Robust Registration of Multimodal Remote Sensing Images With Spectrum Congruency

Jing Huang, Fang Yang, and Li Chai, *Member, IEEE*

***Abstract*—Among the existing registration methods, most feature descriptors are designed with image intensity, gradient information and phase congruency (PC). However, both intensity and gradient are sensitive to image illumination changes, complex intensity differences, noise, etc. Despite the fact that PC is invariant to image illumination and contrast, it does not perform well when images are corrupted with noise and nonlinear radiation distortions. In this article, we propose a novel feature called spectrum congruency (SC), which is robust to noise and variations of image illumination and intensity. SC focuses on exploiting the correlation of the multiscale patches based on their local energy and measures the congruency of the energy distribution in a data-driven transform domain. To demonstrate the superiority of SC, we apply it to multimodal image registration. We construct a histogram-based feature descriptor based on SC, termed as HOSC. Then the HOSC descriptor is integrated with two similarity metrics for multimodal remote sensing image registration. Extensive experimental results on both real and noisy image pairs show that the proposed method presents superior registration accuracy and excellent performance in resisting the nonlinear distortion and noise.**

***Index Terms*—Feature descriptor, local energy, multimodal images, multiscale, nonlinear radiation distortions, registration, spectrum congruency.**

I. Introduction

With the rapid evolution of geospatial information technology, remote sensing images present multimodal forms with various internal characteristics. These multimodal data can provide a complementary information for analysis and interpretation of the region surveyed [1], and have been widely used in comprehensive applications such as modern military surveillance [2], [3], [4], change detection [5], [6], [7], image fusion [8], and 3-D modeling reconstruction [9], [10]. Before fusing the multimodal information, the registration is a prerequisite to align two or more images of roughly the same scene captured by different sensor mechanisms or under different conditions [11]. However, due to the significant differences among modalities and the noise in images, multimodal registration still faces challenges.

Manuscript received 15 November 2022; revised 15 May 2023; accepted 23 May 2023. Date of publication 29 May 2023; date of current version 12 June 2023. This work was supported by the National Natural Science Foundation of China under Grant 62101392 and Grant 62173259. *(Corresponding author: Fang Yang.)*

Jing Huang and Fang Yang are with the Engineering Research Center of Metallurgical Automation and Measurement Technology, Wuhan University of Science and Technology, Wuhan 430081, China (e-mail: jinghuang.work@foxmail.com; yangfang.idif@wust.edu.cn).

Li Chai is with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: chaili@zju.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3281029

Depending on the registration process, most remote sensing image registration methods can be roughly classified into three categories: feature-based, area-based, and their combination [12]. The feature-based methods exploit the features that describe the geometrical or structural characteristic of the image, such as point features, line features, and region features [13]. Then the feature correspondence between the reference and the sensed images is established by a similarity comparison of appropriate feature descriptors.

Feature descriptors are expected to capture a substantial amount of information about the local property of an image and be robust to small deformations or localization errors [14]. Thus, it is essential to extract enough stable and descriptive geometric features. Existing feature descriptors are mainly histogram-based methods designed with image feature map and orientation map. The scale-invariant feature transform (SIFT) [15] is a widely used feature descriptor that is constructed with the histogram of oriented gradient information. Due to its invariance to scale and rotation changes, SIFT has been used for many matching tasks [16], [17], [18]. Inspired by SIFT, many scholars have designed a series of SIFTlike descriptors to improve the matching performance, such as speeded-up robust feature (SURF) [14], scale restriction SIFT (SR-SIFT) [19], uniform robust SIFT [20], local binary descriptor BRIEF [21], and oriented descriptor based on BRIEF (ORB) [22]. However, when these SIFTlike methods are applied to remote sensing images, the matching results are usually unsatisfying. The main reason is that these feature descriptors are designed with gradient or intensity information, which is vulnerable to complex radiometric changes or geometric differences between images. To tackle this problem, Xiang et al. [23] proposed an OS-SIFT method for optical and synthetic aperture radar (SAR) image registration. They used the multiscale ratio of exponentially weighted averages to improve the robustness of speckle noise for SAR images, and applied multiscale Sobel operators to extract features for optical images. Yao et al. [24] proposed a co-occurrence scale space based on co-occurrence filter and developed a new gradient to optimize the impact of nonlinear difference among multimodal images. Hong et al. [25] combined the traditional local binary pattern and the gradient direction information to form a stable descriptor, thus, decreasing the number of mismatched points. However, the feature maps of these improved SIFT-like matching methods are also based on gradient information, which are still not robust for multimodal images.

Later on, many researchers employ phase congruency (PC) to construct feature descriptor for image registration task because it is invariant to the intensity variation of images and consistent with the human visual system [26], [27]. For example, Fu et al.

[28] developed a local feature descriptor by combining the oriented PC information and oriented magnitude binary map to capture the feature properties of local regions. Xiang et al. [29] proposed an SAR-PC model with ratio-based edge detectors to extract spatial properties of optical and SAR images. The SAR-PC model improves the robustness of speckle noise but it retains sensitivity to drastic intensity differences. Li et al. proposed [30] a radiation-variation insensitive feature transform (RIFT) method for multimodal image matching. RIFT applies FAST detector [31] on the maximum moment map of the PC to detect more feature points. And the maximum index map of multiorientation PC sequences is designed to increase the distinction of feature descriptor. Yao et al. [32] designed a histogram of absolute phase consistency gradients (HAPCG) feature descriptor by using magnitude of the maximum and minimum moment of PC and absolute PC orientation. Although these PC-based approaches have achieved encouraging results, the applicability and accuracy of matching are limited because of its essential deficiency. PC is insensitive to image illumination and contrast, but it fails to maintain stable features when images are strongly corrupted with noise and nonlinear radiation distortion. Besides, they generate glitch artifacts when the input image is noisy. This is due to that PC extracts the local energy information by combining the responses of multiscale and multioriented orthogonal filters, which may lead to spurious edges. Therefore, the registration accuracy based on the PC-based descriptors may decrease dramatically when images are corrupted by strong noises, which are very common in remote sensing applications. Moreover, the significant differences among modalities make it difficult to design an appropriate feature descriptor for feature-based multimodal registration.

Different from the feature-based methods, the area-based methods compare the similarity between two images based on the intensity or gradient information of the feature points in a predefined template window. In addition, the area-based methods use the georeferencing techniques as a preprocessing step for coarse registration to remove the obvious translation and rotation differences of the image pairs, so they yield more refined alignment with less errors. The performance of the area-based methods relies heavily on the selection of similarity metrics, which can be computed in the spatial domain and frequency domain [33], [34]. In the spatial domain, the most widely used similarity metrics are the sum of squared differences (SSD), the normalized cross correlation (NCC) and the mutual information (MI). The SSD and NCC are computed directly on image intensity, making them vulnerable to intensity changes and noises and failing in the cases of nonlinear differences and geometric distortions of multimodal images [35]. MI is more robust to nonlinear radiation differences since it measures the statistical dependence between two images and can capture more correlations among pixels [12]. MI-based metrics have been successfully applied in multispectral and multisensor image registration. For example, Chen [36] utilized a new joint histogram estimation algorithm for computing mutual information to register multitemporal remote sensing images. Chen et al. [37] proposed a novel similarity metric for medium-low resolution multisource images, called rotationally invariant regional mutual information (RIRMI). However, the MI-based methods are computationally expensive and very sensitive to the window size for template matching, making them impractical for remote sensing datasets. In the frequency domain, phase correlation is the most popular similarity metric [38], [39], [40], which transforms the image pairs into the Fourier transform domain to obtain phase differences. The phase correlation only considers the phase information, thus, it is insensitive to image content and more robust to the intensity differences and noise. Nevertheless, the frequency-independent noise and geometric deformations across the frequencies make the phase correlation methods perform inaccurate [41]. In a word, the area-based methods are based on the assumption that corresponding image regions have similar intensity contents or patterns and the corresponding performance depend on the similarity metrics. These methods are more adaptable for multispectral images matching, e.g., optical-Infrared images, but they cannot effectively handle the registration of multimodal remote sensing images.

To further improve the registration accuracy and robustness, some studies focus on the combination of feature- and area-based registration methods. They evaluate the similarity of feature descriptors instead of intensity, thus, resisting the nonlinear intensity differences. For example, Gong et al. [42] combined the SIFT descriptor and MI similarity metric to realize a coarse-to-fine registration framework for optical and SAR remote sensing images. In [43], the local self-similarity (LSS) descriptor is integrated as NCC similarity metric to suppress the nonlinear intensity differences among multispectral remote sensing images. Ye and Shen [13] designed a descriptor based on the histogram of oriented PC (HOPC) and develop a novel similarity metric by combining HOPC with NCC similarity to enhance the robustness of multimodal image registration. Later on, Ye et al. [33] presented the channel features of orientated gradients (CFOG) to accelerate the computation efficiency by inducing a 3-D-FFT similarity measures based on SSD and achieves encouraging results. Morrone et al. [44] constructed a novel structural descriptor (SFOC) combined with a fast similarity metric called fast NCC to achieve reliable registration performance. However, as previously mentioned, current feature descriptors are constructed based on gradient or PC, which are sensitive to noise and not effective for multimodal remote sensing registration.

