by the same sensor often have the same systematic error and thus have little effect on the classical image matching algorithms. However, with the diversification of sensors and applications, it is often necessary to fuse the superiority information of different sensors and finally achieve a more accurate and reliable description of the scene. The imaging mechanism of different sensors may be quite different, and the acquired images have different expressions for the same objects, which result in large radiation differences between image pairs. Classical feature matching methods are still able to address linear radiation differences; unfortunately, for NRD, these methods may not work. Generally, the images with large NRD are called multi-modal images. Traditional methods usually use intensity information or gradient information for feature detection and description. However, both image intensity and gradient are very sensitive to NRD. It can be said that the processing of multi-modal images is a bottleneck problem of image matching. At present, if the geometrical geographic information of images is unavailable, no image matching method can be simultaneously applied to optical-optical matching, infrared-optical matching, synthetic aperture radar (SAR)-optical matching, depth-optical matching, map-optical matching, and day-night matching. (2) The radiation transmission error caused by the atmosphere. In the process of electromagnetic wave transmission, the spectral emissivity of ground objects may be distorted by the influence of atmospheric action, solar altitude angle, and illumination conditions. These types of distortions are especially prominent in multi-temporal remote sensing images, which often appear as “different objects with same spectrum” or “same object with different spectra” phenomenon. Such nonlinear differences will reduce the correlation between correspondences, which often results in difficulties in matching. Because multi-modal and multi-temporal remote sensing image data play an important role in target detection, disaster assessment, illegal building detection, and land resource change monitoring, it is imperative to study image matching methods against large radiation differences, especially large NRD.

In this paper, we propose a radiation insensitive feature matching method based on phase congruency (PC) and a maximum index map (MIM), which is called radiation-variation insensitive feature transform (RIFT). First, we detect corner feature points and edge feature points on the

[1] http://www.escience.cn/people/lijiayuan/index.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2959244, IEEE Transactions on Image Processing

PC map. We find that the corner feature usually has better repeatability and there is a higher number of edge features. Thus, the combination of corner features and edge features can improve the stability of feature detection. Then, we construct a MIM based on the log-Gabor convolution sequence, which is much more robust to NRD than traditional gradient maps. Thus, the MIM is very suitable for multi-modal image feature description. However, MIM is very sensitive to rotations. Different rotation angles may result in different MIMs. We analyse the inherent influence of rotations on the values of MIMs and achieve rotation-invariant by the construction of multiple MIMs. Experiments show that RIFT is far superior to invariant feature transform (SIFT) [15] and SAR-SIFT [16] on multi-modal images. To the best of our knowledge, RIFT is the first feature matching algorithm that can achieve good performance on all different types of multi-modal images.

II. Related work

Image matching is one of the most important steps in the automatic production of photogrammetry and remote sensing products, and its results directly affect the applications of image stitching, bundle adjustment, and 3D reconstruction. Reference [1] provides a very systematic summary and classification of traditional image matching methods. According to this summary, image matching methods can be divided into two categories: area-based matching methods and feature-based matching methods.

A. *Area-based Methods*

Area-based matching methods use the original pixel values and specific similarity measures to find the matching relationship between image pairs. Usually, a predefined local window or global image is used to search for matching, and no feature extraction stage is required [17].

One of the drawbacks of area-based methods is that they are only suitable for image pairs containing translation changes. For image pairs containing rotation changes, a circular window is needed to perform the correlation. However, if the image pairs contain complex variations such as rotations, scale changes, geometric distortions, etc., these methods will fail.

Another drawback is that the content inside the local window is not salient. The image content inside the search window may be relatively smooth and lack salient features. If the search window is located in a weakly textured or non-textured area of the image, then this method is likely to get an incorrect match. Therefore, the window selection should be based on the image content, and the portion containing more salient features should be selected as the search window content.

Area-based methods can be roughly divided into three sub-categories, including correlation-based methods [18, 19], Fourier-based methods [20, 21], and mutual information-based methods [22, 23].

B. *Feature-based Methods*

Different from area-based methods, feature-based methods are not directly based on image intensity information. These methods usually first detect salient structural features in the image, which can be point features (corners, line intersections, etc.), line features (lines, contours, etc.), or region features. These features must be salient, easily detectable, and stable. That is, regardless of the effect of image geometric distortions and radiation distortions, there are always enough identical elements in the two feature sets. In the following, we will only review point feature matching methods because point feature is the simplest and the basis of other features.

Features can better describe the structure information of an image, thus reducing the sensitivity to the sensor noise and scene noise. Feature-based methods are generally more robust than area-based methods. In the field of computer vision, the SIFT [15] is one of the most commonly used and effective feature point matching methods. It first constructs a Gaussian scale space and extracts feature points in the scale space. Then, SIFT uses a gradient histogram technique to describe features. Speeded-up robust features (SURF) [24] extracts feature points based on the Hessian matrix and introduces an integration graph technique to improve efficiency. Principal component analysis SIFT (PCA-SIFT) [25] adopts the principal component analysis algorithm to reduce the dimensionality of the SIFT descriptor and extracts the dimensions with larger values. PCA-SIFT greatly reduces the complexity of the original SIFT. Affine-SIFT (ASIFT) [26] extends SIFT to be invariant to affine transformations by simulating two camera axis direction parameters. The ORB (Oriented FAST and Rotated BRIEF) [27] algorithm uses the features from accelerated segment test (FAST) [28] to extract feature points and utilises the directional binary robust independent elementary features (BRIEF) [29] algorithm for feature description. This method has low time complexity and is suitable for real-time applications, but it is not scale-invariant. In the field of photogrammetry and remote sensing, the SIFT algorithm has also been widely adopted due to its robustness to illumination, rotation, scale, noise, and viewpoint changes. However, because remote sensing images are captured by different sensors, at different times, and with different viewpoints, there are large geometric distortions and radiation distortions between remote sensing image pairs. To solve this problem, scholars have proposed many improved algorithms. Uniform robust SIFT (UR-SIFT) [30] studies the distribution of SIFT feature points and proposes an entropy-based feature point selection strategy to improve the distribution uniformity. Uniform competency (UC) detector [31] proposes a novel competency criterion based on a weighted ranking process, which considers the robustness, scales and spatial saliency of a feature. Hence, it has better location distribution and matching quality than Harris [32] detector, SIFT detector, SURF detector and maximally stable extremal region (MSER) [33] detector. SAR-SIFT [16] introduces a new gradient definition based on the specific characteristics of SAR images to improve the robustness to speckle noise. Adaptive binning SIFT (AB-SIFT) [34] adopts an adaptive binning gradient histogram to describe feature points, making it more resistant to local geometric distortions. Sedaghat and Mohammadi [35] combined an improved SURF detector and AB-SIFT for feature matching, and presented a localized graph transformation for outlier removal. Ye et al. [36] proposed a feature detector (MMPC-Lap) and a feature descriptor named local histogram of orientated phase congruency (LHOPC) for multisensor image matching, which is robust to illumination and contrast variations. However, these feature-based methods are sensitive to NRD and are not

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2959244, IEEE Transactions on Image Processing

suitable for multimodal images, such as SAR-optical, depth-optical, map-optical, etc.

This paper aims to solve the problem of radiation distortions in image matching, especially the problem of NRD. Multi-modal images are typical images with NRD. At present, the research of multi-modal image matching mostly focuses on medical images, and there are few studies that address the processing of multimode remote sensing images. However, multi-modal remote sensing image matching has very important theoretical and practical significance. Theoretically, this problem is very difficult, and it is a bottleneck problem of image matching technology. In fact, many applications require automatic matching of multi-modal images, such as information fusion of optical and SAR images. In the next section, we will briefly review the state-of-the-art of multi-modal image matching.

C. Multi-Modal Image Matching

Recently, the multi-modal image matching task has drawn increasingly more attention, and several algorithms have been proposed. For example, local self-similarity descriptor (LSS) [37], partial intensity invariant feature descriptor (PIIFD) [38], distinctive order-based self-similarity descriptor (DOBSS) [39], ARRSI [40], histogram of orientated phase congruency (HOPC) [41], and phase congruency structural descriptor (PCSD) [42]. Among them, ARRSI and HOPC are the most similar to the proposed RIFT.

ARRSI detects feature points and performs normalised cross-correlation (NCC) [43] matching on the maximum moment map of PC. Although both ARRSI and the proposed RIFT use a PC measure for feature detection, RIFT is quite different from ARRSI. First, ARRSI does not construct a feature vector and uses NCC to search matches, which is essentially a template matching method, while RIFT is a feature matching method. Second, RIFT originally proposed a MIM measure for feature description. Third, RIFT is invariant to rotation changes while ARPSI is not.

HOPC extends the PC model to include not only numerical information but also corresponding orientation information. Then, a modified similarity measure, HOPCncc, is presented based on the improved PC measure and NCC. Unfortunately, HOPC suffers from three major problems:

(1) HOPC needs to know the exact geographic information corresponding to the image to perform geometric correction. In practice, however, the geographical information of an image may not be accurate enough or may be unavailable. For example, the geographical information of a satellite image may be hundreds of metres away from its actual geographical location. In such cases, HOPC is completely unusable.

(2) Although HOPC performs feature detection on the reference image, it is essentially a template matching method, rather than a feature-based method, and therefore is sensitive to rotation, scale, etc. Template matching methods usually perform a two-dimensional search, which becomes a one-dimensional search after adding the epipolar constraint. HOPC relies on accurate geographic information, whose search space is small, usually a local window of 20×20 pixels.

(3) HOPC uses Harris detector to detect the feature points. However, Harris is very sensitive to NRD, and it is difficult to be universally suitable for different types of multi-modal images. Especially, when a depth map is used as the reference image, the performance of Harris becomes very poor. Feature detection is the basis of feature matching method, which determines the number of correct matches between two sets of points and point location accuracy. If the number of features is too small, the matching result must be very poor.

In contrast, RIFT does not rely on geographic information. RIFT has good robustness to NRD, regardless of the feature detection or feature description stage, and achieves rotation invariance.

III. Radiation-variation Insensitive Feature Transform (RIFT)

In this section, we will detail the proposed RIFT method, including feature detection and feature description stages.

A. Feature Detection

Classical feature matching methods generally rely on image intensity or gradient information, which is spatial domain information. In addition to spatial domain information, images can also be described using frequency domain information, such as phase information. The theoretical basis of phase is the Fourier transform (FT) theorem [44]. FT can decompose an image into an amplitude component and a phase component. Oppenheim and Lim [45], for the first time, revealed the importance of phase information for the preservation of image features. Phase information has high invariance to image contrast, illumination, scale, and other changes. Further, Morrone and Owens [46] discovered that certain points in the image could cause a strong response on the human visual system, and these points usually have highly consistent local phase information. Hence, the degree of consistency of local phase information at different angles is called the PC measure.

