remote sensing

MDPI

Article

3MRS: An Effective Coarse-to-Fine Matching Method for Multimodal Remote Sensing Imagery

Zhongli Fan [1,†], Yuxian Liu [2,†], Yuxuan Liu [1,*], Li Zhang [1], Junjun Zhang [3], Yushan Sun [1] and Haibin Ai [1]

[1] Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping (CASM), Beijing 100036, China; fzl110466@163.com (Z.F.); zhangl@casm.ac.cn (L.Z.); sunys@casm.ac.cn (Y.S.); Aihb@casm.ac.cn (H.A.)
[2] Shenzhen Investigation & Research Institute Co., Ltd., Shenzhen 518026, China; liuyuxian@sziri.com
[3] Baidu Times Technology (Beijing) Co., Ltd., Beijing 100085, China; Zhangjunjun03@baidu.com
* Correspondence: yxliu@casm.ac.cn
† These authors contributed equally to this study and shared the first authorship.

Abstract: The fusion of image data from multiple sensors is crucial for many applications. However, there are significant nonlinear intensity deformations between images from different kinds of sensors, leading to matching failure. To address this need, this paper proposes an effective coarse-to-fine matching method for multimodal remote sensing images (3MRS). In the coarse matching stage, feature points are first detected on a maximum moment map calculated with a phase congruency model. Then, feature description is conducted using an index map constructed by finding the index of the maximum value in all orientations of convolved images obtained using a set of log-Gabor filters. At last, several matches are built through image matching and outlier removal, which can be used to estimate a reliable affine transformation model between the images. In the stage of fine matching, we develop a novel template matching method based on the log-Gabor convolution image sequence and match the template features with a 3D phase correlation matching strategy, given that the initial correspondences are achieved with the estimated transformation. Results show that compared with SIFT, and three state-of-the-art methods designed for multimodal image matching, PSO-SIFT, HAPCG, and RIFT, only 3MRS successfully matched all six types of multimodal remote sensing image pairs: optical–optical, optical–infrared, optical–depth, optical–map, optical–SAR, and day–night, with each including ten different image pairs. On average, the number of correct matches (NCM) of 3MRS was 164.47, 123.91, 4.88, and 4.33 times that of SIFT, PSO-SIFT, HAPCG, and RIFT for the successfully matched image pairs of each method. In terms of accuracy, the root-mean-square error of correct matches for 3MRS, SIFT, PSO-SIFT, HAPCG, and RIFT are 1.47, 1.98, 1.79, 2.83, and 2.45 pixels, respectively, revealing that 3MRS got the highest accuracy. Even though the total running time of 3MRS was the longest, the efficiency for obtaining one correct match is the highest considering the most significant number of matches. The source code of 3MRS and the experimental datasets and detailed results are publicly available.

Keywords: multimodal image matching; nonlinear intensity deformations; coarse-to-fine matching strategy; reliable transformation estimation; phase congruency

Citation: Fan, Z.; Liu, Y.; Liu, Y.; Zhang, L.; Zhang, J.; Sun, Y.; Ai, H. 3MRS: An Effective Coarse-to-Fine Matching Method for Multimodal Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 478. https://doi.org/10.3390/rs14030478

Academic Editors: Francesca Giannone and Valerio Baiocchi

Received: 14 December 2021
Accepted: 18 January 2022
Published: 20 January 2022

1. Introduction

With the fast development of sensor manufacture and space delivery technology, a multiple platform Earth observation system has been formed, providing various remote sensing data of different spatial, spectral resolutions, and different modalities. The various datasets, including the optical, infrared, LiDAR, and SAR data, encode different aspects of information and compensate for each other.

The joint use of the multimodal images from different types of sensor data can benefit many applications, such as change detection [1–4], object detection [5–7], and land use and

land cover [8–12]. A fundamental prerequisite for the joint use of multimodal images is image matching, which finds accurate correspondences on two or more images with overlapped areas. Although image matching has been studied for decades, reliable matching of multimodal images is still a challenging problem considering the involved significant nonlinear intensity deformations (NID). Therefore, it is crucial to accomplish the task of multimodal remote sensing images matching accurately and robustly.

Current multimodal image matching methods can be roughly classified into three categories: area-based, feature-based, and learning-based methods [13]. In terms of learning-based methods, even though deep learning is a promising technology and some works obtained good results on their datasets [14–17], these methods can hardly be applied to real applications now. Firstly, there are no large-scale and universal multimodal image sets for the training process. Secondly, deep neural networks usually require enormous computing resources and have low efficiency. These limitations restrain the application of learning-based methods in the field of multimodal remote sensing image matching.

Area-based methods, also called template matching, complete the matching process by checking the similarity between two selected template areas. Generally, the traditional area-based methods use the intensity information of images, such as NCC [18] and MI [19]. However, NCC can hardly handle NID, seriously declining its performance. MI shows robustness to NID because of the utilization of statistical information of image intensity distribution, but it is easy to fall into local optima due to ignoring the influence of neighboring pixels. A few studies recently found that the geometric structures and shape features stayed stable across different modal images and proposed using the information to conduct multimodal image matching. Based on HOG [20], Ye et al. [21] proposed the HOPC descriptor by using phase congruency [22] features instead of gradient features, achieving good performance on multimodal image matching. However, both HOG and HOPC descriptors are characterized in a sparse sampling grid, so they are difficult to capture the detailed structural information in the image. With this in mind, Ye et al. [23] further proposed the channel feature of orientated gradients (CFOG). CFOG constructs descriptors in a pixel-by-pixel manner, enhancing the ability to describe detailed structures in images, and its matching performance is significantly better than sparse feature descriptors. Based on CFOG, Fan et al. [24] developed an angle weighted orientation gradients (AWOG) descriptor by distributing the gradient value into two most related orientations and proposed to use three-dimensional phase correlation as a similarity metric, significantly improving the matching performance. However, regardless of the good performance area-based methods achieved, they have a high requirement for the initial matching positions. The matching performance will decrease dramatically if the initial input has a large deviation with the correct matching point. Additionally, the area-based methods are sensitive to scale change and image rotation, so the geometric deformations need to be eliminated roughly in advance, limiting its versatility in various applications.