To address the above issue, we propose a novel combination registration framework for multimodal remote sensing registration by introducing a robust feature perception measurement. The new measurement is called spectrum congruency (SC), which is used to describe the edge features of images. Unlike PC, SC is computed through data-driven bases, and does not need to integrate the filter response value from multiple orientations, thus avoiding the glitch artifact. In addition, SC extracts different frequency information by using multiscale patches and can retain all frequency information. Hence, SC is more adaptable to the input image, which makes the detected features more reliable. To the best of our knowledge, there are no results in the literature regarding applying the data-driven method to measure the local energy and PC model. To take advantage of SC, we construct a robust feature descriptor with the histogram of SC (HOSC) to extract more descriptive geometric features for multimodal image registration. The main contributions of this article can be summarized as follows.

1) A novel feature perception method named SC is proposed. SC is computed via data-driven bases, so it is adaptable to the input image. In addition, SC is not only invariant to changes in brightness or contrast as PC, but also robust to noise, which helps generate more stable and descriptive features.

2) An automatic and robust feature descriptor called HOSC is constructed for multimodal registration. HOSC is designed with the histogram of SC magnitude and gradient orientation. We integrate HOSC with robust similarity metrics to deal with the complex radiation distortion and nonlinear intensity differences, and achieve satisfying registration results.
3) Extensive experiments on both noise-free and noisy data demonstrate that, the proposed descriptor HOSC is more effective and robust than the state-of-the-art methods in terms of the evaluation indicators and visual effects.

The rest of this article is organized as follows. Section II introduces the relevant knowledge of PC model. Section IV presents the proposed registration framework based on SC for multimodal remote sensing images. Section V analyzes the parameters and shows the experimental results. Finally, Section VI concludes this article.

II. Phase Congruency

In [45] and [46], the authors found that biologically or physically, the edges and corners in images could be defined as places where the Fourier components are maximally congruent in phase. This phenomenon is called the phase congruency (PC). Venkatesh and Owens [47] found that the points of maximum PC locates at the peaks of local energy and it has been proved that local energy is equal to PC scaled by the sum of the Fourier amplitudes. That is

$$PC(x) = \frac{E(x)}{\Sigma_n A_n(x) + \epsilon} \tag{1}$$

where $A_n(x)$ represents the amplitude of the n_{th} Fourier component and ϵ is a small positive constant to prevent the expression instability. The local energy $E(x)$ of the 2-D image f at position x can be obtained by

$$E(x) = \sqrt{F^2(x) + H^2(x)} \tag{2}$$

where $F(x)$ is the signal without its direct-current (dc) component and $H(x)$ is the Hilbert transform of $F(x)$.

Later, Kovesi [27] improved the PC modal by using the quadrature wavelet filters which enable one to compute the frequency information at a given spatial location. And the local energy can be measured by convolving the signal with the pair of quadrature filters over scales and orientations. Let M^e_{no} and M^o_{no} denotes the even and odd symmetric wavelets at a scale n and orientation o. The corresponding response vector of each quadrature pair of filters is given by

$$[e_{no}(x), o_{no}(x)] = [f(x) * M^e_{no}, f(x) * M^o_{no}]. \tag{3}$$

Thus, local energy can be calculated

$$E(x) = \sqrt{\left(\sum_n \sum_o e_{no}(x)\right)^2 + \left(\sum_n \sum_o o_{no}(x)\right)^2} \tag{4}$$

and the Fourier amplitudes at a given scale n and orientation o is given by

$$A_{no}(x) = \sqrt{e_{no}(x)^2 + o_{no}(x)^2}. \tag{5}$$

Thus, the PC model is defined as:

$$PC(x) = \frac{\sum_o \sum_n W_o(x) \lfloor E(x) - T_o \rfloor}{\sum_o \sum_n A_{no}(x) + \epsilon} \tag{6}$$

where T_o is the noise compensation term estimated by subtracting noise power spectrum in each orientation of local energy. $\lfloor \cdot \rfloor$ denotes that the enclosed quantity is itself if it is positive or zero otherwise. W_o is the weighted function which maintains the significant distribution of frequency and suppresses spurious responses where the spread of filter responses is narrow.

The classical PC model computes the energy and amplitude in the transform domain with oriented filters, such as the Fourier transform [26], the log-Gabor transform [27], and the monogenic signal [48], [49]. The integration of multiple oriented filters over all scales and orientations brings some spurious edges and glitch artifacts, especially in noisy images. Although the noise compensation is considered, the image features and magnitude of PC will be seriously reduced if the noise is strong.

III. Spectrum Congruency

As stated in Section II, the performance of the traditional PC will drop dramatically when images are corrupted by noises. In this section, we will propose a data-driven model to extract frequency information and preserve more edge features of images.

A. SC via Local Energy

Traditionally, the local energy is computed via the integration of response values of pairs of quadrature filters. These quadrature filters are generally fixed bases and created from a mother wavelet to obtain particular frequencies of images. Since fixed bases are not adaptable to input signals, we propose to find the data-driven bases in appropriate transform domain.

The transform domain is embedded in a Hilbert space $\mathbf{H}$ and composed of a set of orthonormal bases $\{\mathbf{v}_n\}_{n=1}^N, \mathbf{v}_n \in \mathbb{R}^N$, where N is the dimension of $\mathbf{H}$. Note that the wavelet transform can be used in the multiscale analysis naturally because the wavelet base is scalable, however, our method is data-driven, and scaling of the bases will lose the complete and orthogonal property of the bases. Therefore we develop a multiscale framework to access the local frequency information of images by scaling the patches around each center pixel. For a certain pixel x of a 2-D image f, we firstly extract a set of image patches centered around this pixel with different sizes $S_1, S_2, \ldots, S_m$ that are sorted in ascending order. These patches can be denoted as $P_1, P_2, \ldots, P_K$, where $P_k \in \mathbb{R}^{\sqrt{S_k} \times \sqrt{S_k}}$. We can pick the coarse frequencies by downsampling these patches to the smallest patchsize S_1

$$P'_k(x) = P_k(x) \downarrow_\rho \tag{7}$$

where $\downarrow$ means the downsampling operation and $\rho = \frac{S_k}{S_1}$ is the downsampling ratio. Hereby, the scaled patches $\{P'_k\}$ are assumed to describe the different frequency components with different scaling ratio. The patches with smallest size P'_1 (the same as P_1) represent the high frequency components because these patches contain more content of the center pixel with its neighborhood. As k increases, the scaled patches P'_k describe the lower frequencies because downsampling process has removed high-frequency information gradually.

The resampling process is similar to the calculation of PC model by using wavelet. The wavelet-based methods pick up the low-, mid-, and high-frequency information according to the scaled filters on the frequency domain, while we extract different frequency components directly from the signal in the spatial domain. In addition, the noise adhering to features can be

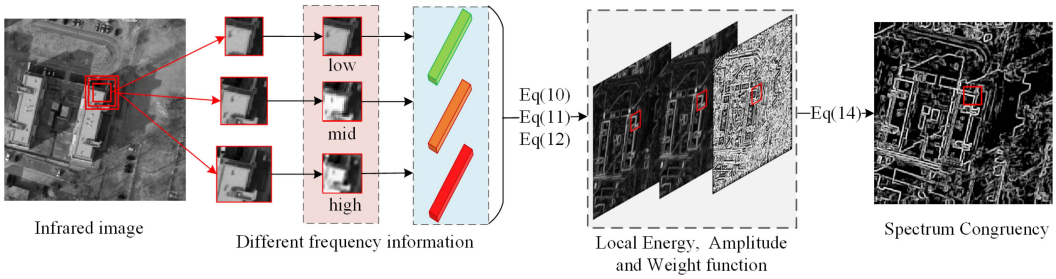


Fig. 1. Flowchart of the proposed SC.

subtracted in resamping process, making the image information more robust to noise.