1) Recall on Log-Gabor Filter

The 2D log-Gabor filter (2D-LGF) [47, 48] can generally be obtained by Gaussian spreading of the vertical direction of the log-Gabor filter (LGF). Thus, the 2D-LGF function is defined as follows,

$$L(\rho,\theta,s,o)=\exp(\frac{-(\rho-\rho_s)^2}{2\sigma_\rho^2})\exp(\frac{-(\theta-\theta_{so})^2}{2\sigma_\theta^2}) \qquad (1)$$

where (ρ,θ) represents the log-polar coordinates; s and o are the scale and orientation of 2D-LGF, respectively; (ρ_s,θ_{so}) is the centre frequency of 2D-LGF; σ_ρ and σ_θ are the bandwidths in ρ and θ, respectively.

LGF is a frequency domain filter, whose corresponding spatial domain filter can be obtained by inverse Fourier transform. In the spatial domain, 2D-LGF can be represented as [41, 49],

$$L(x,y,s,o)=L^{even}(x,y,s,o)+iL^{odd}(x,y,s,o) \qquad (2)$$

where the real part $L^{even}(x,y,s,o)$ and the imaginary part $L^{odd}(x,y,s,o)$ stand for the even-symmetric and the odd-symmetric log-Gabor wavelets, respectively.

2) Recall on Phase Congruency (PC)

Let $I(x,y)$ denote a 2D image signal. Convolving the image $I(x,y)$ with the even-symmetric and the odd-symmetric wavelets yields the response components $E_{so}(x,y)$ and $O_{so}(x,y)$,

$$\begin{aligned}&[E_{so}(x,y),O_{so}(x,y)]\\&=[I(x,y)*L^{even}(x,y,s,o),I(x,y)*L^{odd}(x,y,s,o)]\end{aligned} \quad (3)$$

Then, the amplitude component $A_{so}(x,y)$ and the phase component $\phi_{so}(x,y)$ of $I(x,y)$ at scale s and orientation o can be obtained by,

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

$$\phi_{so}(x,y)=\arctan(O_{so}(x,y)/E_{so}(x,y)) \quad (5)$$

Considering the analysis results in all directions and all orientations, and introducing the noise compensation term T, the final 2D PC model is (more details about the 2D PC model can be found in Reference [50]):

$$PC(x,y)=\frac{\sum_s\sum_o w_o(x,y)\lfloor A_{so}(x,y)\Delta\Phi_{so}(x,y)-T\rfloor}{\sum_s\sum_o A_{so}(x,y)+\xi} \quad (6)$$

where $w_o(x,y)$ is a weighting function; ξ is a small value; $\lfloor\cdot\rfloor$ operator prevents the enclosed quantity from getting a negative value; that is, it takes zero when the enclosed quantity is negative. $\Delta\Phi_{so}(x,y)$ is a phase deviation function, whose definition is,

$$\begin{aligned}A_{so}(x,y)\Delta\Phi_{so}(x,y)&=(E_{so}(x,y)\bar{\phi}_E(x,y)+O_{so}(x,y)\bar{\phi}_O(x,y))\\&-|(E_{so}(x,y)\bar{\phi}_O(x,y)-O_{so}(x,y)\bar{\phi}_E(x,y))|\end{aligned} \quad (7)$$

where,

$$\bar{\phi}_E(x,y)=\sum_s\sum_o E_{so}(x,y)/C(x,y) \quad (8)$$

$$\bar{\phi}_O(x,y)=\sum_s\sum_o O_{so}(x,y)/C(x,y) \quad (9)$$

$$C(x,y)=\sqrt{(\sum_s\sum_o E_{so}(x,y))^2+(\sum_s\sum_o O_{so}(x,y))^2} \quad (10)$$

3) Corner and Edge Features

Several image matching methods [51-53] have adapted the PC measure for feature detection. Differently, RIFT combines corner features and edge features. In addition, these methods generally use SIFT or a local binary pattern (LBP) for feature description, which are very sensitive to NRD and not suitable for multi-modal images.

Based on equation (6), we can obtain a very precise edge map, i.e., the PC map. However, this formula ignores the effect of orientation changes on the PC measure [54]. To get the relations between the PC measure and orientation changes, we compute an independent PC map $PC(\theta_o)$ for each orientation o, where θ_o is the angle of orientation o. We then calculate the moments of these PC maps and analyse the moment changes with the orientation

According to the moment analysis algorithm [55], the axis corresponding to the minimum moment is called the principal axis, and the principal axis usually indicates the direction information of the feature; the axis corresponding to the maximum moment is perpendicular to the principal axis, and the magnitude of the maximum moment generally reflects the distinctiveness of the feature. Before calculation of the minimum and maximum moments, we first compute three intermediate quantities,

$$a=\sum_o(PC(\theta_o)\cos(\theta_o))^2 \quad (11)$$

$$b=2\sum_o(PC(\theta_o)\cos(\theta_o))(PC(\theta_o)\sin(\theta_o)) \quad (12)$$

$$c=\sum_o(PC(\theta_o)\sin(\theta_o))^2 \quad (13)$$

Then, the principal axis ψ, minimum moment M_ψ, and maximum moment m_ψ are given by,

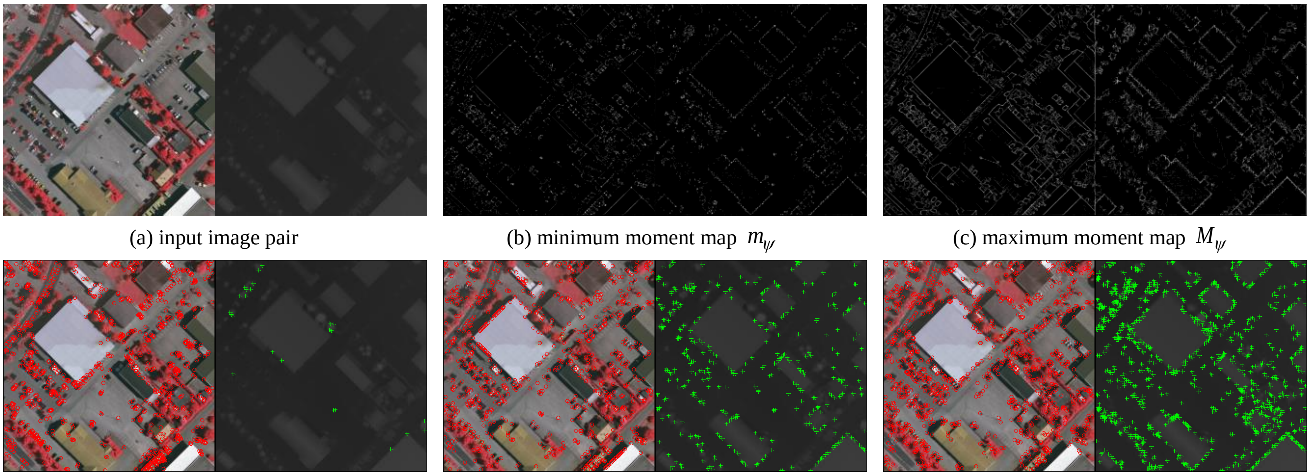
$$\psi=\frac{1}{2}\arctan(\frac{b}{a-c}) \quad (14)$$

$$M_\psi=\frac{1}{2}(c+a+\sqrt{b^2+(a-c)^2}) \quad (15)$$

$$m_\psi=\frac{1}{2}(c+a-\sqrt{b^2+(a-c)^2}) \quad (16)$$

The minimum moment m_ψ is equivalent to the cornerness measure in the corner detector. In other words, if the value of m_ψ at a point is large, the point is most likely to be a corner feature; and the maximum moment M_ψ is the edge map of an image, which can be used for edge feature detection. Specifically, we first compute m_ψ and M_ψ of the PC maps. For the minimum moment map m_ψ, local maxima detection and non-maximal suppression are performed, and the remaining points are accepted as corner features. Because edge structure information has better resistance to radiation distortions, we also use the maximum moment map M_ψ to detect the edge feature points, that is, perform FAST feature detection on M_ψ. Therefore, the proposed method integrates corner features and edge features for feature matching.

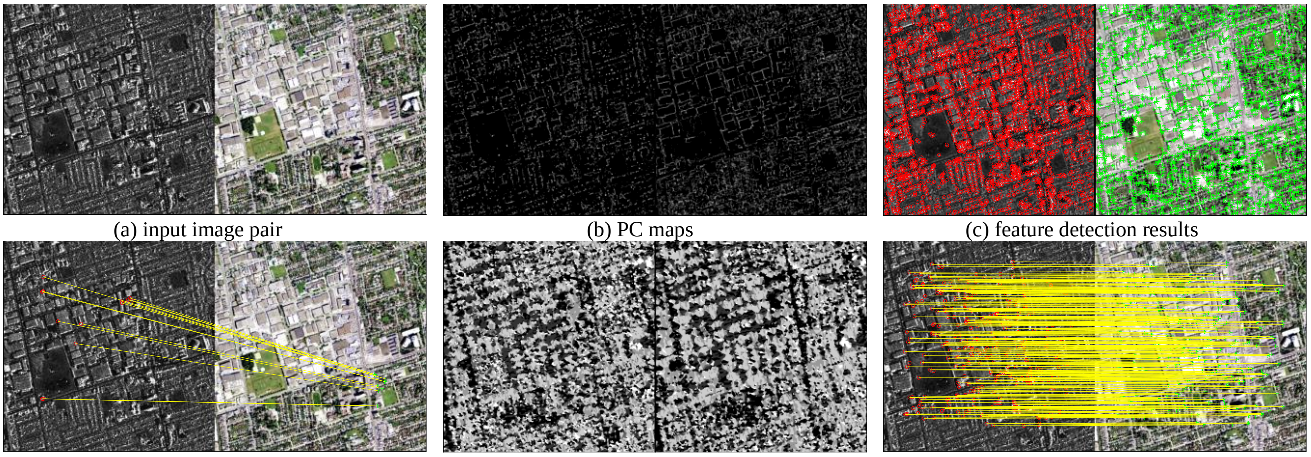
Fig. 1 shows an example of feature detection, where Fig. 1(a) is a pair of multi-modal images (an optical satellite image and a LIDAR depth map); Fig. 1(b) and Fig. 1(c) are the minimum moment map m_ψ and maximum moment map M_ψ, respectively; Fig. 1(d) is the result of FAST [28] feature detection; Fig. 1(e) and Fig. 1(f) are our corner features and edge features, respectively. From the results, we can draw several conclusions: (1) Comparing Fig. 1(b) and Fig. 1(f), we find that traditional feature detectors based on image intensity or gradient (such as FAST or Harris detectors) are very sensitive to NRD, while the moments of PC measures have good invariance to NRD. A large number of reliable feature points can be obtained by performing the same FAST or Harris detector on the maximum moment of PC maps. (2) From Fig. 1(e), it can be seen that the obvious corner features can be obtained. However, the number of feature points is relatively small. (3) From Fig. 1(f), it can be seen that more feature points can be detected on the maximum moment map, but many corner features are missed.