Feature-based methods detect salient features on the images and match them based on the similarity of the feature descriptors. One of the representative feature-based methods is scale-invariant feature transform (SIFT) [25], widely applied to matching optical remote sensing images considering its good robustness to scale and illumination change and rotation. However, SIFT tends to fail when there is a large NID between the image pair. Towards the matching of multi-source remote sensing images, Ma et al. [26] proposed the PSO-SIFT algorithm, which optimizes the way of calculation of image gradients to increase the robustness to intensity difference and introduce an enhanced matching strategy using multiple aspects of information of the feature points to increase the number of image correspondences. To accomplish the task of multimodal remote sensing image matching, Yao et al. [27] put forward the histogram of absolute phase consistency gradients (HAPCG) algorithm. They first used an anisotropic filter to preprocess images, constructing an anisotropic weighted moment equation based on image phase consistency. Then, they extended the phase consistency model, built an absolute phase consistency orientation gradient, and established the HAPCG descriptor. Relatively robust matching of multimodal

images is achieved with HAPCG. To increase the robustness to large NID, Li et al. [28] proposed a radiation-variation insensitive feature transform (RIFT) algorithm that employs phase congruency and introduces a maximum index map (MIM) descriptor. They first detected the salient corner and edge feature points on the phase congruency map and then constructed the MIM descriptor based on a log-Gabor convolution image sequence, obtaining superior performance than SIFT. In a word, feature-based methods demonstrate more flexibility in multimodal remote sensing image matching considering the good resistance against scale change and image rotation and low requirements for initial image pose conditions. However, the repetition rate of directly extracted feature points is relatively low because of the large NID between multimodal images. For example, the inlier ratio for RIFT is only about 15~20%. Moreover, the accuracy of feature-based methods is always lower than area-based methods, considering the unstable localization accuracy of feature points.

In this paper, we propose a coarse-to-fine multimodal remote sensing image matching method (3MRS) based on the 2D phase congruency model to overcome the large NID. In the stage of coarse matching, feature points are detected on a maximum moment map computed from multi-scale and multi-oriented phase congruency (PC) images, considering the map can reflect the apparent corner and edge features. Then, feature description is conducted by finding an index map for fast and effective image matching. Finally, the fast sample consensus (FSC) algorithm [29] is employed to remove the outliers. Besides, the coarse affine transformation model between images can be estimated based on the obtained correspondences. In the stage of fine matching, we proposed a novel template feature based on the log-Gabor convolution image sequence to rematch the extracted feature points and use a 3D phase correlation as the similarity measure, significantly improving the matching rate of extracted feature points and obtaining a more significant number of correspondences with high accuracy. Since the template feature is built on the log-Gabor convolution image sequence obtained during the coarse matching process, there is no extra time cost, improving the computation efficiency. To testify the performance of 3MRS, we compare the results of 3MRS with four state-of-the-art multimodal remote sensing image matching methods: SIFT, PSO-SIFT, HAPCG, and RIFT, on six types of multimodal image datasets: optical–optical, optical–infrared, optical–depth, optical–map, optical–SAR, and day–night and each type of dataset contains ten image pairs. Results reveal that 3MRS successfully matches all image pairs. Besides, 3MRS obtains 1365 correct matches on average and an accuracy of 1.47 pixels, while those of the corresponding best results of the comparative methods are 313 correct matches and 2.45 pixels, respectively, demonstrating significant performance improvement.

This study is structured as follows: Section 2 gives the principle of our proposed method, 3MRS. Section 3 presents the experiments and results concerning 3MRS. Section 4 discusses several important aspects related to 3MRS. Finally, this study is concluded in Section 5.

2. Methodology

Even though the coarse-to-fine framework has been studied and applied to matching same or different source optical satellite images [30–32], it is barely applied to tackle the matching of multimodal remote sensing imagery. Therefore, we propose the 3MRS algorithm to explore the potential of fulfilling the task with a coarse-to-fine pipeline. Unlike the exact source of optical satellite images, there is a large NID between the multimodal images. The core is to ensure that each stage of the method is robust to NID.

Following this thought, we conduct coarse and fine matching based on 2D phase congruency considering its high invariance to NID. Specifically, we detect feature points on a maximum moment map calculated from the multi-scale and multi-oriented PC images and construct the feature descriptor by applying a distributed histogram method on an index map calculated from convolved images of all orientations. After that, feature matching and mismatch elimination are performed to obtain a few reliable matches, which can be

used to estimate the rough affine transformation model between images. In the stage of fine matching, taking the positions predicted through the estimated transformation model, we construct template features from the multi-oriented log-Gabor convolution sequences calculated using both even-symmetric and odd-symmetric log-Gabor filters and further convolve the template features with a 3D Gaussian-like kernel. Then, 3D phase correlation matching is applied to build correspondences, and outlier removal is employed to refine the matches. In detail, the pipeline of 3MRS is given in Figure 1.