Suppose the ideal transform domain is spanned by a set of complete orthogonal bases $\{\mathbf{v}_n\}_{n=1}^{S_1}, \mathbf{v}_n \in \mathbb{R}^{S_1}$, then the local energy and amplitude is computed as follows. First, remove the dc component from each patch $P'_{k(x)}$ surrounded at each target pixel x by subtracting their mean value $\overline{P'_k(x)}$, respectively

$$X_k(x) = P'_k(x) - \overline{P'_k(x)}. \tag{8}$$

Then vectorize these patches into a vector denoted as $\{\mathbf{x}_k(x)\}_{k=1}^{K}, \mathbf{x}_k(x) \in \mathbb{R}^{S_1}$, and extract the specific frequency components by projecting each vector to $\{\mathbf{v}_n\}$

$$\mathbf{y}_k(x) = \left[\mathbf{x}_k(x)^T\mathbf{v}_1, \mathbf{x}_k(x)^T\mathbf{v}_2, \ldots, \mathbf{x}_k(x)^T\mathbf{v}_{S_1}\right]^T \tag{9}$$

where $\mathbf{y}_k(x) = [y_k^1(x), y_k^2(x), \ldots, y_k^{S_1}(x)]^T$ means the projection term of the vector from the kth scale, and $y_k^s(x) = \mathbf{x}_k(x)^T\mathbf{v}_s$ is the s_{th} element of $\mathbf{y}_k(x)$. The summation of local energy can be expressed as follows:

$$E(x) = \sqrt{\left(\sum_k y_k^1(x)\right)^2 + \left(\sum_k y_k^2(x)\right)^2 + \cdots + \left(\sum_k y_k^{S_1}(x)\right)^2}. \tag{10}$$

The corresponding local amplitude can be computed as

$$\sum_k A_k(x) = \sum_k \sqrt{\left(y_k^1(x)\right)^2 + \left(y_k^2(x)\right)^2 + \cdots + \left(y_k^{S_1}(x)\right)^2}. \tag{11}$$

To obtain a good localization of features, it is important to suppress superior responses of no significant frequency components. This can be realized by carrying out a sigmoid function to the width of frequencies

$$W(x) = \frac{1}{1 + e^{\beta(c - s(x))}} \tag{12}$$

where β and c control the cutoff value of weight function. The width of frequencies $s(x)$ is defined as

$$s(x) = \frac{1}{M}\left(\frac{\sum_k A_k(x)}{A_{\max}(x) + \epsilon} - 1\right). \tag{13}$$

Let M be the number of scales, and $A_{max}(x)$ be the maximum amplitude at point x on the image. Then the SC based on the multiscale patches is defined as follows:

$$SC(x) = \frac{W(x)\lfloor E(x) - T\rceil}{\sum_k A_k(x) + \epsilon}. \tag{14}$$

The term T is the noise compensation which can be measured as the mean value of local energy response scaled by a small constant, denoted as $T = \alpha\overline{E}$. ϵ is a small constant avoiding a zero denominator. Fig. 1 illustrates the flow process of the proposed SC model.

B. Bases Selection

In this section, we show that the SC of an image defined by (14) is invariant to representations under different domains as long as the orthonormal bases are used.

Let $\{\mathbf{v}_n\}_{n=1}^{N}$, $\{\mathbf{u}_n\}_{n=1}^{N}$ be two arbitrary sets of orthonormal bases of a domain $\Omega \subset \mathbb{R}^N$. For a set of vectors $\mathcal{X} = \{\mathbf{x}_k\}_{k=1}^{K}, \mathbf{x}_k \in \Omega$, the energy and amplitude of its projection on $\{\mathbf{v}_n\}$ and $\{\mathbf{u}_n\}$ are the same

$$\begin{cases} E(\mathcal{X}_{\mathbf{v}_n}) = E(\mathcal{X}_{\mathbf{u}_n}) \\ A_k(\mathcal{X}_{\mathbf{v}_n}) = A_k(\mathcal{X}_{\mathbf{v}_n}) \end{cases} \tag{15}$$

where $E(\mathcal{X}_{\mathbf{v}_n})$ and $E(\mathcal{X}_{\mathbf{u}_n})$ correspond to the energy of $\mathcal{X}$ on $\{\mathbf{v}_n\}$, and $\{\mathbf{u}_n\}$, respectively. $A_k(\mathcal{X}_{\mathbf{v}_n})$ and $A_k(\mathcal{X}_{\mathbf{v}_n})$ are the amplitude at scale k of $\mathcal{X}$ on $\{\mathbf{v}_n\}$, and $\{\mathbf{u}_n\}$, respectively.

SC is invariant for the signal represented by any orthonormal bases $\{\mathbf{v}_n\}_{n=1}^{N}$. Hence, in our case, to facilitate the computation, we use the column vectors of the identity matrix $I_d \in \mathbb{R}^{S_1 \times S_1}$ as the bases. The detailed description of SC is presented in Algorithm 1.

C. Antinoise Performance of SC

To verify the effectiveness of proposed method, we compare SC with gradient and traditional PC on both synthetic images and real remote sensing images.

1) Synthetic Image: Fig. 2 demonstrates the feature results on a synthetic image, corrupted by a mixture of Salt & Pepper noise and Gaussian noise. The corrupted ratio of Salt & Pepper noise is $d = 0.01$, and the standard deviations δ of Gaussian noises ranges from 5 to 20 with incremental value of 5. With the increasing of the standard deviation δ, the gradient map becomes monotonically unclearer. Specially, when δ reaches 20, the gradient feature of the triangular is nearly invisible since the triangular has lower contrast. PC is more robust than gradient but is also affected by noise and generates the glitch artifact along the edges. In addition, we can see from Fig. 2(c) that the contour of the triangular and circle becoming fuzzier with the increase of noise intensity. This means that the

Algorithm 1: Framework of SC via Multiscale Local Patches.

Input:
Image f, PatchSize=$\{S_1, S_2, S_3, \ldots, S_K\}$;
Output:
Feature map SC;
1: **for** each pixel x in f **do**
2: Extract K patches centered at x: $P_1(x)$, $P_2(x)$, $P_3(x)$,..., $P_K(x)$
3: Remove the dc components of each patch
4: **for** each S_k in PatchSize **do**
5: Downsample $P_k(x)$ to form a new patch $P_{S_k new}(x)$ with the same size of S_1
6: Project each patch $P_{S_k new}(x)$ to a set of complete and orthogonal bases $\{\mathbf{v}\}_n$;
7: Compute E_k (10) and A_k, respectively, according to (11)
8: **end for**
9: Compute the spectrum congruency $SC(x)$ via (14)
10: **end for**
11: **return** SC

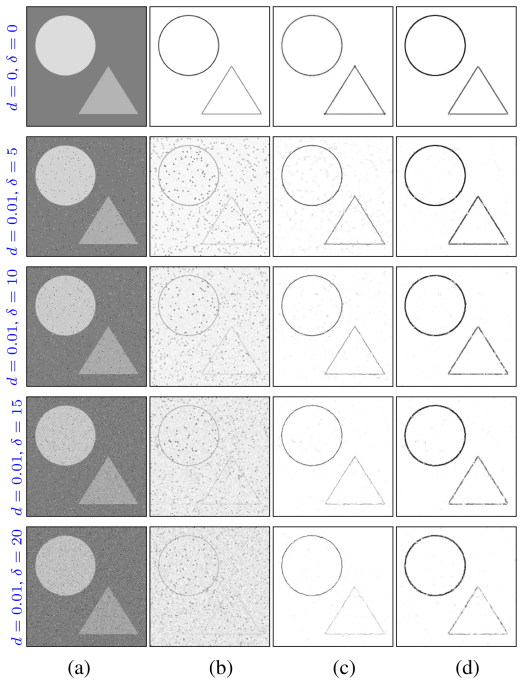


Fig. 2. From left to right: (a) Original image and the noise images with $d = 0.01$ and δ varies from 0, 5, 10, 15, 20. (b)–(d) Corresponding edge maps obtained by gradient (b), PC (c), and SC (d), respectively.

PC value decreases gradually as noise increases. SC is least affected by noise and the edge features detected by SC are more stable. This result shows that our proposed SC holds a more robust antinoise performance compared with PC in noisy instances.

2) Remote Sensing Images: We test the performance of SC on a real infrared remote sensing image. The original infrared image is about an urban area with buildings. The noisy image is obtained by corrupting the real image by a mixture of Gaussian and Salt & Pepper noise with Gaussian standard deviation $\delta = 5$ and the percentage of corrupted ratio $d = 0.05$. Fig. 3 depicts the feature results obtained by gradient, PC and SC.

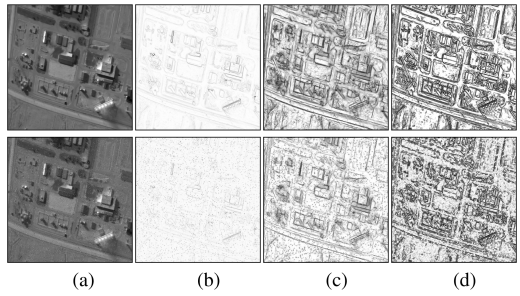


Fig. 3. Edge detection results on the real (Top) and noisy infrared remote sensing image (bottom), from left to right. (a) Source images. (b) Gradient map. (c) PC map. (d) SC map.