(a) input image pair (b) minimum moment map m_ψ (c) maximum moment map M_ψ

(d) FAST features (e) m_ψ corner features (f) M_ψ edge features

Fig. 1. Feature detection.



(a) input image pair (b) PC maps (c) feature detection results

(d) matching based on PC measure (e) MIMs (f) matching based on MIM

Fig. 2. Comparison of MIM with PC.

Thus, the combination of the characteristics of the minimum moment map corner features and the maximum moment map edge features not only ensures the high repeatability of the features but also the large number of features, which lays a foundation for subsequent feature matching.

B. Feature Description

Once feature points are detected, feature description is needed to increase the distinction between features. Classical feature descriptors generally use image intensity or gradient distribution to construct feature vectors. However, as mentioned earlier, both intensity and gradient are very sensitive to NRD. These descriptors are not suitable for the multi-modal image matching task. In the above, we analysed the characteristics of the PC measure, whose advantage is the robustness to NRD. Intuitively, using a PC map instead of an intensity map or gradient map for feature description is more suitable. However, experimental results do not reach our expectations. Specifically, we selected an image pair that was composed of an SAR satellite image and an optical satellite image for testing (see Fig. 2(a)). We first compute the PC maps (see Fig. 2(b)) and detect corner and edge features from each image (see Fig. 2(c)). Then, for each feature, we construct a feature vector based on the distribution histogram technique similar to SIFT. The matching result based on the PC map description is shown in Fig. 2(d).

From Fig. 2, we can see that the number and distribution of extracted features are quite good; however, the matching result is very poor, as the matches are almost all outliers. It shows that the PC map is not quite suitable for feature description. The reasons may be as follows. First, there is less information from the PC map since most pixel values in the PC map are close to zero. It is not robust enough for feature description. Second, the PC map is sensitive to noise because it mainly contains edges, which causes the feature description to be inaccurate. With such analyses, we present a MIM measure instead of a PC map for feature description.

1) Maximum Index Map (MIM)

A MIM is constructed via the log-Gabor convolution sequence. The convolution sequence was obtained in the PC map calculation stage. Therefore, the computation complexity of a MIM is very small. Fig. 3 illustrates the construction of a MIM. Given an image $I(x,y)$, we first convolve $I(x,y)$ with a 2D-LGF to obtain the response components $E_{so}(x,y)$ and $O_{so}(x,y)$; and then calculate the amplitude $A_{so}(x,y)$ at scale s and orientation o. For orientation o, the amplitudes of all N_s scales are summed to obtain a log-Gabor layer $A_o(x,y)$,

$$A_o(x,y)=\sum_{n=1}^{N_s} A_{so}(x,y) \qquad (17)$$

The log-Gabor convolution sequence is obtained by arranging the log-Gabor convolution layers in order, which is a multi-channel convolution map $\{A_o^{\omega}(x,y)\}_1^{N_o}$, where N_o is the number of orientations; the superscript $\omega = 1,2,...,N_o$ represents the different channels of the log-Gabor convolution sequence. Thus, for each pixel position (x_j, y_j) of the convolution map, we can get an N_o-dimensional ordered array $\{A_o^{\omega}(x_j,y_j)\}_1^{N_o}$. Then, find the maximum value $A_{\max}(x_j,y_j)$ and its corresponding location channel $\omega_{\max}$ in this array by $[A_{\max}(x_j,y_j),\omega_{\max}]=\max\{\{A_o^{\omega}(x_j,y_j)\}_1^{N_o}\}$. We set $\omega_{\max}$ as the pixel value of position (x_j,y_j) in the MIM.

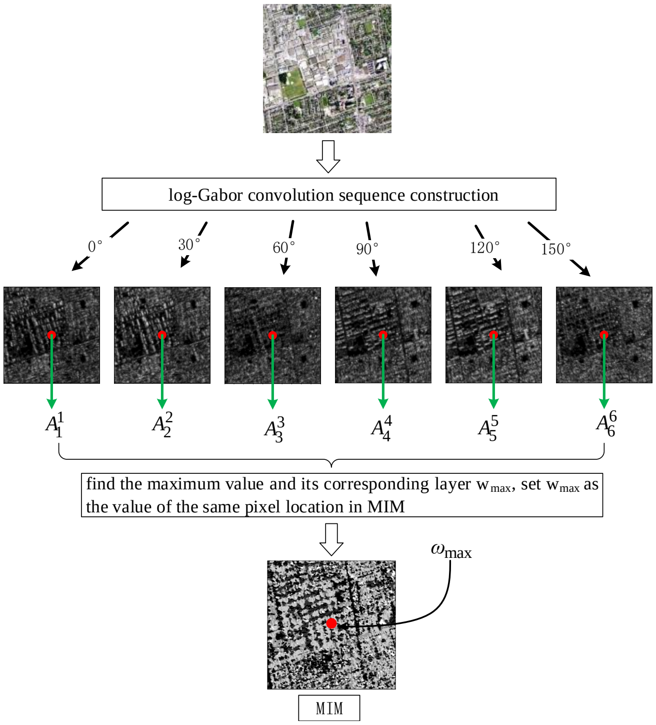

log-Gabor convolution sequence construction
0°
30°
60°
90°
120°
150°
A_1^1
A_2^2
A_3^3
A_4^4
A_5^5
A_6^6
find the maximum value and its corresponding layer $w_{\max}$, set $w_{\max}$ as the value of the same pixel location in MIM
$\omega_{\max}$
MIM

Fig. 3. The construction of a MIM.

After obtaining the MIM, we use a distribution histogram technique similar to SIFT for feature vector description. In detail, for each feature point, we select a local image patch with $J \times J$ pixels centred at the feature and use a Gaussian function whose standard deviation is equal to $J/2$ to assign weights for each pixel. This process avoids abrupt changes in the feature description if the window position changes. We then divide the local patch into 6×6 sub-grids and build a distribution histogram with N_o bins for each sub-grid because the values of MIM range from 1 to N_o. The feature vector is obtained by concatenating all the histograms. Thus, the dimension of the feature vector is 6×6×N_o. To gain invariance to illumination changes, we finally normalise the feature vector.

A matching example based on the MIM description is given in Fig. 2, where Fig. 2(e) is the MIM corresponding to Fig. 2(a), and Fig. 2(f) is the matching result. In this experiment, we set $N_o = 6$ and use the MIM instead of a traditional gradient for the feature description. We regard the feature point pairs with minimal Euclidean distance as potential matches and apply the NBCS [10] method for outlier removal. As seen, the proposed method can extract a large number of reliable matches with relatively uniform distribution, even in the SAR and optical image pair. The imaging mechanisms of SAR and the optical sensors are quite different, which results in large NRD between the SAR image and optical image. Thus, it shows that the proposed MIM descriptor is very suitable for the multi-modal image matching task and is much better than traditional feature matching methods.

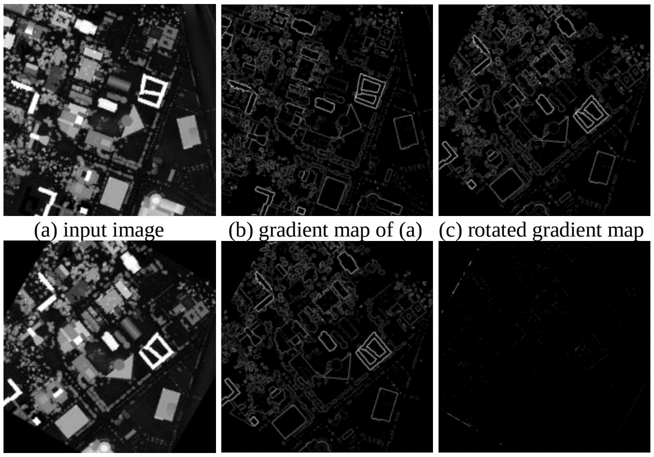
2) *Rotation Invariance*

The previous section analysed the possibility and validity of the MIM for feature description and described the feature vector construction details. However, the description method assumes that there are no rotations between an image pair; that is, the rotation changes are not considered. Thus, if there is a rotation change in the image pair, the above method will no longer be suitable. Therefore, special processing must be performed to make it rotationally invariant. The most straightforward idea is to use the dominant orientation method similar to SIFT. However, after extensive experiments, we found that rotation invariance cannot be achieved by a dominant orientation method.

To analyse the reasons, two experiments are performed. Fig. 4 analyses the effect of the rotations on the gradient map, where Fig. 4(a) is a LIDAR point cloud depth map; Fig. 4(d) is obtained by rotating Fig. 4(a) clockwise 30°; Fig. 4(b) and Fig. 4(e) are the gradient maps of Fig. 4(a) and Fig. 4(d), respectively. To eliminate the rotational difference between Fig. 4(b) and Fig. 4(e), Fig. 4(b) is rotated clockwise by 30° to obtain Fig. 4(c); Fig. 4(f) is the difference between Fig. 4(b) and Fig. 4(e). According to Fig. 4(f), the gradient maps after removing the rotation difference are basically the same, indicating that the rotation has no influence on the values of the gradient map. Therefore, by calculating the main orientation of the feature point, the rotation difference between the local image patches can be eliminated, thereby achieving rotation invariance. Similarly, the above analysis is also performed on the MIM, as shown in Figure 5. Fig. 5(b) and Fig. 5(e) are the MIMs of Fig. 5(a) and Fig. 5(d), respectively; Fig. 5(c) is obtained by rotating Fig. 5(b) clockwise 30°; Fig. 5(f) is the difference between Fig. 5(b) and Fig. 5(e). The dominant orientation method can be applied only if Fig. 5(c) is similar enough to Fig. 5(e). However, most of the values of Fig. 5(f) are not close to zero, indicating that there is not only a rotation difference between Fig. 5(c) and Fig. 5(e) but also a numerical difference, and this numerical difference is caused by the rotations. Thus, to achieve rotation invariance, we must determine the relationship between rotations and the values of the MIM.

As mentioned previously, the MIM is constructed based on the log-Gabor convolution sequence, and the convolution layer is closely related to the orientations. Therefore, if the start layer of the log-Gabor convolution sequence is different, then the constructed MIM is completely different.