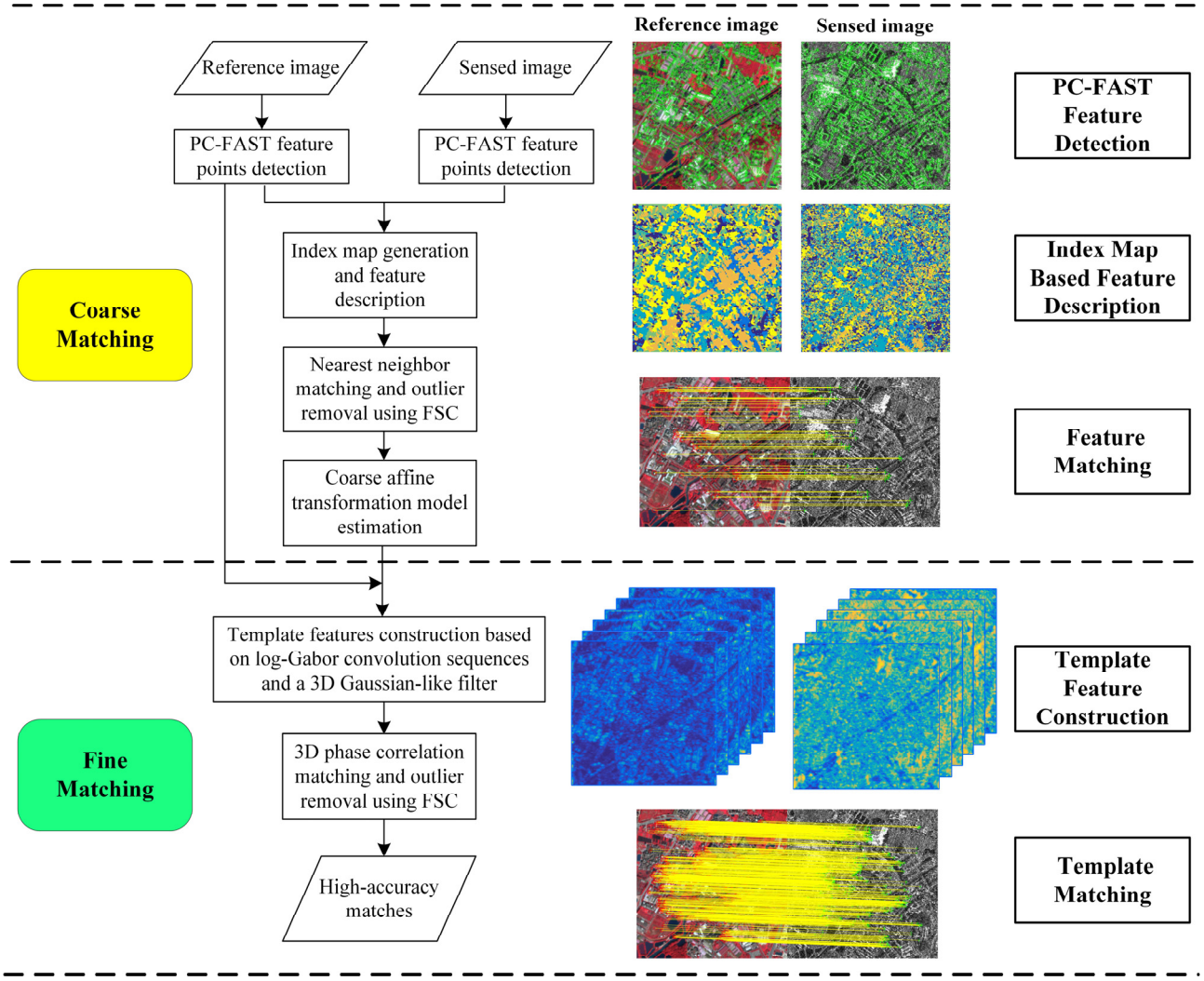


Figure 1. The pipeline of 3MRS.

2.1. Coarse Matching

The goal of coarse matching is to provide an affine transformation model between images for the following optimization process, which can be estimated by a few reliable matches. In detail, the coarse matching process mainly contains the two steps of feature detection and feature description, in which the 2D phase congruency model is adopted considering its good robustness against different image conditions.

2.1.1. Point Feature Detection

Phase congruency (PC) [22] is applied for feature points detection. Given a 2D image $I(x, y)$, the 2D PC model can be expressed as follows:

$$PC(x,y) = \frac{\sum_s \sum_o w_o(x) \lfloor A_{so}(x,y)\Delta\Phi_{so}(x,y) - T \rfloor}{\sum_s \sum_o A_{so}(x,y) + \varepsilon} \quad (1)$$

where $w_o(x)$ is a weighted function; $A_{so}(x,y)$ and $\Phi_{so}(x,y)$ are the amplitude component and phase component of $I(x,y)$ at scale s and orientation o, respectively; T compensates the noise; ε is a small value; and $\lfloor \cdot \rfloor$ is a truncation function, the value is itself when it is larger than 0; otherwise, the value is 0. Note that $A_{so}(x,y)$ and $\Phi_{so}(x,y)$ can be calculated with the even-symmetric and odd-symmetric responses obtained through convolving the image $I(x,y)$ with an even-symmetric and an odd-symmetric 2D log-Gabor filter [33], respectively.