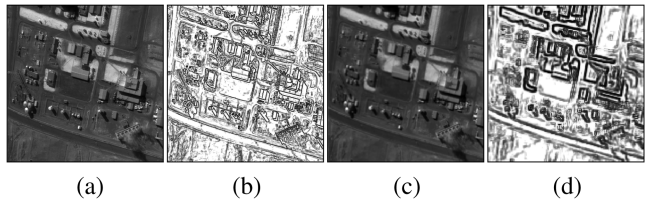
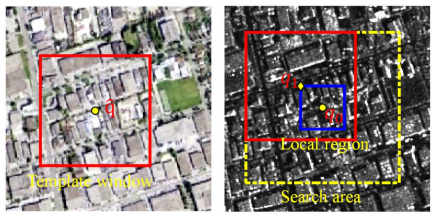


Fig. 4. SC results on the optical image with varying resolution, from left to right: (a) Source image. (b) SC map of source image. (c) Down-sampling image $\rho = 0.3$. (d) SC map of (c).

For the real infrared image, we observe that PC and SC are more robust to image illumination and contrasts than the gradient magnitude. However, as mentioned above, the edge features detected by PC contains some false shadow and glitch artifact in the flat region due to the low resolution and complex intensity changes of the infrared image. The proposed detector extracts more smooth and descriptive geometric features compared with PC. For the noisy image, the gradient and PC map are easily affected by the Salt & Pepper noise since they are sensitive to significant complex intensity and noise. Although SC is a little affected by the noise, it still provides much more essential and complete structural information of images.

Fig. 4 depicts the SC results of an optical image (a) and its low resolution version (c) by compressing the image with a down-sampling ratio $\rho = 0.3$. It can be seen that the low resolution image loses some of its detail and clarity, resulting in blurred edges. Despite the low resolution of the image, SC is still able to extract smooth and complete contours of the structural shapes, as shown in Fig. 4(d). This means that even with lower pixel density, SC can effectively capture the overall shape and boundary of the object in the image, while maintaining its smoothness and completeness.

SC is based on the local energy of multiscale patches so it does not need to consider the influence of the filter orientations. This avoids the integration of values from multiple orientations, thus eliminating the glitch artifact and providing a much simpler way to measure the edge strength. Besides, the multiscale patches based on downsampling operation can reduce the spurious noise in images, making SC highly robust to noise. The unique transform domain with data-driven bases make SC more suitable to perceive the image perception and preserve more descriptive geometric features. Therefore, we explore SC to construct the feature descriptor to improve the registration accuracy of multimodal remote images.



(a) (b)

Fig. 5. Template window and search region in similarity evaluation processing. (a) Reference image. (b) Sensed image. Red box: Template window, blue box: Local region, yellow dashed box: Search region.

IV. Multimodal Registration Framework Based on Spectrum Congruency

Image registration aims to find the optimal spatial geometric mapping from the sensed image S to reference image R, and it can be expressed as

$$E\left(\hat{T}\right) = arg\max_{\hat{T}}\left[\Psi(F_R, \hat{T}(F_S))\right] \tag{16}$$

where F_R and F_S denote the set of feature points in the reference image and sensed image, respectively. $\hat{T}$ denotes the spatial geometric mapping and $\Psi(\cdot)$ is the structural similarity metric between feature points. The feature descriptor and similarity metric play significant roles in image registration. In this section, we propose a structure descriptor called HOSC and introduce the corresponding similarity metric.

A. HOSC: Structural Feature Descriptor

Based on SC, a novel structural descriptor named HOSC for multimodal remote sensing images registration is proposed. HOSC is a local histogram-based descriptor to identify local object appearance and shape by the projection of intensity feature orientation. It combines the histogram of SC magnitude and gradient orientation to represent the local image structure. Fig. 6 presents the whole process of constructing the HOSC feature descriptor. The detailed process is as follows.

1) For each pixel, we first extract the feature map and orientation map of local region and divide them into overlapping blocks. Each block consists of $N_b \times N_b$ cells region and each cell contains $N_c \times N_c$ pixels.
2) The orientation of each cell is partitioned into several bins and weighted by SC amplitude with a trilinear interpolation. We normalize the histogram for all cells in one block by the ℓ_2 norm to eliminate the effect of illumination changes. The feature vector is vectorized from the block.
3) The feature vectors of all blocks (or pixels) are arranged to form a 3-D orientation histogram.

Next, we evaluate the similarity between two images on the basis of HOSC structural properties to detect corresponding matching points.

B. Similarity Metric Based on Structural Properties

The matching method is a template-based framework, which defines a template in the reference image and then finds the optimal correspondence in the local search region of sensed image by evaluating similarity measures. Fig. 5 presents the template window and search region in similarity evaluation processing. Suppose that the point $\hat{q}$ is a feature point in the reference image and the point q_0 is the candidate point in the sensed image. We need to find the target matching point in the local region centered around q_0 by calculating the feature representations similarity of each pixel in local region. As shown, the yellow diamond point q_1 denotes the first point in local region. The correlation between $\hat{q}$ and q_1 can be measured by the similarity with HOSC within template window centered around them.

As mentioned above, combining the feature descriptor with similarity metrics can help resist the nonlinear intensity differences. In this article, we use two matching metrics (NCC and the FFT-SSD) to evaluate the similarity of HOSC descriptor for registration, denoted as HOSCncc and FHOSC, respectively. The HOSCncc is defined as

$$HOSC_{ncc} = \frac{\sum_x (R_A(x) - \bar{R}_A)(R_B(x-b) - \bar{R}_B(x-b))}{\sqrt{\sum_x (R_A(x) - \bar{R}_A)^2 \sum_x (R_B(x-b) - \bar{R}_B(x-b))^2}} \tag{17}$$

where $R_A(x)$ and $R_B(x-b)$ denote HOSC descriptor of the template window A and B at location x and $x-b$, and b is the translated vector over local region between R_A and R_B. $\bar{R}_A$ and $\bar{R}_B$ are the means of R_A and R_B.

The FFT-SSD is the SSD similarity metric calculated by using 3-D FFT, which can improve computational efficiency. The FHOSC is given by

$$FHOSC = \left\{3DF^{-1}[3DF(R_A(x)) \cdot 3DF^*(R_B(x-b))]\right\} \tag{18}$$

where $3-DF$, $3-DF^{-1}$, and $3-DF^*$ denote the forward, inverse, and the complex conjugate of 3-D FFT, respectively.

C. Proposed Registration Framework Based on HOSC

Before registration, the reference image and the sensed image are coarsely rectified using the georeferencing technique and resampled to the common spatial coordinate system. This can help eliminate obvious translation and rotation differences between multimodal images. Then the refinement process generally consists of four steps—feature point detection, feature matching, outlier elimination, and image rectification.

1) Feature point detection: Apply the block-Harris detector [43] to obtain the evenly distributed feature points in the reference image. The Harris detector response values are ranked in a descending order and the top K points are identified as feature points.
2) Feature matching: For each feature point in the reference image, define a local search region in the sensed image based on the georeferencing information. Compare each pixel in the searching window to find the best matching point according to the similarity metric integrated with the proposed feature descriptor. The corresponding points are regarded as a pair of control points (CPs).
3) Outlier elimination: Dislodge the outliers using the global consistency check method [50], and remove the mismatching CPs by the iterative refining procedure.
4) Image rectification: Estimate the transformation model and rectify the sensed image.

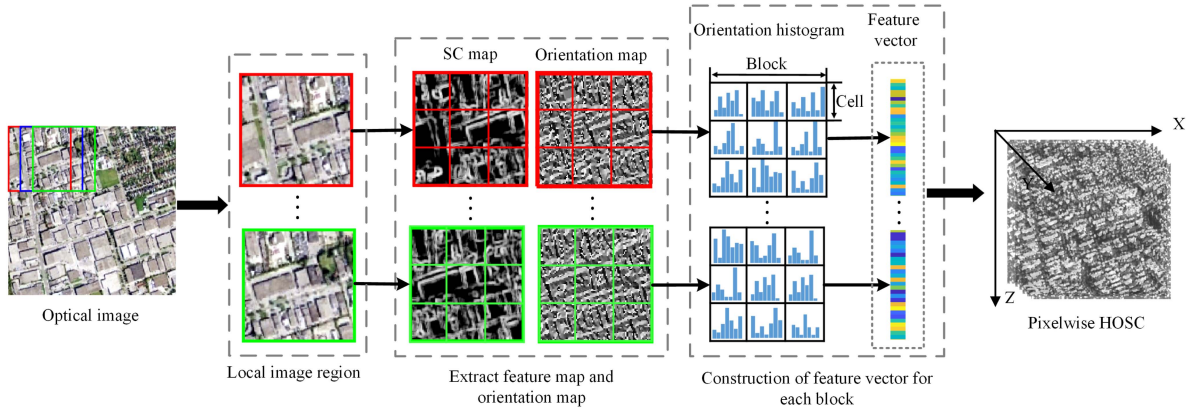


Fig. 6. Flowchart of the construction of HOSC.