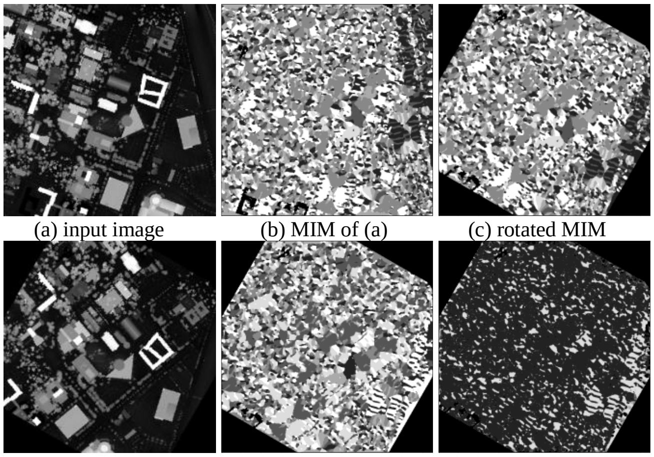
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2959244, IEEE Transactions on Image Processing



(a) input image (b) gradient map of (a) (c) rotated gradient map

(d) rotated image (e) gradient map of (d) (f) error map

Fig. 4. The effect of rotations on the gradient map.

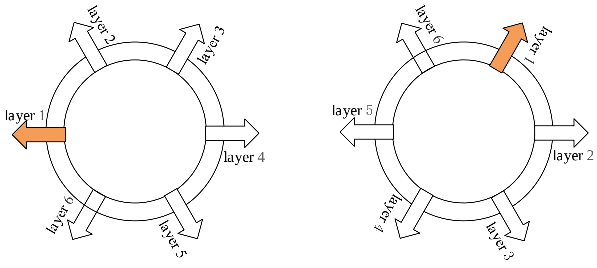


(a) input image (b) MIM of (a) (c) rotated MIM

(d) rotated image (e) MIM of (d) (f) error map

Fig. 5. The effect of rotations on a MIM.

In other words, if two images are to be successfully matched, the log-Gabor convolution sequences corresponding to the two images must be highly similar, and each layer of the log-Gabor convolution sequence needs to be similar. In fact, the log-Gabor convolution sequence can be thought of as an end-to-end annular structure, as shown in Fig. 6. Assume that Fig. 6(a) is a 6-layers log-Gabor convolution sequence ring (noted by S_A) obtained from the original image (image in Fig. 4(a)), where the first layer is the 0° direction convolution result (the initial layer of the convolution sequence); the second layer is the 30° direction convolution result, and so on. The sixth layer is the 150° direction convolution result. However, if we rotate the image by an angle (as shown in Fig. 6(b)), and still use the 0° direction convolution result as the initial layer to construct the convolution sequence (obtaining convolution sequence S_B), due to the effect of the rotation, the content of the initial layer of S_A will be quite different from S_B. In fact, which layer should be used as the initial layer is not known because it is highly related to the rotation angle. Considering that N_o is small and generally set to 6, we use the simplest traversal strategy, listing all possible scenarios. In detail, we first construct a convolution sequence S_A and a convolution sequence S_B for the reference image and the target image, respectively. For S_A of the reference image, we directly construct a MIM (MIM^{S_A}); for S_B of the target image, we successively transform the initial layer of S_B to reconstruct a set of convolution sequences $\{S_w^B\}_1^{N_o}$ with different initial layers and then calculate a MIM from each convolution sequence to obtain a set of MIMs $\{MIM_w^{S_B}\}_1^{N_o}$. In general, there is always a MIM in set $\{MIM_w^{S_B}\}_1^{N_o}$ that is similar to MIM^{S_A}. To verify this conclusion more intuitively, we perform an experiment on the image in Fig. 4(d) to obtain the MIM set $\{MIM_w^{S_B}\}_1^{N_o}$ ($N_o = 6$). Fig. 7 shows all the MIMs in the set $\{MIM_w^{S_B}\}_1^{N_o}$. The initial layers of $\{S_w^B\}_1^{N_o}$ are the first to sixth layers of the convolution sequence S_B. As seen, if the initial layers are different, the resulting MIM is completely different. Fig. 8 shows the difference between each subfigure in Fig. 7 and Fig. 5(c). It is found that when the sixth layer is used as the initial layer, the constructed MIM $MIM_6^{S_B}$ is very consistent with MIM^{S_A}, which verifies the above conclusion.



(a) convolution sequence S_A (b) convolution sequence S_B

Fig. 6. Convolution sequence ring.

The above process substantially eliminates the effect of rotations on the values of the MIM. Then, the dominant orientation method can be directly applied to gain rotation invariance. In summary, the proposed RIFT algorithm builds a feature vector for each keypoint of the reference image and N_o feature vectors for each keypoint of the target image.

IV. Experimental results

To verify the effectiveness of the proposed RIFT method, we select several multi-modal datasets for qualitative and quantitative evaluation. We compare our RIFT algorithm against four state-of-the-art algorithms, i.e., SIFT, SAR-SIFT, LSS, and PIIFD. LSS is only a feature descriptor; hence, we use the proposed detector to extract features. For a fair comparison, the sizes of local description patches for these five methods are set to the same; and the implementation of each compared method are obtained from the authors' personal website. The parameters of each method are fine-tuned to obtain the best performance and are consistent in all experiments.

A. Datasets

Six types of multi-modal image datasets are selected as experimental sets, including optical-optical, infrared-optical, SAR-optical, depth-optical, map-optical, and day-night. Each type of dataset contains 10 image pairs for a total of 60 multi-modal image pairs. The sample data are shown in Fig. 9.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2959244, IEEE Transactions on Image Processing

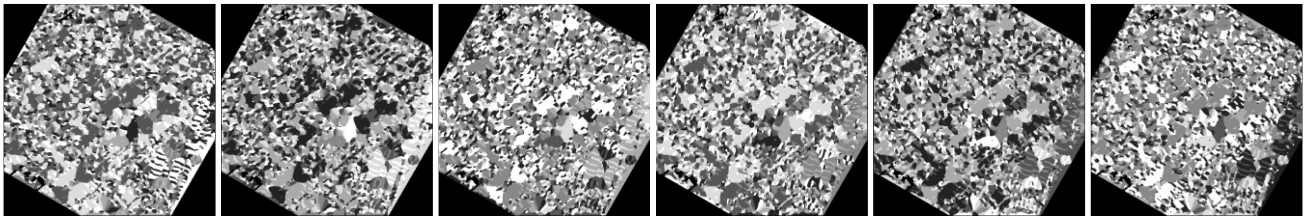


Fig. 7. MIM set $\{MIM_w^{S_B}\}_1^{N_o}$ ($N_o = 6$). The initial layers of are the first to sixth layers of the convolution sequence S_B .

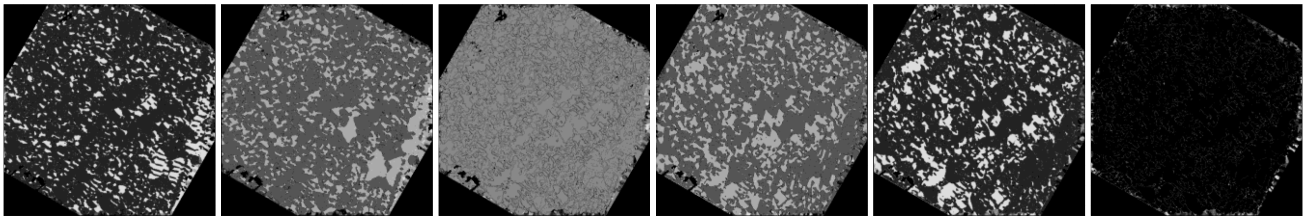
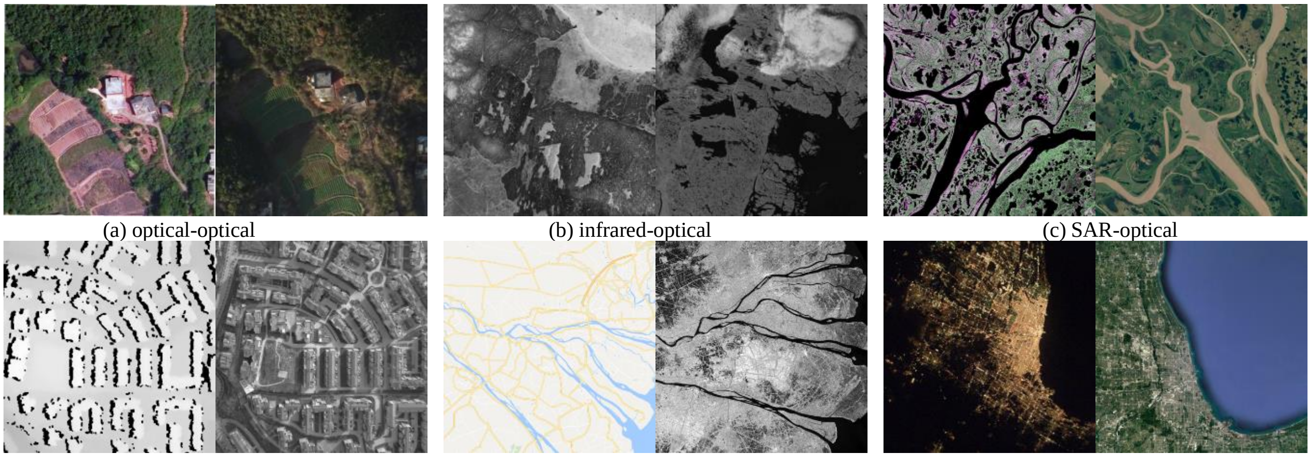


Fig. 8. Error map between each subfigure in Fig. 7 and Fig. 5(c).



(a) optical-optical

(b) infrared-optical

(c) SAR-optical

(d) depth-optical

(e) map-optical

(f) day-night

Fig. 9. Sample data.

These image pairs include not only multi-sensor images and multi-temporal images but also artificially produced images, such as rasterised map data; not only images under good lighting conditions (daytime images) but also nighttime remote sensing images; not only high spatial resolution images but also low and medium spatial resolution images, whose GSD ranges from 0.1 metres to hundreds of metres; not only satellite images but also UAV images and even close-range images; not only urban area images but also countryside and mountain forest images. There are serious distortions between these image pairs, especially radiation distortions, which will create great challenges for the image matching algorithms. Such challenges can test the validity and robustness of the proposed RIFT algorithm more comprehensively. It should be noted that RIFT is currently not scale invariant. Therefore, the two images of each image pair need to be resampled to have approximately the same GSD.