Compared with image gradients, the PC map has much better noise resistance. Figure 2 demonstrates the computed gradient maps and PC maps of an optical aerial image with/without noises. We can see that, with the presence of large amounts of noises, the image gradient is so severely affected by noises that the gradient map can hardly reflect any useful information, while the PC map is much less affected by noises and keeps the image structures well.

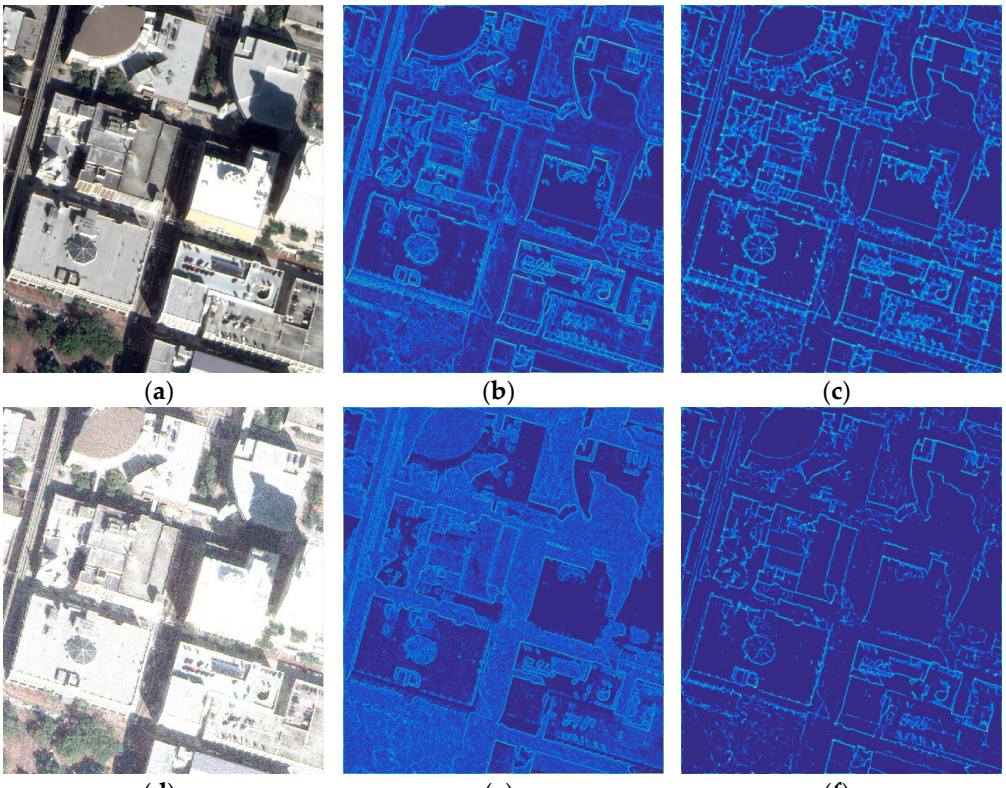


Figure 2. Performance of gradients and PC with the presence of noises: (**a**,**d**) are the images with/without noises; (**b**,**e**) are the computed gradient maps on (**a**,**d**); and (**c**,**f**) are the computed PC maps on (**a**,**d**), respectively.

Then, the feature points are detected based on a PC measure. Following the moment analysis algorithm [34], the principal axis Φ corresponds to the minimum moment m_Φ and indicated feature direction. While the axis corresponds to the maximum moment M_Φ is perpendicular to the principal axis. The magnitude of the M_Φ and m_Φ can be calculated as follows:

$$\Phi = \frac{1}{2}\arctan\left(\frac{B}{A-C}\right) \quad (2)$$

$$M_\Phi = \frac{1}{2}\left(C + A + \sqrt{B^2 + (A - C)^2}\right) \tag{3}$$

$$m_\Phi = \frac{1}{2}\left(C + A - \sqrt{B^2 + (A - C)^2}\right) \tag{4}$$

where

$$A = \sum_o (PC(\theta_o)\cos(\theta_o))^2 \tag{5}$$

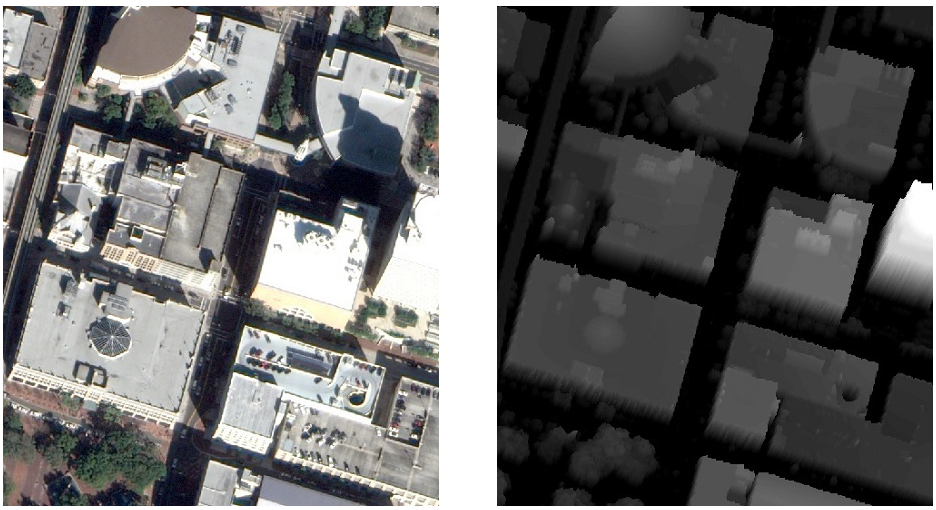
$$B = 2\sum_o (PC(\theta_o)\cos(\theta_o))(PC(\theta_o)\sin(\theta_o)) \tag{6}$$

$$C = \sum_o (PC(\theta_o)\sin(\theta_o))^2 \tag{7}$$

where $PC(\theta_o)$ is the PC map under the orientation of θ_o.