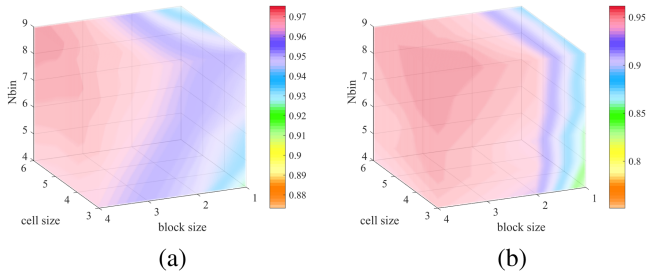


Fig. 7. Average CMR values versus the number of cells, blocks, and orientation bins of FHOSC (a) and HOSCncc (b).

V. Experimental Results and Analysis

In this section, the performance of the proposed feature descriptor are demonstrated on both noise-free and noisy multimodal remote sensing images. We evaluate and discuss the registration results compared with some state-of-the-art methods.

A. Experimental Setting

1) Datasets: In the experiment, five modalities of sensing images are employed. These images are used to form four different pairs as optical-infrared, Lidar-optical, optical-SAR, and optical-map. These images are captured by different imaging mechanisms with different resolution and exhibit diverse land covers such as urban, rivers, and flat areas. As mentioned before, the two images of each pair have been coarsely rectified and resampled to the same ground sample distance (GSD) to eliminate obvious scale, rotation, and translation differences. However, these images still have significant nonlinear radiometric differences and complex image intensity fields, bringing substantial challenges for matching tasks. The detailed information of these multimodal images is list in Table I.

2) Implementation Details: In terms of feature detection among multimodal images, three patch scales are selected to calculate SC: 3×3, 5×5, and 7×7. The smallest scale patches are set to be 3×3 that can sufficiently capture local neighborhood information of pixels. We set $\alpha = 0.2$ for the threshold in (14) to estimate noise energy in this article. In the feature points detection processing, 300 uniformly distributed interesting points are detected by block-based Harris detector in the reference image. The error threshold is set to be 1.5 pixels to eliminate the CPs with large errors. When constructing the structural descriptor, the local region size is set to be equal to the block size. In this way, the feature vector of each block denotes the structure information of the one target point, providing a pixelwise representation of the image. In template matching processing, according to previous literature [13], [33], [51], the correct matching ratio (CMR) value gets better with the increase of the template size and it can achieve good performance when the template size is around 100×100. However, the higher CMR value is at the cost of higher computational complexity and run time. Hence, to balance the cost and performance, we set the template window 100×100 pixels and the search region 20×20 pixels. In addition, we adopt the template matching scheme designed in [13] to collect the feature descriptors at the interval of five pixels.

3) Evaluation Criteria: We analyze the performance of the proposed registration framework via three indices: the number of correct matches (NCM), correct matching ratio (CMR), and matching accuracy. The CMR is defined as: CMR = NCM/NM, NM is the total number of CPs. 40–60 evenly distributed CP pairs are selected manually to assess the matching accuracy in terms of the rmse of these points.

For a set of CP pairs $(x_i; x_i'), i = 1, \ldots N_p$, an affine or projective transformation is used to calculate the corresponding point of x_i, denoted as $\hat{x}_i$. The rmse of these CPs can be calculated as

$$RMSE = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} \|x_i' - \hat{x}_i\|^2}. \tag{19}$$

4) Parameter Analysis: The proposed feature descriptor is related to three parameters: the block size $N_b \times N_b$, the cell size $N_c \times N_c$, and the gradient orientation bins O. Fig. 7 shows the average CMR values versus the parameters for FHOSC and HOSCncc tested on ten pairs of multimodal remote sensing images. It can be observed that both for FHOSC and HOSCncc, the average CMR value increases gradually as the number of orientation bins and block size increases. However, the FHOSC and HOSCncc behave differently when the cell size changes, i.e., the average CMR value of FHOSC is highest when $N_c = 6$, but the average CMR value of HOSCncc achieves best when $N_c = 3$. Since the bigger cell size will lead to more computation time, to make a tradeoff between calculating efficiency and CMR value, we set $N_b = 4$, $N_c = 3$, and $O = 8$ as default parameters.

TABLE I
DETAILED INFORMATION OF MULTIMODAL IMAGES USED IN EXPERIMENT

Category	Case	Image Source	Size	GSD	Data	Location
Optical-Infrared	1	Landsat optical Landsat infrared	500×485 500×485	N/A N/A	N/A N/A	Land area
	2	Daedalus optical Daedalus infrared	512×512 512×512	0.5m 0.5m	04/2000 04/2000	Urban area
LiDAR-Optical	3	LiDAR intensity WorldView 2 visible	621×617 621×621	2m 2m	10/2010 10/2011	Urban area
	4	LiDAR depth WorldView 2 visible	524×524 524×524	2.5m 2.5m	06/2012 06/2012	Urban area
Optical-SAR	5	Google Earth TerraSAR-X	800×800 818×800	0.92m 0.92m	N/A N/A	Road area
	6	Google Earth TerraSAR-X	628×618 628×618	3m 3m	03/2009 01/2008	Urban area
	7	Landsat 5 TM band 3 TerraSAR-X	600×600 600×600	30m 30m	05/2007 03/2008	Suburban area
	8	Google Earth GaoFen-3	528×524 534×524	3m 3m	06/2020 06/2020	Field area
Optical-Map	9	Google Earth Google Earth	550×550 550×550	N/A N/A	N/A N/A	Park area
	10	Google Earth Google Earth	600×600 600×600	N/A N/A	N/A N/A	River area

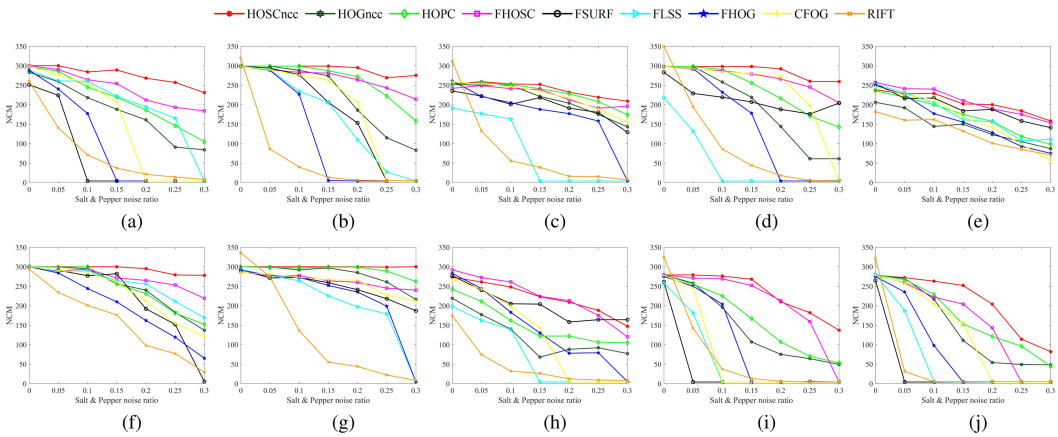


Fig. 8. NCM values of descriptor similarity metric versus different intensity of Salt & Pepper noise for real multimodal images. (a) Case 1. (b) Case 2. (c) Case 3. (d) Case 4. (e) Case 5. (f) Case 6. (g) Case 7. (h) Case 8. (i) Case 9. (j) Case 10.

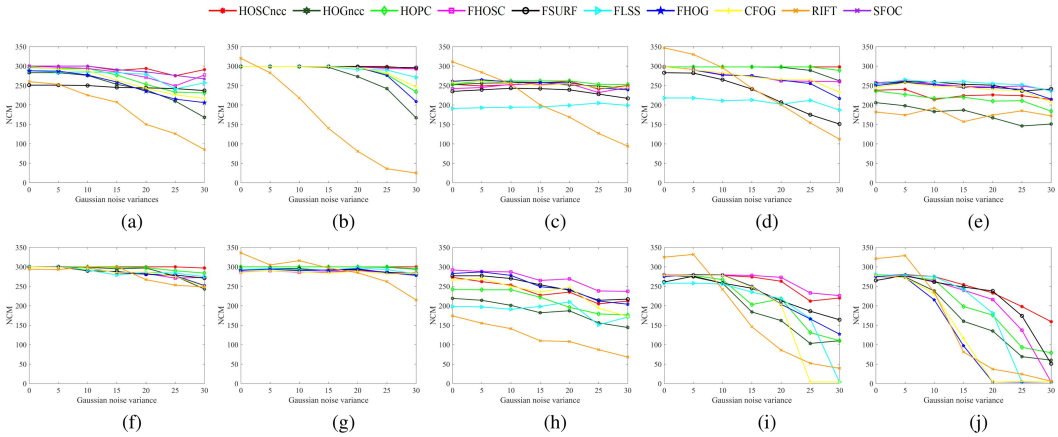


Fig. 9. NCM values of descriptor similarity metric versus different standard deviation of Gaussian noise for real multimodal images. (a) Case 1. (b) Case 2. (c) Case 3. (d) Case 4. (e) Case 5. (f) Case 6. (g) Case 7. (h) Case 8. (i) Case 9. (j) Case 10.