For better quantitative evaluation, we need to obtain a ground truth geometric transformation between each image pair. However, due to the interference of various factors, the real datasets usually do not have a true ground truth geometric transformation. The approximate ground truth geometric transformation is generally used for evaluation. Specifically, for each image pair, we select five evenly distributed correspondences with sub-pixel accuracy and use these correspondences to estimate an accurate affine transformation as the approximation of the ground truth geometric transformation. We first perform feature matching on this image pair (RIFT/SIFT/SAR-SIFT) and remove the outliers based on the NBCS [10] method; then, we calculate the residuals of these image correspondences under the estimated affine transform and regard the correspondences with residuals less than 3 pixels as the correct matches. We use the number of correct matches (NCM), root mean square error (RMSE), mean error (ME), and success rate (SR) as the evaluation metrics. Note that if the NCM of an image pair is less than four, the matching is considered to have failed.

B. Parameter Study

The proposed RIFT method contains three main parameters, namely, N_s , N_o , and J . Parameter N_s is the number of convolution scales of the log-Gabor filter, and its value is usually greater than 1. Parameter N_o is the number of convolution orientations of the log-Gabor filter. In general, the higher the number of orientations is, the richer the amount of information of the constructed MIM and the higher the computational complexity. Parameter J is the size of the local image patch used for feature description. If the local patch is too small, it contains insufficient information, which does not adequately reflect the distinctiveness of the feature. In contrast, if the image patch is too large, it is easily affected by the local geometric distortion. Therefore, suitable parameters are very important. This section performs a parameter study and sensitivity analysis based on a map-optical dataset. We design three independent experiments to learn parameters N_s, N_o, and J, where each experiment has only one parameter as a variable and other parameters are fixed values. The experimental setup details are summarised in Table 1. For each parameter, we use NCM and SR as the evaluation metrics. The experimental results are reported in Table 2~Table 4.

From the experimental results, we can infer that (1) larger values of N_o mean that richer information of the constructed MIM, and thus more NCM, can be obtained; however, larger values of N_o also mean that the number of convolution sequences increases, which will greatly increase the computational complexity of the algorithm. From Table 2, when N_o reaches 6, the SR of RIFT reaches 100%. However, increasing the number of orientations only slightly improves the NCM. Therefore, to take into account both the matching performance and computational complexity of RIFT, we set N_o to 6. (2) From Table 3, we can see that small values of N_s result in low SR accuracy and large values of N_s result in poor NCM performance. When $N_s = 4$, RIFT achieves the best performance in both SR and NCM metrics. Although the results of $N_s = 3$ are only slightly different from the results of $N_s = 4$, the number of scales is different from the number of directions, and increasing the scales does not significantly increase the computational complexity. Therefore, we set N_s to 4. (3) The influence of the parameter J on RIFT is similar to N_s. If the value of J is small, the amount of information is not rich enough, and the SR and NCM metrics will be poor; however, if the value of J is large, due to the effect of local geometric distortions, the performance of NCM will decrease. As shown in Table 4, RIFT achieves the best performance when $J = 96$. To reduce computational complexity, we select $J = 72$. Based on the experimental results and analyses, these parameters are fixed to $N_o = 6$, $N_s = 4$, $J = 72$ in the following experiments.

Table 1 The details of parameter settings.

experiments	variable	fixed parameters
parameter N_o	$N_o = [4,\ 5,\ 6,\ 7,\ 8]$	$N_s = 3$, $J = 96$
parameter N_s	$N_s = [2,\ 3,\ 4,\ 5,\ 6]$	$N_o = 6$, $J = 96$
parameter J	$J = [48,\ 72,\ 96,\ 120,\ 144]$	$N_o = 6$, $N_s = 3$

Table 2 The results of parameter N_o.

metric	N_o, $N_s = 3$, $J = 96$				
	4	5	6	7	8
NCM	50.4	84.9	114.8	120.5	121.4
SR/%	60	70	100	100	100

Table 3 The results of parameter N_s.

metric	N_s, $N_o = 6$, $J = 96$				
	2	3	4	5	6
NCM	81	114.8	119.8	102.5	89.6
SR/%	80	100	100	100	100

Table 4 The results of parameter J.

metric	J, $N_o = 6$, $N_s = 3$				
	48	72	96	120	144
NCM	91.9	111.6	119.8	116.6	98.7
SR/%	60	100	100	100	100

Table 5 Evaluation feature detectors on multi-modal image datasets.

metric	FAST	Harris	Brisk	SIFT	SURF	MSER	RIFT
N^c	472.8	169.5	93.2	110.9	236.8	78.8	652.2
Rep /%	17.4	13.0	8.1	8.8	13.1	8.6	22.9
$r_{N_c>100}$	65.0	55.0	36.7	40.0	81.7	25.0	100.0
$r_{Rep>10\%}$	61.7	61.7	38.3	33.3	76.7	36.7	91.7

C. Detector evaluation

We compare the detector of RIFT with other six well-known feature detectors, including FAST [28], Harris [32], Brisk [56], SIFT [15], SURF [24], and MSER [33]. We use the repeatability Rep and the number of correspondences N^c that can be established for detected features as evaluation metrics. The repeatability is a ratio between N^c and the average number of features detected in two images I_1 and I_2 [57],

$$Rep = \frac{N^c}{(n_1 + n_2)/2} = \frac{\left|\{\| \mathbf{x}_i^1 - \mathbf{H}\mathbf{x}_i^2 \| < 3\}_{i=1}^{n_1}\right|}{(n_1 + n_2)/2} \tag{18}$$

where $\mathbf{H}$ is the ground truth transformation between I_1 and I_2; $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$ are homogeneous coordinates of a feature in I_1 and I_2, respectively; n_1 and n_2 are the number of features in I_1 and I_2, respectively; $\left|\{\| \mathbf{x}_i^1 - \mathbf{H}\mathbf{x}_i^2 \| < 3\}_{i=1}^{n_1}\right|$ returns the number of matches that satisfy $\| \mathbf{x}_i^1 - \mathbf{H}\mathbf{x}_i^2 \| < 3$.

We use the whole 60 multi-modal image pairs as the evaluation dataset and compute the average repeatability and the average number of correspondences for each detector. The results are summarized in Table 5. To show the stability of each detector, Table 5 also reports two ratio metrics, i.e., $r_{N_c>100}$ and $r_{Rep>10\%}$, whose definitions are as follows,

$$\begin{cases} r_{N_c>100} = \left|\{N_i^c > 100\}_1^{N_{ip}}\right| \Big/ N_{ip} \\ r_{Rep>10\%} = \left|\{Rep_i > 10\%\}_1^{N_{ip}}\right| \Big/ N_{ip} \end{cases} \tag{19}$$

where N_{ip} is number of image pairs in the evaluation dataset.

As shown, all the detectors have low repeatability on the multi-modal image datasets because of serious NRDs, which means that matching multi-modal image pairs is a very challenging task. Among these six detectors, FAST and SURF perform much better than others. FAST has higher repeatability and obtains more correspondences than SURF, while getting lower $r_{N_c>100}$ and $r_{Rep>10\%}$ values. Considering the performance and the efficiency of FAST, we apply it on the maximum moment map to detect edge features. The detector of RIFT ranks best in all the four metrics. It gains 5.5 percentages in terms of Rep and 179.4 correspondences in terms of N^c compared with FAST, which ranks second.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2959244, IEEE Transactions on Image Processing

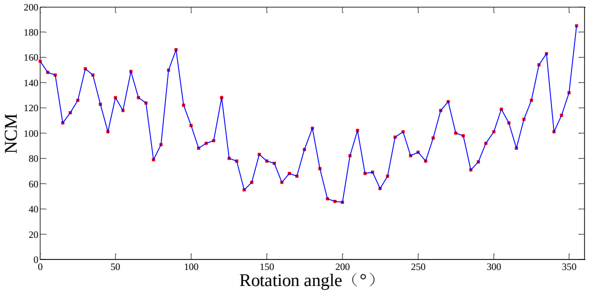
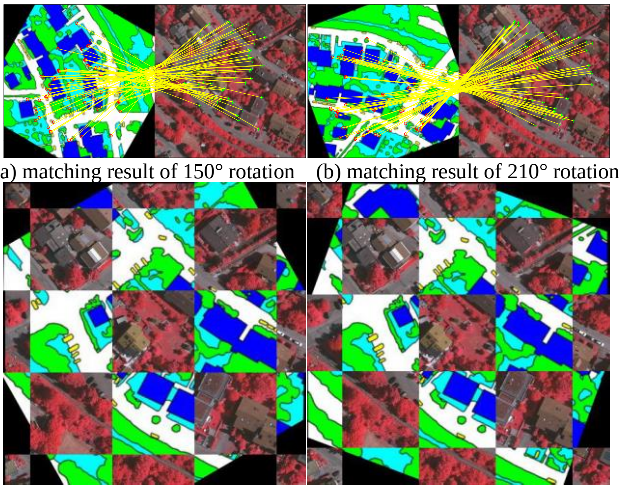


Fig. 10. Rotation invariance test.


(a) matching result of 150° rotation (b) matching result of 210° rotation

(a) registration result of 150° rotation (b) registration result of 210° rotation

Fig. 11. Matching and registration results.

D. Rotation Invariance Test

Rotation invariance is an important property of the proposed RIFT, which is also a major advantage compared to the HOPC method. The calculation of MIMs and PC maps are both related to orientations. The proposed RIFT algorithm generally performs log-Gabor convolution filtering along six directions, i.e., 0°, 30°, 60°, 90°, 120°, and 150°. The angles of these directions only range from 0° to 150°, which inevitably raises the concern: "If the rotation angle between the image pairs is not within this range, is the proposed RIFT still robust?"