Theoretically, the minimum moment map m_ψ represents the corner map and the maximum moment map M_ψ represents the edge map of the image. However, the PC corner map is a strict subset of the PC edge map [22]. Thus, this provides a simplified way to integrate edge and corner information of the image. Namely, we do not detect corner features on m_ψ and only detect features on M_ψ which includes both corner and edge features. Moreover, the FAST algorithm [35] is applied on M_ψ to obtain a large number of distinctive features.

Figure 3 gives the feature extraction results on a pair of optical-depth images. Figure 3b,c show the feature points extracted by applying the FAST algorithm on the original image and the PC map, respectively. We can see that the feature points obtained from using the original images are very sparse and unstable under the presence of NID. In contrast, lots of reliable corner and edge feature points with good distribution are detected based on the maximum moment maps, proving the excellent invariance of the PC measure to NID.



(**a**)

Figure 3. *Cont.*

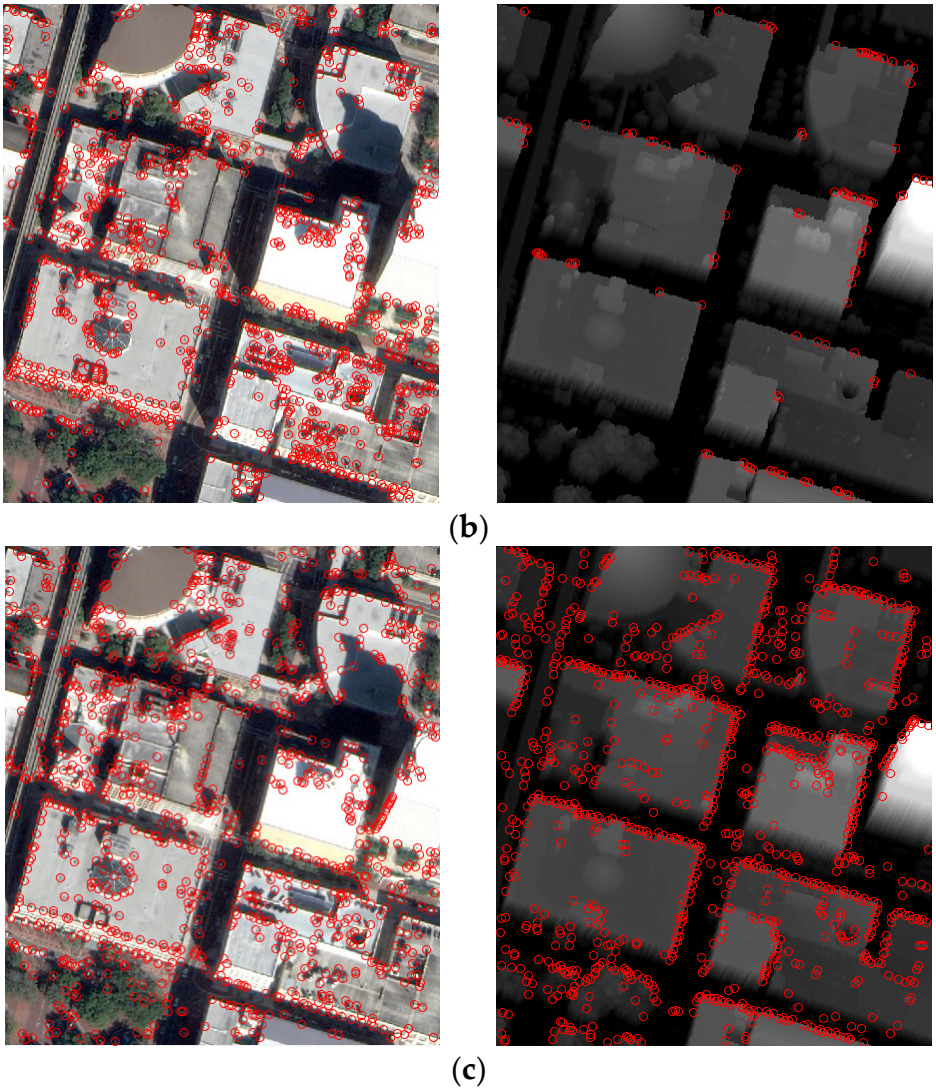


Figure 3. Comparison of feature detection using FAST and PC-FAST detectors. (**a**) The original optical (**left**) and depth (**right**) image pair; (**b**) FAST feature points based on original images; and (**c**) PC-FAST feature points based on maximum moment maps.