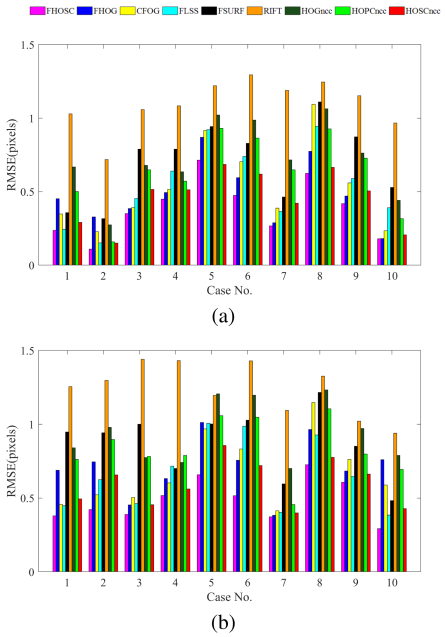


Fig. 10. RMSEs of different feature descriptor. (a) Real multimodal images results. (b) Noisy multimodal images results.

B. Analysis of Feature Descriptor

To test the robustness of feature descriptors, we analyze the performance under Salt & Pepper noise with noise different intensity d and Gaussian noise with standard deviation δ, respectively. We compare the NCM results of the proposed feature descriptor with some the state-of-the-art methods, including CFOG, FHOG, FLSS, and FSURF in terms of 3-DFFT-SSD similarity metric [33]; HOGncc and HOPCncc in terms of NCC similarity metric [13] and RIFT [30].

1) Analysis of Salt and Pepper Noise Sensitivity: Fig. 8 shows the NCM value of these descriptors versus Salt & Pepper noise with a range of $d \in [0, 0.3]$. As can be seen, HOSCncc can extract more stable CPs and have the highest NCM value for almost all image sets and FHOSC outperforms the other 3-DFFT-SSD similarity metrics. Feature descriptors in terms of NCC are more robust to Salt & Pepper noise than that of 3-DFFT-SSD since 3-DFFT-SSD is calculated by FFT transform which is easily affected by Salt & Pepper noise.

For the noise-free multimodal images pairs ($d = 0$), CFOG, FHOG, HOPCncc perform comparably to FHOSC and HOSCncc, especially for image pairs with less radiation deformation, such as Optical-Infrared pairs (Cases 1–2), Lidar-optical pairs (Cases 3-4), and optical-map pairs (Cases 9–10). As d increases, CFOG, FHOG, FLSS, FSURF, and HOGncc fail to extract enough common features among multimodal images. This is because these descriptors based on gradient or image intensity are more likely to suffer from Salt & Pepper noise. PC is more robust than gradient so that HOPCncc can get the higher NCM value compared with HOGncc. However, PC can not keep its property when noise level becomes higher. RIFT can extract more points on real multimodal images since it takes both the corner features and PC edges into consideration so that it ensures the quantity of the feature points. Nevertheless, the NCM of RIFT is easily affected by noise, i.e., when the noise is strong, the NCM drops dramatically. On the contrary, FHOSC and HOSCncc can resist nonlinear radiometric differences and Salt & Pepper noise due to that SC is robust to noise.

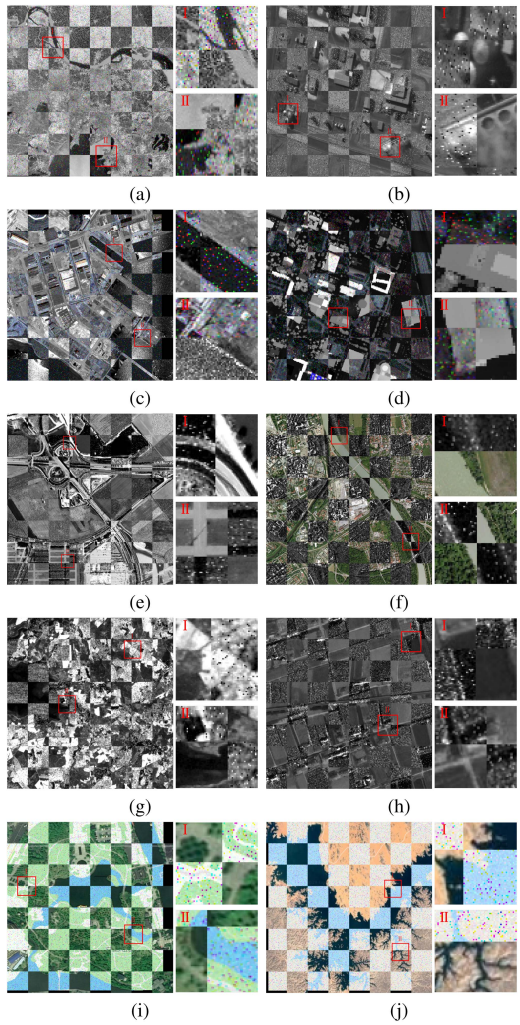


Fig. 11. Fusion results of noisy multimodal images. (a) Case 1. (b) Case 2. (c) Case 3. (d) Case 4. (e) Case 5. (f) Case 6. (g) Case 7. (h) Case 8. (i) Case 9. (j) Case 10.

2) Analysis of Gaussian Noise Sensitivity: Fig. 9 represents the NCM result of feature descriptors versus Gaussian noise with a range of $\delta \in [0, 30]$. It can be observed that the 3-D-FFT based descriptor can extract enough feature points when the level δ is low or the image intensity does not change drastically, such as Case 2 and Case 7. CFOG can better resist Gaussian noise compared with other gradient-based methods since CFOG is weighted by a 3-D Gaussian kernel instead of a triangular kernel, making it more robust to Gaussian noise. When δ increases, CFOG performs worse than FHOSC, especially when images have significant nonlinear intensity differences intrinsically. FLSS, FSURF, and FHOG are not stable for these multimodal images since they are sensitive to noise or image intensity changes. For NCC-based feature descriptor, HOSCncc shows the highest NCM result followed by HOPCncc and HOGncc for almost all multimodal images. The NCM curve of RIFT decreases sharply compared with other 3-D-FFT based and NCC-based methods. One reason is that PC is not robust