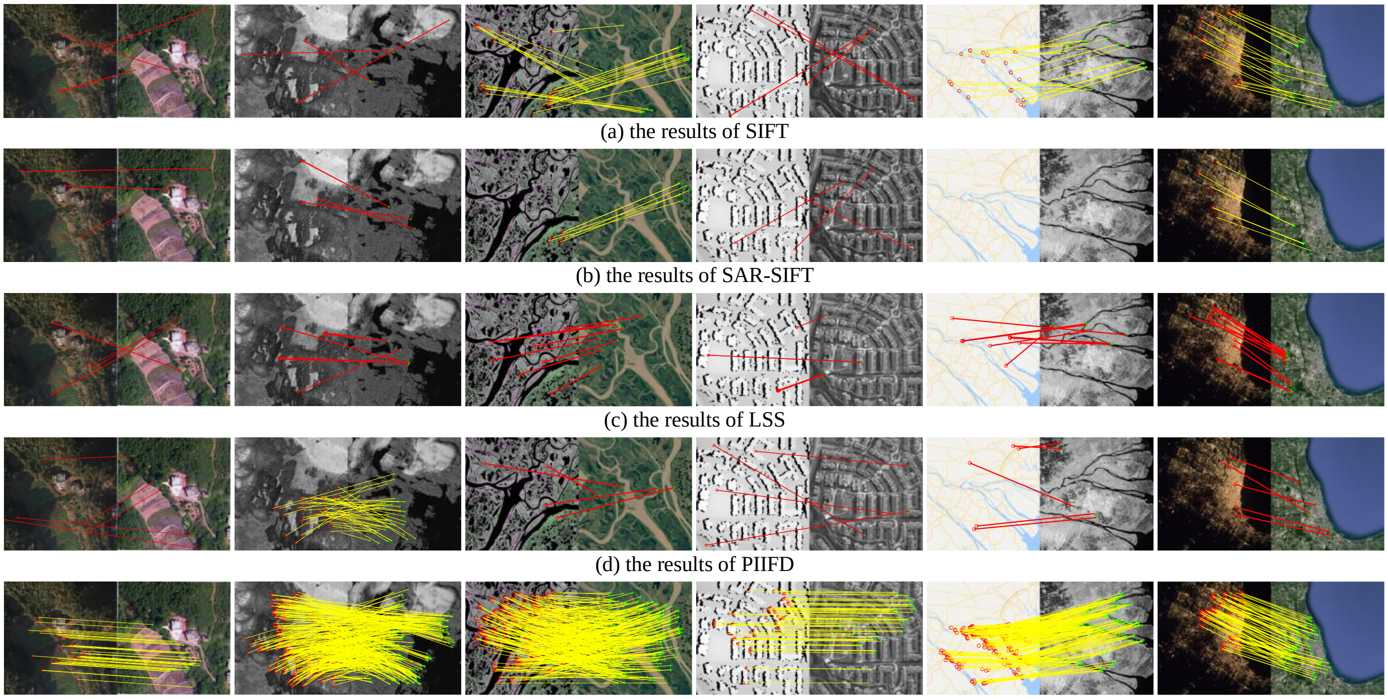
In fact, the proposed RIFT has very good rotation invariance, not only for the rotations between [0°~150°] but also for the rotations in the entire 360° range. To verify this conclusion, an image pair was selected from the map-optical dataset for experimentation. This image pair does not suffer from rotation changes. First, we rotate the map of this image pair. The rotation angles are from 0° to 359° with an interval of 5°. Thus, a total of 72 maps are obtained (the rotation angles are $[0^\circ, 5^\circ, 10^\circ, ..., 345^\circ, 350^\circ, 355^\circ, 359^\circ]$). These 72 maps and the optical image constitute 72 image pairs. Then, these images are processed one by one using RIFT, and their corresponding NCMs are plotted in Fig. 10. The red dots in the figure represent the NCMs. It can be clearly seen that although the NCMs under different rotation angles are different, all the NCMs are greater than 40, indicating that the proposed RIFT can successfully match all the image pairs, and the matching SR accuracy is 100%, which also verifies that the proposed RIFT has good rotation invariance for rotations in the entire 360° interval. Meanwhile, the differences in the NCMs also indicate that the dominant orientation calculation of RIFT may not be optimal, and a more robust feature main orientation calculation method will further improve the matching performance of the proposed RIFT, which will become one of our key research topics in the future. Fig. 11 shows the experimental results for 150° rotation and 210° rotation. Among them, the first row is the results of feature matching (yellow lines in the figure represent correct matches), and the second row is the registration results. It can be seen that the NCMs are large; the distribution of matching points is relatively uniform, and the registration accuracy is very high.

E. Matching Performance Test

Qualitative comparisons: we select the first image pairs from the six multi-modal datasets for evaluation, as shown in Fig. 9. Among them, Fig. 9(a) contains translation, small rotation, and small-scale changes; Fig. 9(b) includes a translation change and a 90° rotation change; Fig. 9(c), Fig. 9(e), and Fig. 9(f) includes both translation and rotation changes; Fig. 9(d) only contains a translation change. Since these image pairs are all multi-modal image pairs, the imaging mechanism is quite different, and these image pairs contain severe NRD. Therefore, matching on these image pairs is very challenging. Fig. 12 plots the results of SIFT, SAR-SIFT, LSS, PIIFD, and the proposed RIFT, respectively.

As seen, SIFT algorithm fails to match on the first, second, and fourth image pairs in Fig. 12(a). The SR accuracy is 50%. However, even if the matching is successful, the NCMs are also small, i.e., 24, 24, and 23. Because SIFT algorithm uses gradient histograms for feature description, the matching results depend heavily on the similarity of the gradient maps of the image pair. The above analysis shows that the gradient map is very sensitive to NRD, which is the fundamental reason for its poor matching performance on multi-modal images. SAR-SIFT algorithm fails to match on the first, second, fourth, and fifth image pairs in Fig. 12(b). The SR accuracy is only 33.3%. Similarly, the NCMs for SAR-SIFT are also small, 6 and 8, respectively. Although SAR-SIFT redefines the concept of the gradient to fit the SAR image matching task, the redefined gradient is even more sensitive to NRD. In addition, SAR-SIFT uses a multiscale Harris detector for feature detection. The detector usually obtains fewer feature points and cannot obtain any keypoints on some images. For example, the number of feature points detected by SAR-SIFT on the Google map in Fig. 12(b) is 0; thus, the NCM must be 0. LSS fails on all these six image pairs. PIIFD only matches successfully on the second image pair. In contrast, the proposed RIFT algorithm matches successfully on all these six image pairs, whose SR accuracy is 100%. The NCMs of RIFT are large, i.e., 41, 368, 335, 66, 230 and 93. The average NCM of RIFT is approximately 7.8 times that of SIFT.

The matching performance of RIFT on the image pairs with NRD is far superior to the current popular feature matching methods. There are two main reasons: (1) RIFT uses a PC map instead of image intensity for feature detection, and considers both the feature repetition rate and the feature number, which lays a foundation for subsequent matching. (2) RIFT adopts the log-Gabor convolution sequence to construct a MIM instead of a gradient map for feature description. The MIM has very good robustness to NRD, thus ensuring the accuracy of the feature vectors. Fig. 13 shows more results of RIFT.



(a) the results of SIFT

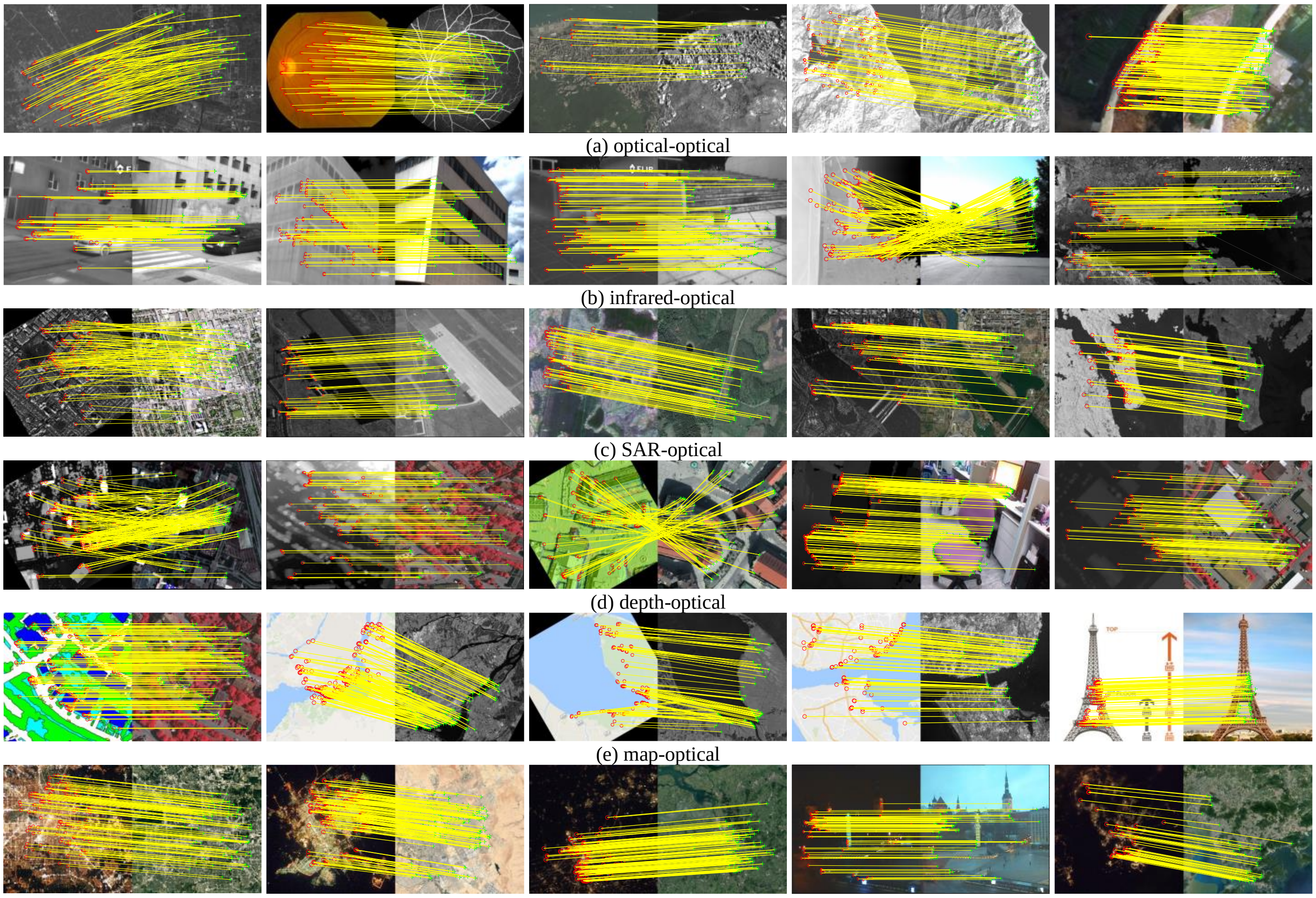
(b) the results of SAR-SIFT

(c) the results of LSS

(d) the results of PIIFD

(f) the results of RIFT

Fig. 12. Qualitative comparison results on the sample data. The red circles and the green crosshairs in the figure indicate the feature points on the reference image and the target image, respectively; the yellow lines and the red lines indicate the correct matches and the outliers, respectively.