2.1.2. Index-Map-Based Feature Description

After obtaining the feature points, a feature descriptor is needed for each point to increase the feature discrimination and robustness against changes such as NID and rotation. Here, an index map is applied based on the log-Gabor convolution sequence. Moreover, the map is used for feature description. Given that the sequence has been obtained during the generation of the PC maps, the construction of the index map only needs very few calculations and time. Precisely, for each orientation o, the amplitude components at all scales are summed up to obtain a log-Gabor layer $A_o(x, y)$.

$$A_o(x, y) = \sum_{s=1}^{N} A_{so}(x, y) \tag{8}$$

A log-Gabor convolution sequence can be built by arranging the log-Gabor layers at all orientations. Then, the index map is constructed by searching the index of the maximum value in all orientations:

$$indexMap(x, y) = OI(\max(A_o(x, y))) \tag{9}$$

where $\max(\cdot)$ is applied to locate the maximum value in a sequence of convolved images, and $OI(\cdot)$ is used to get the index of the maximum value in the image sequence. Figure 4 demonstrates the generation process of an index map of a typical optical aerial image. Then, we construct the feature descriptor using a distributed histogram method similar to SIFT on the obtained index map.

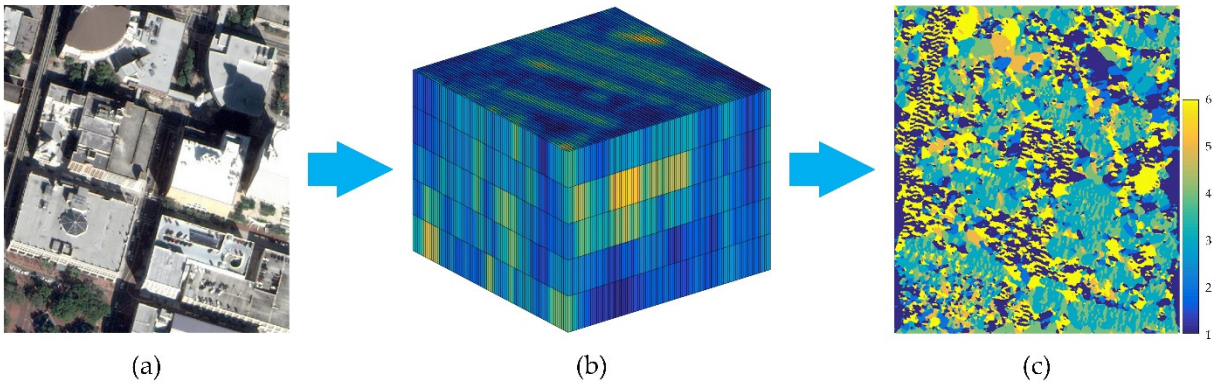


Figure 4. Index map generation: (**a**) original image; (**b**) log-Gabor convolution image sequence; and (**c**) index map.

After obtaining the feature descriptors, the nearest neighbor matching strategy is applied to obtain the initial correspondences, and the FSC algorithm is employed to eliminate outliers. Based on the obtained matches, the affine transformation model between images is calculated, which can be applied to provide initial correspondence locations for the following fine matching.

2.2. Fine Matching

Even though a few correct matches can be achieved in the coarse matching, the matching rate is low, and most of the extracted distinctive feature points are wasted. Besides, the accuracy of the matches is limited by that of the detected feature points. Therefore, we proposed a novel template-based method to refine the initial matching result. As displayed in Figure 5, the initial correspondences for all extracted feature points are firstly calculated using the affine transformation parameters obtained in coarse matching. Then, we take the initial correspondences as input and compute a template feature based on a window image centered at the initial predicted location. Finally, we match the template features with a 3D phase correlation measure and remove the outliers with the FSC algorithm. In this way, the number and accuracy of the obtained matches are significantly improved.

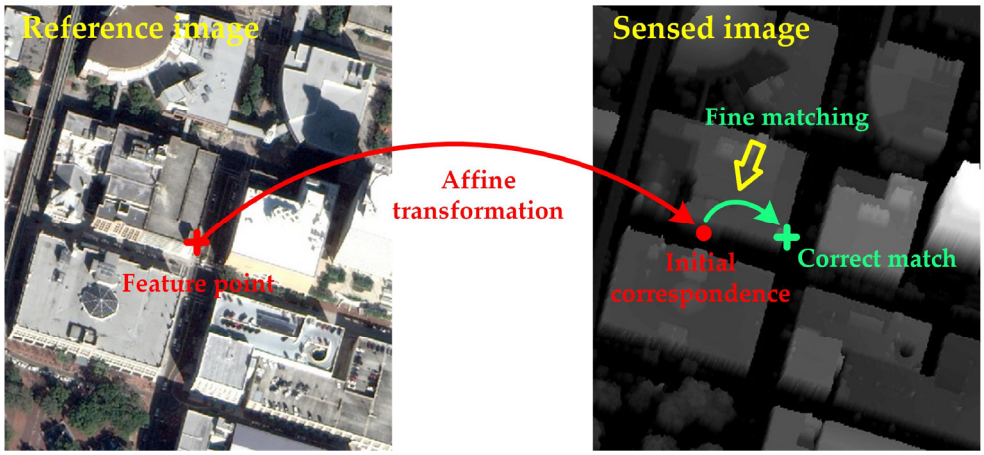


Figure 5. Fine matching.