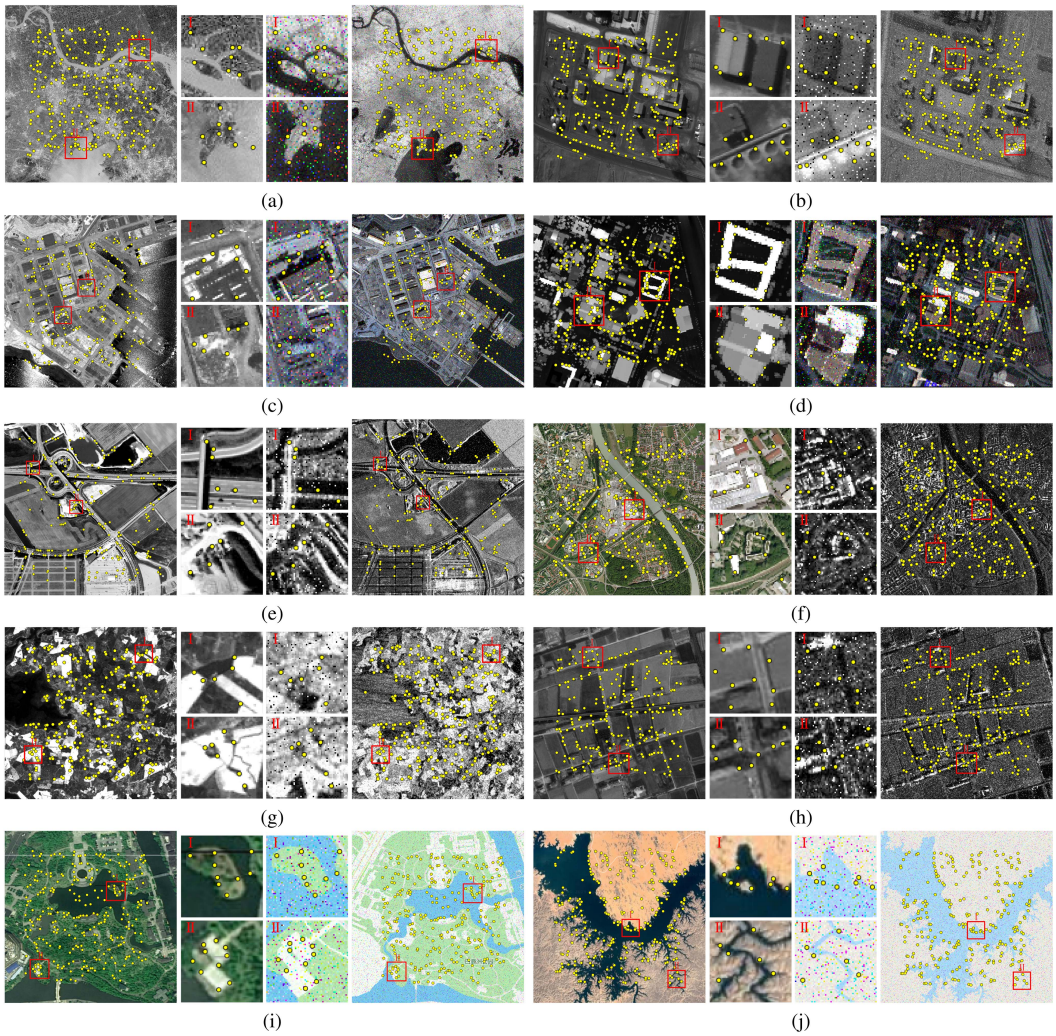


Fig. 12. CPs detection result of FHOSC. (a) Case 1. (b) Case 2. (c) Case 3. (d) Case 4. (e) Case 5. (f) Case 6. (g) Case 7. (h) Case 8. (i) Case 9. (j) Case 10.

to noise and the other one is RIFT uses a global pointwise matching strategy, which would generate mismatches. In addition, we compare our method with a more recent approach, i.e., SFOC [44]. To testify the antirobustness of our method and SFOC, we conduct a comparison with SFOC according to the same experimental settings and results described in [44]. The comparison result of Case 1 is shown in Fig. 9(a), the SFOC uses an NCC-based similarity, and we can see that HOSCncc (red curve) has comparable or better performance than SFOC (purple curve).

In summary, the proposed feature descriptors based on SC feature are more robust and effective for multimodal remote sensing registration.

3) Analysis of Matching Accuracy: Fig. 10 presents the comparison of matching accuracy of both noise-free and noisy multimodal remote sensing images. According to the NCM results, We add a mixture of Gaussian and Salt & Pepper noise with $\delta = 10$ and $d = 0.05$ to the sensed images of optical-infrared, Lidar-optical, and optical-SAR image pairs, and only add Gaussian noise with $\delta = 10$ for the sensed images of optical-map pairs to ensure each method can detect enough points.

We can see that, RIFT gives the worst results in almost all cases for both real and noisy image pairs. Different from other template-based methods, RIFT is a pointwise feature matching method and finds the corresponding candidates by calculating the minimum similarity that traverses the whole set of features points. The search area of RIFT is much larger than the template-based methods, so RIFT mismatches more points and returns lower accuracy. In terms of 3-D-FFT similarity methods, CFOG and FHOG can handle images without complex radiation differences or less polluted by noise, such as Lidar-optical image pairs (Cases 3–4) and optical-map pairs (Cases 9–10). As shown, CFOG performs slightly better than FHOG, FLSS and FSURF in these image pairs. This is because CFOG adopts 3-D Gaussianlike kernel to obtain the feature channels and resist the influence of Gaussian noise. However, CFOG also fails to cope with the issue of Pepper & Salt noise as well as the complex distortions existing in Optical-SAR image pairs. FSURF is easily influenced by nonlinear differences in almost all cases and the matching result is unsatisfying. FLSS performs better on optical-infrared image pairs (Case 1–2), but it fails for other modalities. FLSS is constructed based on image intensity so it can cover some detailed textures of images. On the other hand, it is easily affected by image intensity variance. In terms of NCC-based methods, HOPCncc achieves the lower rmses results compared with HOGncc, but it also performs poorly.

By contrast, FHOSC and HOSCncc achieve lower rmse results and are more stable to deal with multimodal images registration. Although HOSCncc can get higher NCM than FHOSC, FHOSC can achieve lower rmses and get more accurate registration results. The main reason is that the histogram of FHOSC is more distinguishable, then it can achieve more accurate matches for multimodal images. Therefore, we use FHOSC to present the registration result in the following experiment.

C. Registration Result

We depict the visual registration results to examine the performance of our proposed matching method for multimodal remote sensing images. Fig. 12 shows the corresponding CPs result extracted by FHOSC. The multimodal remote sensing images are polluted with the mixture of Gaussian noise and Pepper & Salt noise with $\delta = 10$ and $d = 0.05$. It can be seen that these images have significant nonlinear radiation distortions and the sensed images are greatly corrupted by noise. While FHOSC can detect sufficient evenly distributed CPs and these points hold precise positioning. Fig. 11 shows the fusion results of image pairs in checkerboard mosaic and enlarged subimages. It can be observed that these multimodal images are all well aligned and present a good matching performance. This result verifies the robustness and effectiveness of our proposed registration method.

D. Time Complexity

The proposed registration method mainly involves three steps: calculating SC, constructing the HOSC, and computing the similarity between the feature points. For an image $f \in \mathbb{R}^{N\times N}$, the time complexity of the first two steps is $\mathcal{O}(N^2)$ and $\mathcal{O}(N^2)$, respectively. Regarding the computation of similarity, two similarity metrics are considered, i.e., NCC and FFT-SSD. In Section V-A2, 300 uniformly distributed feature points are used in the reference image, and the search window in the sensed image is of size 20×20. For each pair of points to be registered, the dimension of the feature point in a template window is $n_v = d_f \times 100 \times 100$, where $d_f = 128$ is the dimension of HOSC. So the time complexity of using NCC is $\mathcal{O}(300 \times 400 \times n_v^2)$, and of using FFT-SSD is $\mathcal{O}(300 \times 400 \times n_v \times \log n_v)$. Thus, no matter which similarity is used, in this article, the most time-consuming step is still the similarity computation. The main difference between our method and the other methods lies in that our method uses the newly proposed SC, and the time complexity of other methods also depends mainly on similarity computation. Although the time complexity of SC $\mathcal{O}(N^2)$ is slightly higher than that of PC $\mathcal{O}(NlogN)$, it hardly affects the overall time complexity of the whole process.

VI. Conclusion

In this article, we propose a novel feature called SC based on local energy of multiscale patches. SC is consistent with the human visual system on perceiving the features, invariant to image illumination contrast and robust to noise. Besides, SC can encode the stable feature structure benefited from the data-driven transform in multiscale framework. We apply SC to extract features of multimodal images and propose a robust feature descriptor HOSC based on SC magnitude and gradient orientation. FFT-SSD similarity and NCC similarity are integrated with HOSC to handle the complex radiation distortion and nonlinear intensity differences between multimodal images. Extensive experiments demonstrate that, compared with state-of-the-art multimodal registration methods, the proposed method presents superior registration accuracy, especially for the strongly nonlinear distortion and noisy cases. In addition, the FFT-SSD similarity is more effective than NCC similarity, making FHOSC exhibit better registration results compared with HOSCncc in matching accuracy and calculation efficiency.

References

[1] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[2] B. R. Abidi, N. R. Aragam, Y. Yao, and M. A. Abidi, "Survey and analysis of multimodal sensor planning and integration for wide area surveillance," *ACM Comput. Surv.*, vol. 41, no. 1, pp. 1–36, 2009.

[3] T. D. Räty, "Survey on contemporary remote surveillance systems for public safety," *IEEE Trans. Syst., Man, Cybern., Part C (Appl. Rev.)*, vol. 40, no. 5, pp. 493–515, Sep. 2010.

[4] S. Liu, H. Liu, V. John, Z. Liu, and E. Blasch, "Enhanced situation awareness through CNN-based deep multimodal image fusion," *Opt. Eng.*, vol. 59, no. 5, 2020, Art. no. 053103.

[5] R. Touati, M. Mignotte, and M. Dahmane, "Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based Markov random field model," *IEEE Trans. Image Process.*, vol. 29, pp. 757–767, 2019.

[6] R. Touati, M. Mignotte, and M. Dahmane, "Anomaly feature learning for unsupervised change detection in heterogeneous images: A deep sparse residual model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 588–600, 2020.

[7] S. Chirakkal, F. Bovolo, A. R. Misra, L. Bruzzone, and A. Bhattacharya, "A general framework for change detection using multimodal remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10665–10680, 2021.

[8] Y. Sun, Z. Fu, C. Sun, Y. Hu, and S. Zhang, "Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5404418.

[9] K. Makantasis, A. Doulamis, N. Doulamis, and M. Ioannides, "In the wild image retrieval and clustering for 3D cultural heritage landmarks reconstruction," *Multimedia Tools Appl.*, vol. 75, pp. 3593–3629, 2016.