(a) optical-optical

(b) infrared-optical

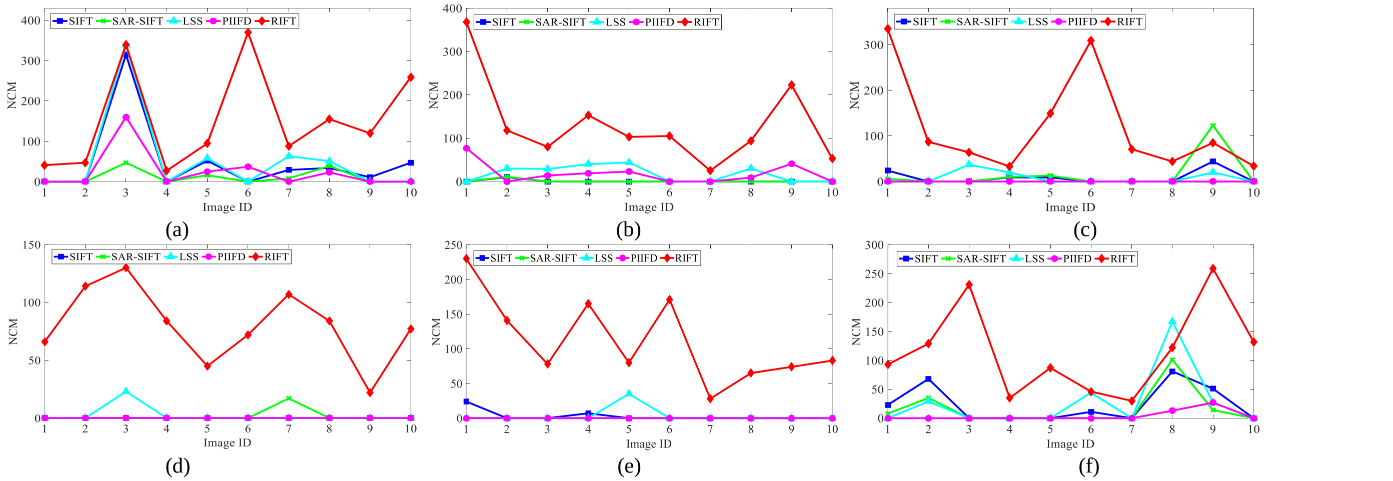
(c) SAR-optical

(d) depth-optical

(e) map-optical

(f) day-night

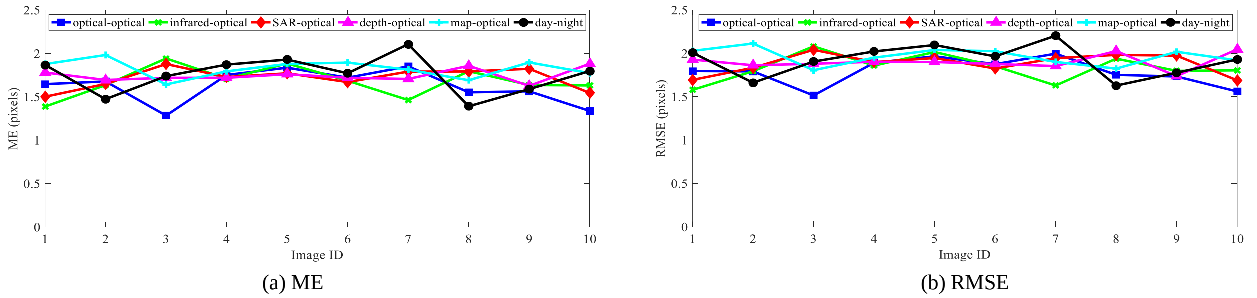
Fig. 13. More results of the proposed RIFT. For better visualization, no more than 100 matches are displayed.



(a) (b) (c)

(d) (e) (f)

Fig. 14. Comparisions on NCM metric.



(a) ME (b) RMSE

Fig. 15. The ME and RMSE of the proposed RIFT.

Table 6 Comparisons on SR metric.

method	SR/%					
	optical-optical	infrared-optical	SAR-optical	depth-optical	map-optical	day-night
SIFT	60	10	40	0	20	50
SAR-SIFT	40	10	40	10	0	40
LSS	40	50	30	10	10	40
PIIFD	40	60	0	0	0	20
RIFT	100	100	100	100	100	100

Table 7 Quantitative evaluation results of RIFT.

metric	optical-optical	infrared-optical	SAR-optical	depth-optical	map-optical	day-night
NCM	154.2	128.2	121.1	80.1	111.5	116.4
ME/pixels	1.62	1.68	1.71	1.75	1.82	1.75
RMSE/pixels	1.79	1.84	1.88	1.91	1.96	1.92

Quantitative comparisons: Fig. 14 is the quantitative results of the NCM metric, where Fig. 14(a)~Fig. 14(f) show the results of the three comparison methods on the six multi-modal datasets. As seen, SIFT performs better on the optical-optical dataset and the day-night dataset than the other 4 datasets because of its resistance to illumination changes. In the optical-optical dataset, the difference in the imaging mechanism between images is smaller than that of the other four datasets, and matching is relatively easy. The day-night dataset is also essentially an optical-optical dataset. The difference is that the light conditions of the day-night dataset are more complex. SIFT performs the worst on the depth-optical dataset. The SR accuracy is zero, and no correct matches are obtained. The reasons may be as follows: (1) SIFT uses gradient information for feature description. The gradient can reflect the structural information (edge information) of an image to a certain extent. However, in a depth map or disparity map, the edge structure is relatively weak. (2) SIFT detects feature points directly based on intensity. The number of extracted feature points is small, and the distribution is poor, (especially in-depth maps and disparity maps, as shown in Fig. 1), resulting in poor matching performance. In most of the successfully matched image pairs, the NCMs of SIFT are very small (smaller than 50). In some images, there are only a few correctly matching points. The performance of SAR-SIFT is similar to that of the SIFT algorithm, and its performance on the optical-optical dataset, the SAR-optical dataset, and the day-night dataset is superior to the other 4 datasets. As described above, the difference in the imaging mechanism between the image pairs of the optical-optical and day-night dataset is relatively small. Because the SAR-SIFT algorithm is specifically designed for SAR image matching and the gradient concept is redefined, it may be more suitable for the SAR-optical dataset. SAR-SIFT performs the worst on the infrared-optical dataset and the map-optical dataset and almost fails completely. The radiation characteristics of the infrared-optical datasets are quite different, and the radiation characteristics of most objects are completely opposite. As shown in Fig. 9(b), black objects in the optical image appear white in the infrared image. Therefore, the redefined gradient may be more sensitive to this inverse difference. As previously analysed, a multi-scale Harris detector has difficulty extracting feature points on the map of the map-optical dataset, which will inevitably lead to a matching failure. The NCMs of SAR-SIFT are also very small in most of the successfully matched image pairs. However, on a few image pairs, such as image pair 9 of the SAR-optical dataset, the NCM obtained by SAR-SIFT is even larger than the proposed RIFT. In general, the matching performance of SAR-SIFT is extremely unstable. LSS performs much better on the infrared-optical dataset than the depth-optical and map-optical datasets. PIIFD totally fails on the depth-optical and map-optical datasets. In contrast, the proposed RIFT successfully matches all the image pairs of the six datasets, and the NCMs are much greater than 50 on most of the image pairs. The matching performance of RIFT is very stable and robust, and it is hardly affected by the type of radiation distortions. RIFT is far superior to other methods.

Table 6 summarises the matching SRs of the six methods on each data set. As shown, SIFT has the highest SR on the optical-optical dataset, which is 60%; the SRs of SAR-SIFT on the optical-optical, the SAR-optical, and the day-night datasets are all 40%; both LSS and PIIFD achieves the highest SRs on the infrared-optical dataset, i.e., 50% and 60%, respectively; and the SRs of the proposed RIFT on all datasets are 100%. The average SRs of SIFT, SAR-SIFT, LSS, PIIFD, and RIFT on all six datasets are 30%, 23.3%, 30%, 20%, and 100%, respectively. Compared with SIFT and LSS, the proposed RIFT improves by 70 percent. Fig. 15 plots the average ME and RMSE of the proposed RIFT on each image pair. Because other methods have insufficient SR accuracy, their corresponding ME and RMSE are not calculated. As seen, the MEs and RMSEs of RIFT are between 1 pixel and 2.5pixels. There are many reasons for these errors, such as ground truth geometric model estimation error, estimated geometric model error, and the error of feature point positioning. Table 7 reports the NCM, ME, and RMSE of the proposed RIFT on each dataset. From the table, the NCMs of RIFT are relatively large and very stable, all of which are approximately 100; the matching precision is high, where the ME is approximately 1.75 pixels, and the RMSE is approximately 1.9 pixels. As mentioned earlier, RIFT is not affected by the type of radiation distortions. The average NCM, ME, and RMSE over all 60 image pairs are 119.3, 1.72 pixels, and 1.88 pixels, respectively.

Summarising the above qualitative and quantitative experimental results, we can draw the following conclusions: The proposed algorithm is specially designed for NRD problems, including feature detection and feature description. Therefore, the proposed algorithm has very good resistance to NRD and is not affected much by the type of radiation distortions. The proposed method achieved very good NCMs and matching accuracy on all six datasets. The matching performance of RIFT is far superior to the current classical feature matching methods. The proposed RIFT is a feature matching algorithm that has rotation invariance and is suitable for a variety of multi-modal images.

F. Running Time and Limitations

Table 7 reports the average running time of each compared method on the whole 60 image pairs, which is calculated on a laptop with an Intel i7-8550U @ 1.8GHz CPU, 8 GB of RAM. SIFT and LSS are implemented in C++ while others are implemented in Matlab. RIFT$^+$ represents RIFT without rotation invariance. RIFT* deals with the rotation invariance stage based on parallel computing.

As can be seen, the running time of RIFT is about 2 times of SAR-SIFT and 5 times of PIIFD. The computational complexity is one of the limitations of RIFT. Comparison RIFT with RIFT$^+$, we find that the rotation invariance stage has the highest complexity in RIFT. Fortunately, if the rotation prior is known, the rotation invariance stage can be easily disabled by setting a flag parameter in RIFT. In addition, the rotation invariance stage is very suitable for parallel computing and can be easily implemented. As shown, RIFT* only costs a third of the running time of RIFT. If we rewrite RIFT* by C++, the running speed can be increased by an order of magnitude.

RIFT does not build a scale space for feature detection and description. Thus, it is sensitive to large scale and viewpoint variations. To cope with this issue, we can first build a Gaussian scale space that is similar to SIFT; then, apply a well-designed feature detector such as the uniform competency detector [31] on the maximum moment map and use an adaptive binning histogram technique [34] on the MIM to achieve robustness to geometric variations. As mentioned earlier, this paper only focuses on NRD problem in feature matching. We will implement these strategies in a common feature matching framework to achieve the robustness to both geometric and radiation variations.

Table 7 Running time comparisons (seconds).