2.2.1. Template Feature Construction

Similar to the other template-based methods, we first open a same-size window on each of the candidate images. Then, we build a feature vector based on the log-Gabor convolution image sequence, which is calculated using both even-symmetric and odd-symmetric log-Gabor filters for each pixel in the window. Unlike the index map, which only utilizes the index information, we employ the whole sequence that contains more detailed structural information, increasing the feature distinguishability. Specially, we simply arrange the log-Gabor convolution sequence in the order of orientation in the z-direction, and a cube-like 3D image displayed in Figure 6b can be obtained. Since the log-Gabor convolution sequence has been obtained during the generation process of PC maps, no more time is required for this step.

After that, a 3D Gaussian-like kernel is applied to the 3D image cube to reduce the influence of local distortions caused by geometric and intensity deformation. Precisely, the 3D Gaussian-like kernel consisted of a 2D Gaussian kernel in the xy plane whose diameter and standard deviations are 3 pixels and 0.5, respectively, and a kernel of $d_z = [1, 3, 1]^T$ in the z-direction. This process can be described with the following equations:

$$A_o^\sigma(x, y) = g_{xy}^\sigma * A_o(x, y) \tag{10}$$

$$T_o(x, y) = d_z * A_o^\sigma(x, y) \tag{11}$$

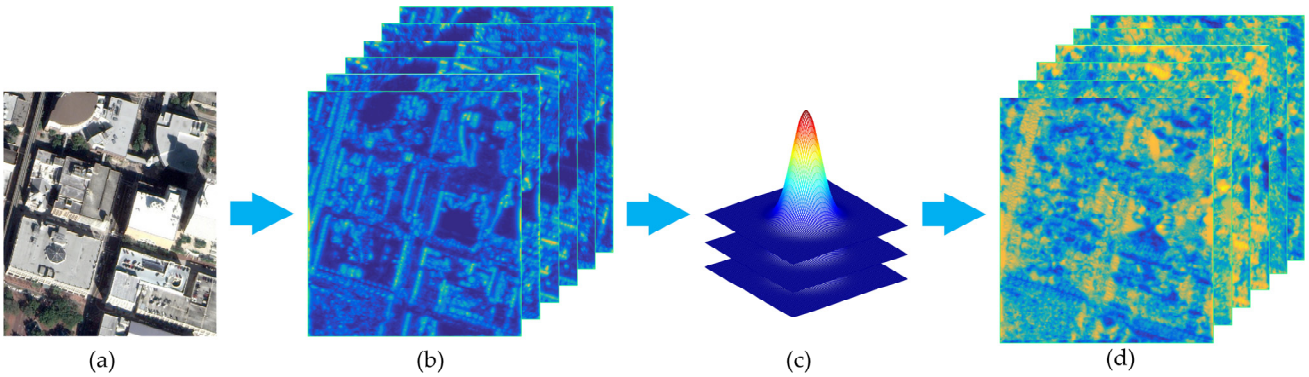
where $A_o(x, y)$ is the 3D image feature obtained from arranging the log-Gabor layer images in a specific order, g_{xy}^σ is a Gaussian kernel in the xy plane, d_z is a kernel in the z-direction, and $T_o(x, y)$ is the template feature after filtering.

At last, we conduct normalization on the z-direction to further increase the robustness of the feature vector. Specifically, the L2 norm is used to normalize the feature vector, which can be expressed as follows:

$$T_i(x, y) = \frac{T_i(x, y)}{\sqrt{\sum_{i=1}^{6} |T_i(x, y)|^2 + \varepsilon}} \tag{12}$$

where ε is a small constant value.

The feature vectors of the pixels in the window form the template feature, which is still a 3D image cube. Figure 6 displays the generation process of a template feature.



(a) (b) (c) (d)

Figure 6. Template feature generation: (**a**) an optical aerial image; (**b**) preliminary template features; (**c**) a 3D Gaussian-like kernel; and (**d**) template features.

2.2.2. 3D Phase Correlation Matching

Considering the template feature is three-dimensional, much computation is required when using the traditional space-based similarity metric, such as SSD and NCC. Therefore, we utilized the 3D phase correlation instead, which has been proven to have high efficiency and keep high matching accuracy [24,36,37].

Given that $t_r(x, y, z)$ and $t_s(x, y, z)$ are the template features of the reference image and sensed image, respectively, their geometric relationship is as follows:

$$t_r(x, y, z) = t_s(x - x_0, y - y_0, z) \tag{13}$$

where (x_0, y_0) is the offset between two image windows, and z is the dimensionality of the template feature.

$T_r(u, v, w)$ and $T_s(u, v, w)$ can be obtained by performing a 3D fast Fourier transform of $t_r(x, y, z)$ and $t_s(x, y, z)$. Moreover, their correlation can be expressed as follows according to the Fourier shift theorem:

$$T_r(u, v, w) = T_s(u, v, w)e^{-i(ux_0+vy_0)\vec{\gamma}} \tag{14}$$

where $\vec{\gamma}$ is a 3D unit vector. Specifically, their cross-power spectrum can be expressed as follows:

$$T_r(u, v, w)T_s(u, v, w)^* = e^{-i(ux_0+vy_0)\vec{\gamma}} \tag{15}$$

where $*$ denotes complex conjugate.

Then, a correlation function $\delta(x - x_0, y - y_0)$ can be obtained by conducting inverse Fourier transform to the cross-power spectrum. Generally, the peak position of the correlation function will appear in (x_0, y_0). Therefore, the matching point can be found by searching the local maximum of the following equation:

$$F^{-1}\left(T_r(u, v, w)T_s(u, v, w)^*\right) = \delta(x - x_0, y - y_0)\vec{\gamma} \tag{16}$$

where $F^{-1}(\cdot)$ represents the inverse transform of the 3D fast Fourier transform.