[10] M. Cao, H. Gao, and W. Jia, "Stable image matching for 3D reconstruction in outdoor," *Int. J. Circuit Theory Appl.*, vol. 49, no. 7, pp. 2274–2289, 2021.

[11] L. G. Brown, "A survey of image registration techniques," *ACM Comput. Surv.*, vol. 24, no. 4, pp. 325–376, 1992.

[12] R. Feng, H. Shen, J. Bai, and X. Li, "Advances and opportunities in remote sensing image geometric registration: A systematic review of state-of-the-art approaches and future research directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 120–142, Dec. 2021.

[13] Y. Ye and L. Shen, "HOPC: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 9–16, 2016.

[14] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

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

[16] X. Hu, Y. Tang, and Z. Zhang, "Video object matching based on SIFT algorithm," in *Proc. Int. Conf. Neural Netw. Signal Process.*, 2008, pp. 412–415.

[17] F. Alhwarin, C. Wang, D. Ristić-Durrant, and A. Gräser, "Improved SIFT-features matching for object recognition," in *Proc. Visions Comput. Sci.-BCS Int. Academic Conf.*, 2008, pp. 179–190.

[18] S. Malathi and C. Meena, "Partial fingerprint matching based on SIFT features," *Int. J. Comput. Sci. Eng.*, vol. 2, no. 4, pp. 1411–1414, 2010.
[19] Z. Yi, C. Zhiguo, and X. Yang, "Multi-spectral remote image registration based on SIFT," *Electron. Lett.*, vol. 44, no. 2, pp. 107–108, 2008.
[20] A. Sedaghat, M. Mokhtarzade, and H. Ebadi, "Uniform robust scale-invariant feature matching for optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4516–4527, Nov. 2011.
[21] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a local binary descriptor very fast," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, Jul. 2012.
[22] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.
[23] Y. Xiang, F. Wang, and H. You, "OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3078–3090, Jun. 2018.
[24] Y. Yao, Y. Zhang, Y. Wan, X. Liu, X. Yan, and J. Li, "Multi-modal remote sensing image matching considering co-occurrence filter," *IEEE Trans. Image Process.*, vol. 31, pp. 2584–2597, 2022.
[25] Y. Hong, C. Leng, X. Zhang, Z. Pei, I. Cheng, and A. Basu, "HOLBP: Remote sensing image registration based on histogram of oriented local binary pattern descriptor," *Remote Sens.*, vol. 13, no. 12, 2021, Art. no. 2328.
[26] P. Kovesi, "Image features from phase congruency," *Videre: J. Comput. Vis. Res.*, vol. 1, no. 3, pp. 1–26, 1999.
[27] P. Kovesi, "Phase congruency: A low-level image invariant," *Psychol. Res.*, vol. 64, no. 2, pp. 136–148, 2000.
[28] Z. Fu, Q. Qin, B. Luo, H. Sun, and C. Wu, "HOMPC: A local feature descriptor based on the combination of magnitude and phase congruency information for multi-sensor remote sensing images," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1234.
[29] Y. Xiang, R. Tao, F. Wang, H. You, and B. Han, "Automatic registration of optical and SAR images via improved phase congruency model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5847–5861, 2020.
[30] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2019.
[31] E. Tola, V. Lepetit, and P. Fua, "A fast local descriptor for dense matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[32] Y. Yao, Y. Zhang, Y. Wan, X. Liu, and H. Guo, "Heterologous images matching considering anisotropic weighted moment and absolute phase orientation," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 46, no. 11, pp. 1727–1736, 2021.
[33] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, "Fast and robust matching for multimodal remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9059–9070, Nov. 2019.
[34] B. Zhu, L. Zhou, S. Pu, J. Fan, and Y. Ye, "Advances and challenges in multimodal remote sensing image registration," *IEEE J. Miniaturization Air Space Syst.*, vol. 4, no. 2, pp. 165–174, Jun. 2023.
[35] J. N. Sarvaiya, S. Patnaik, and S. Bombaywala, "Image registration by template matching using normalized cross-correlation," in *Proc. Int. Conf. Adv. Comput., Control, Telecommunication Technol.*, 2009, pp. 819–822.
[36] H.-M. Chen, P. K. Varshney, and M. K. Arora, "Performance of mutual information similarity measure for registration of multitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 11, pp. 2445–2454, Nov. 2003.
[37] S. Chen, X. Li, L. Zhao, and H. Yang, "Medium-low resolution multisource remote sensing image registration based on SIFT and robust regional mutual information," *Int. J. Remote Sens.*, vol. 39, no. 10, pp. 3215–3242, 2018.
[38] I. Ito and H. Kiya, "DCT sign-only correlation with application to image matching and the relationship with phase-only correlation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2007, pp. I-1237–I-1240.
[39] X. Wan, J. G. Liu, and H. Yan, "The illumination robustness of phase correlation for image alignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5746–5759, Oct. 2015.
[40] Y. Li, J. Wang, and K. Yao, "Modified phase correlation algorithm for image registration based on pyramid," *Alexandria Eng. J.*, vol. 61, no. 1, pp. 709–718, 2022.
[41] X. Tong et al., "Image registration with fourier-based image correlation: A comprehensive review of developments and applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 10, pp. 4062–4081, Oct. 2019.
[42] M. Gong, S. Zhao, L. Jiao, D. Tian, and S. Wang, "A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 4328–4338, Jul. 2014.
[43] Y. Ye and J. Shan, "A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences," *ISPRS J. Photogrammetry Remote Sens.*, vol. 90, pp. 83–95, 2014.
[44] Y. Ye, B. Zhu, T. Tang, C. Yang, Q. Xu, and G. Zhang, "A robust multimodal remote sensing image registration method and system using steerable filters with first-and second-order gradients," *ISPRS J. Photogrammetry Remote Sens.*, vol. 188, pp. 331–350, 2022.
[45] M. C. Morrone, J. Ross, D. C. Burr, and R. Owens, "Mach bands are phase dependent," *Nature*, vol. 324, no. 6094, pp. 250–253, 1986.
[46] M. C. Morrone and R. A. Owens, "Feature detection from local energy," *Pattern Recognit. Lett.*, vol. 6, no. 5, pp. 303–313, 1987.
[47] S. Venkatesh and R. Owens, "An energy feature detection scheme," in *Proc. IEEE Int. Conf. Image Process.*, 1989, pp. 553–557.
[48] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Trans. Signal Process.*, vol. 49, no. 12, pp. 3136–3144, Dec. 2001.
[49] X.-G. Luo, H.-J. Wang, and S. Wang, "Monogenic signal theory based feature similarity index for image quality assessment," *AEU- Int. J. Electron. Commun.*, vol. 69, no. 1, pp. 75–81, 2015.
[50] L. Yu, D. Zhang, and E.-J. Holden, "A fast and fully automatic registration approach based on point features for multi-source remote-sensing images," *Comput. Geosciences*, vol. 34, no. 7, pp. 838–848, 2008.
[51] 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.

Jing Huang is currently working toward the Ph.D. degree in control science and engineering with the Engineering Research Center of Metallurgical Automation and Measurement Technology, Wuhan University of Science and Technology, Wuhan, China.

Her research interests include multimodal image registration and feature detection.

Fang Yang received the master's degree in information and telecommunication engineering from Wuhan University, Wuhan, China, in 2013, and the Ph.D degree in applied mathematics from Université Paris Dauphine, PSL, Paris, France, in 2017.

She is currently an Associate Professor with Wuhan University of Science and Technology, Wuhan, China. Her research interests include image restoration, feature extraction, geodesic method, and PDE-based image analysis.

Li Chai (Member, IEEE) received the B.S. degree in applied mathematics and the M.S. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 1994 and 1997, respectively, and the Ph.D. degree in electrical engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2002.

From 2002 to 2007, he was with Hangzhou Dianzi University, Hangzhou, China. He worked as a Professor with the Wuhan University of Science and Technology, Wuhan, China from 2008 to 2022. In August 2022, he joined Zhejiang University, Hangzhou, China, where he is currently a Professor with the College of Control Science and Engineering. He has been a Postdoctoral Researcher or Visiting Scholar with Monash University, Melbourne, Australia, Newcastle University, Callaghan, Australia, and Harvard University, Cambridge, MA, USA. He has authored or coauthored over 100 fully refereed papers in prestigious journals and leading conferences. His research interests include distributed optimization, filter banks, graph signal processing, and networked control systems.

Dr. Chai was the recipient of the Distinguished Young Scholar of the National Science Foundation of China. He serves as the Associate Editor of IEEE TRANSACTIONS ON CIRCUIT AND SYSTEMS II: EXPRESS BRIEFS, *Journal of Control and Decision*, and *Journal of Image and Graphs*.