SIFT (C++)	SAR-SIFT (Matlab)	LSS (C++)	PIIFD (Matlab)	RIFT (Matlab)	RIFT$^+$ (Matlab)	RIFT* (Matlab)
1.93	16.90	112.74	5.92	32.05	4.54	13.34

V. Conclusion

In this paper, we proposed a radiation-variation insensitive feature matching method called RIFT. This method has rotation invariance and is suitable for a variety of multi-modal images. We first introduced the concept of PC. The initial motivation of RIFT is derived from the fact that PC has good radiation robustness. After analysing and summarising the drawbacks of the current methods, we described the details of RIFT. In feature detection, we obtained both corner features and edge features based on a PC map, taking into account the number of features and repetition rate. We proposed a MIM instead of gradient for feature description, which has very good robustness to NRD. We also analysed the inherent influence of rotations on the values of MIM and achieved rotation-invariant by the constructing multiple MIMs. In the experiment, we performed qualitative and quantitative comparisons to verify the reliability and superiority of RIFT. We also analysed the limitations of RIFT.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) (No. 41901398), Natural Science Foundation of Hubei Province (No. 2019CFB167), and China Postdoctoral Science Foundation (No. 2018M640734).

References

[1] B. Zitova, and J. Flusser, “Image registration methods: a survey,” *Image and vision computing,* vol. 21, no. 11, pp. 977-1000, 2003.

[2] M. F. Demirci, A. Shokoufandeh, Y. Keselman, L. Bretzner, and S. Dickinson, “Object recognition as many-to-many feature matching,” *International Journal of Computer Vision,* vol. 69, no. 2, pp. 203-222, 2006.

[3] D. Nister, and H. Stewenius, "Scalable recognition with a vocabulary tree." pp. 2161-2168.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, “ORB-SLAM: a versatile and accurate monocular SLAM system,” *IEEE Transactions on Robotics,* vol. 31, no. 5, pp. 1147-1163, 2015.

[5] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, “MonoSLAM: Real-time single camera SLAM,” *IEEE transactions on pattern analysis and machine intelligence,* vol. 29, no. 6, pp. 1052-1067, 2007.

[6] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, “Medical image registration,” *Physics in medicine & biology,* vol. 46, no. 3, pp. R1, 2001.

[7] J. Li, Q. Hu, and M. Ai, “4FP-Structure: A Robust Local Region Feature Descriptor,” *Photogrammetric Engineering & Remote Sensing,* vol. 83, no. 12, pp. 813-826, 2017.

[8] J. Li, Q. Hu, M. Ai, and R. Zhong, “Robust feature matching via support-line voting and affine-invariant ratios,” *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 132, pp. 61-76, 2017.

[9] J. L. Q. H. M. Ai, “LAM: Locality affine-invariant feature matching,” *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 154, pp. 28-40, August 2019.

[10] J. Li, Q. Hu, and M. Ai, “Robust feature matching for geospatial images via an affine-invariant coordinate system,” *The Photogrammetric Record,* vol. 32, no. 159, pp. 317-331, 2017.

[11] J. Li, Q. Hu, and M. Ai, “Robust feature matching for remote sensing image registration based on $ L_ {q} $-estimator,” *IEEE Geoscience and Remote Sensing Letters,* vol. 13, no. 12, pp. 1989-1993, 2016.

[12] J. Li, Q. Hu, R. Zhong, and M. Ai, “Exterior Orientation Revisited: A Robust Method Based on l q-norm,” *Photogrammetric Engineering & Remote Sensing,* vol. 83, no. 1, pp. 47-56, 2017.

[13] J. A. Richards, and J. Richards, *Remote sensing digital image analysis*: Springer, 1999.

[14] R. A. Schowengerdt, *Remote sensing: models and methods for image processing*: Elsevier, 2006.

[15] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision,* vol. 60, no. 2, pp. 91-110, 2004.

[16] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, “SAR-SIFT: a SIFT-like algorithm for SAR images,” *IEEE Transactions on Geoscience and Remote Sensing,* vol. 53, no. 1, pp. 453-466, 2015.

[17] A. Gruen, “Development and status of image matching in photogrammetry,” *The Photogrammetric Record,* vol. 27, no. 137, pp. 36-57, 2012.

[18] W. K. Pratt, “Correlation techniques of image registration,” *IEEE transactions on Aerospace and Electronic Systems*, no. 3, pp. 353-358, 1974.

[19] A. Mahmood, and S. Khan, “Correlation-coefficient-based fast template matching through partial elimination,” *IEEE Transactions on image processing,* vol. 21, no. 4, pp. 2099-2108, 2012.

[20] E. De Castro, and C. Morandi, “Registration of translated and rotated images using finite Fourier transforms,” *IEEE Transactions on pattern analysis and machine intelligence,* no. 5, pp. 700-703, 1987.

[21] X. Tong, Z. Ye, Y. Xu, S. Liu, L. Li, H. Xie, and T. Li, “A novel subpixel phase correlation method using singular value decomposition and unified random sample consensus,” *IEEE Transactions on Geoscience and Remote Sensing,* vol. 53, no. 8, pp. 4143-4156, 2015.

[22] P. Viola, and W. M. Wells III, “Alignment by maximization of mutual information,” *International journal of computer vision,* vol. 24, no. 2, pp. 137-154, 1997.

[23] F. P. Oliveira, and J. M. R. Tavares, “Medical image registration: a review,” *Computer methods in biomechanics and biomedical engineering,* vol. 17, no. 2, pp. 73-93, 2014.

[24] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (SURF),” *Computer vision and image understanding,* vol. 110, no. 3, pp. 346-359, 2008.

[25] Y. Ke, and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors." pp. II-506-II-513 Vol. 2.

[26] J.-M. Morel, and G. Yu, “ASIFT: A new framework for fully affine invariant image comparison,” *SIAM journal on imaging sciences,* vol. 2, no. 2, pp. 438-469, 2009.

[27] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF." pp. 2564-2571.

[28] E. Rosten, R. Porter, and T. Drummond, “Faster and better: A machine learning approach to corner detection,” *IEEE transactions on pattern analysis and machine intelligence,* vol. 32, no. 1, pp. 105-119, 2008.

[29] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, “BRIEF: Computing a local binary descriptor very fast,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 7, pp. 1281-1298, 2012.

[30] A. Sedaghat, M. Mokhtarzade, and H. Ebadi, “Uniform robust scale-invariant feature matching for optical remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing,* vol. 49, no. 11, pp. 4516-4527, 2011.

[31] A. Sedaghat, and N. Mohammadi, “Uniform competency-based local feature extraction for remote sensing images,” *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 135, pp. 142-157, 2018.

[32]C. G. Harris, and M. Stephens, "A combined corner and edge detector." pp. 10-5244.

[33] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and vision computing,* vol. 22, no. 10, pp. 761-767, 2004.

[34] A. Sedaghat, and H. Ebadi, “Remote sensing image matching based on adaptive binning SIFT descriptor,” *IEEE transactions on geoscience and remote sensing,* vol. 53, no. 10, pp. 5283-5293, 2015.

[35] A. Sedaghat, and N. Mohammadi, “High-resolution image registration based on improved SURF detector and localized GTM,” *International Journal of Remote Sensing,* vol. 40, no. 7, pp. 2576-2601, 2019.

[36] Y. Ye, J. Shan, S. Hao, L. Bruzzone, and Y. Qin, “A local phase based invariant feature for remote sensing image matching,” *ISPRS journal of photogrammetry and remote sensing,* vol. 142, pp. 205-221, 2018.

[37] E. Shechtman, and M. Irani, "Matching Local Self-Similarities across Images and Videos." p. 3.

[38] 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 Transactions on Biomedical Engineering,* vol. 57, no. 7, pp. 1707-1718, 2010.

[39] A. Sedaghat, and A. Alizadeh Naeini, “DEM orientation based on local feature correspondence with global DEMs,” *GIScience & Remote Sensing,* vol. 55, no. 1, pp. 110-129, 2018.

[40] A. Wong, and D. A. Clausi, “ARRSI: Automatic registration of remote-sensing images,” *IEEE Transactions on Geoscience and Remote Sensing,* vol. 45, no. 5, pp. 1483-1493, 2007.

[41] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, “Robust registration of multimodal remote sensing images based on structural similarity,” *IEEE Transactions on Geoscience and Remote Sensing,* vol. 55, no. 5, pp. 2941-2958, 2017.

[42] 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 Transactions on Geoscience and Remote Sensing,* vol. 56, no. 9, pp. 5368-5379, 2018.

[43] J. P. Lewis, "Fast normalized cross-correlation." pp. 120-123.

[44] R. N. Bracewell, and R. N. Bracewell, *The Fourier transform and its applications*: McGraw-Hill New York, 1986.

[45] A. V. Oppenheim, and J. S. Lim, “The importance of phase in signals,” *Proceedings of the IEEE,* vol. 69, no. 5, pp. 529-541, 1981.

[46] M. C. Morrone, and R. A. Owens, “Feature detection from local energy,” *Pattern recognition letters,* vol. 6, no. 5, pp. 303-313, 1987.

[47] S. Fischer, F. Šroubek, L. Perrinet, R. Redondo, and G. Cristóbal, “Self-invertible 2D log-Gabor wavelets,” *International Journal of Computer Vision,* vol. 75, no. 2, pp. 231-246, 2007.

[48] J. Arrospide, and L. Salgado, “Log-Gabor filters for image-based vehicle verification,” *IEEE Transactions on Image Processing,* vol. 22, no. 6, pp. 2286-2295, 2013.

[49] F.-h. Wang, and J.-q. Han, “Iris recognition based on 2D Log-Gabor filtering,” *Journal of System Simulation,* vol. 20, no. 6, pp. 1808-11, 2008.

[50] P. Kovesi, “Phase congruency: A low-level image invariant,” *Psychological research,* vol. 64, no. 2, pp. 136-148, 2000.

[51] J. Fan, Y. Wu, F. Wang, Q. Zhang, G. Liao, and M. Li, “SAR image registration using phase congruency and nonlinear diffusion-based SIFT,” *IEEE Geoscience and Remote Sensing Letters,* vol. 12, no. 3, pp. 562-566, 2015.

[52] D. Dall’Alba, and P. Fiorini, “BIPCO: ultrasound feature points based on phase congruency detector and binary pattern descriptor,” *International journal of computer assisted radiology and surgery,* vol. 10, no. 6, pp. 843-854, 2015.

[53] D. Fan, Y. Ye, L. Pan, and S. Yan, "A remote sensing adapted image registration method based on SIFT and phase congruency." pp. 326-331.

[54] P. Kovesi, "Phase congruency detects corners and edges."

[55] B. Horn, *Robot vision*: MIT press, 1986.

[56] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK: Binary robust invariant scalable keypoints." pp. 2548-2555.

[57] K. Mikolajczyk, and C. Schmid, "Indexing based on scale invariant interest points." pp. 525--531.