The size of the template window can significantly affect the matching performance. Considering the complex change between multimodal images, we use a relatively larger size than that in same-modal image matching, which is 101 $\times$ 101 pixels in this study, to ensure the matching accuracy and robustness. At last, the FSC algorithm is used for outlier removal.

3. Experiments and Results

In order to verify the superior of 3MRS, we compare it with four state-of-the-art algorithms: SIFT, PSO-SIFT, HAPCG, and RIFT. For a fair comparison, the codes of the comparative methods provided by the authors are applied, and the parameters are set according to the recommendations of the providers.

3.1. Data Description

The CoFSM datasets provided by Yao et al. [27] are used, including six types of multimodal image pairs: optical–optical, optical–infrared, optical–depth, optical–map, optical–SAR, and day–night. There are significant NID and slight geometric differences due to different time phases, lighting conditions, and sensor differences between the image pairs. Each type of image pair contains ten image pairs, and each image pair has about 10 to 30 high-precision correspondences manually selected by the provider. Note that the correspondences can be used to estimate the true transformation model H of the image pair, which are taken as ground truth for the following evaluation process.

3.2. Evaluation Indices

To comprehensively testify the proposed method, we present large amounts of qualitative and quantitative experimental results. For quantitative evaluation, four indices are utilized, which are success ratio (SR), the number of correct matches (NCM), root mean square error (RMSE), and the running time.

1. SR refers to the ratio of the number of successfully matched image pairs to the total number of image pairs in a type of image pair. This index reflects the robustness of a matching method to a specific type of multimodal image pair.
2. To count the number of correct matches, we first use the obtained matches to estimate a transformation between an image pair. Then, the matches with residual errors of less than three pixels are taken as correct matches, and the number of correct matches is NCM. Additionally, the image pair with NCM smaller than three is deemed a matching failure. Considering the significant NID between multimodal remote sensing images, three pixels are a relatively strict threshold.
3. Taking the correct matches as input, the coordinates (x_1, y_1) on one image can be converted to (x_1', y_1') on the other image of the image pair using H. If the coordinates of the corresponding matching point of (x_1, y_1) are (x_2, y_2), RMSE can be calculated with (17). RMSE reflects the matching accuracy of the correct matches. The smaller the value of RMSE, the higher the accuracy. In addition, the image pairs with RMSE larger than five are deemed a matching failure.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{NCM}} \sum_{i=1}^{\mathrm{NCM}} \left[\left(x_1^{i'} - x_2^{i} \right)^2 + \left(y_1^{i'} - y_2^{i} \right)^2 \right]} \tag{17}$$

$$\left[x_1^{i'}, y_1^{i'}, 1 \right]^T = H \cdot \left[x_1^{i}, y_1^{i}, 1 \right]^T \tag{18}$$

4. With respect to efficiency, we not only count the total running time $\mathrm{T_{total}}$ but also the time $\mathrm{T_{one}}$ used for obtaining one correct match. Specifically, $\mathrm{T_{one}}$ can be calculated as follows:

$$\mathrm{T_{one}} = \frac{\mathrm{T_{total}}}{\mathrm{NCM}} \tag{19}$$

3.3. Qualitative Results

We selected one representative image pair from each of the six types of multimodal image datasets and visualized their experimental results. These image pairs contain various significant NID due to the time phase or imaging mechanism. Therefore, it is very challenging to match these image pairs automatically. Figure 7 shows the comparative visualization matching results of SIFT, PSO-SIFT, HAPCG, RIFT, and 3MRS.

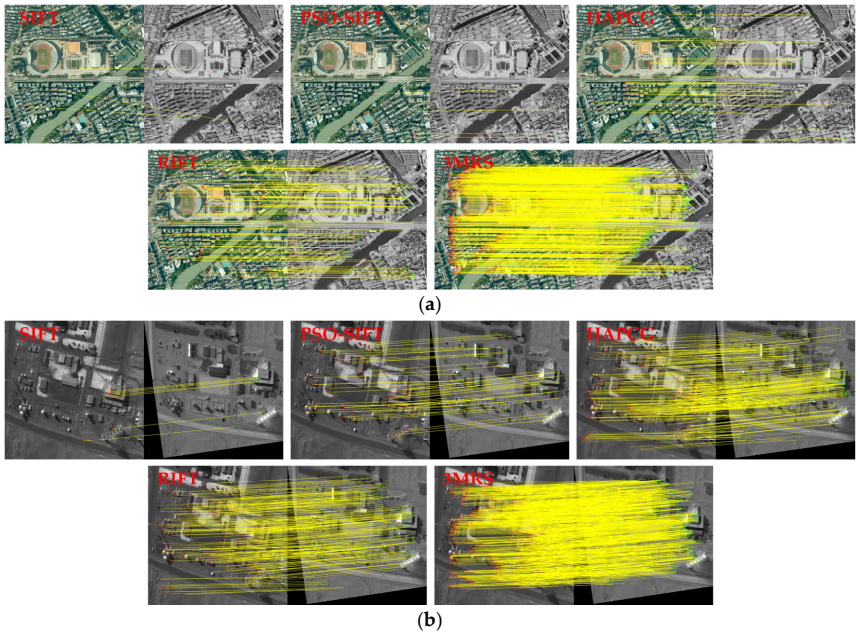


Figure 7. *Cont.*

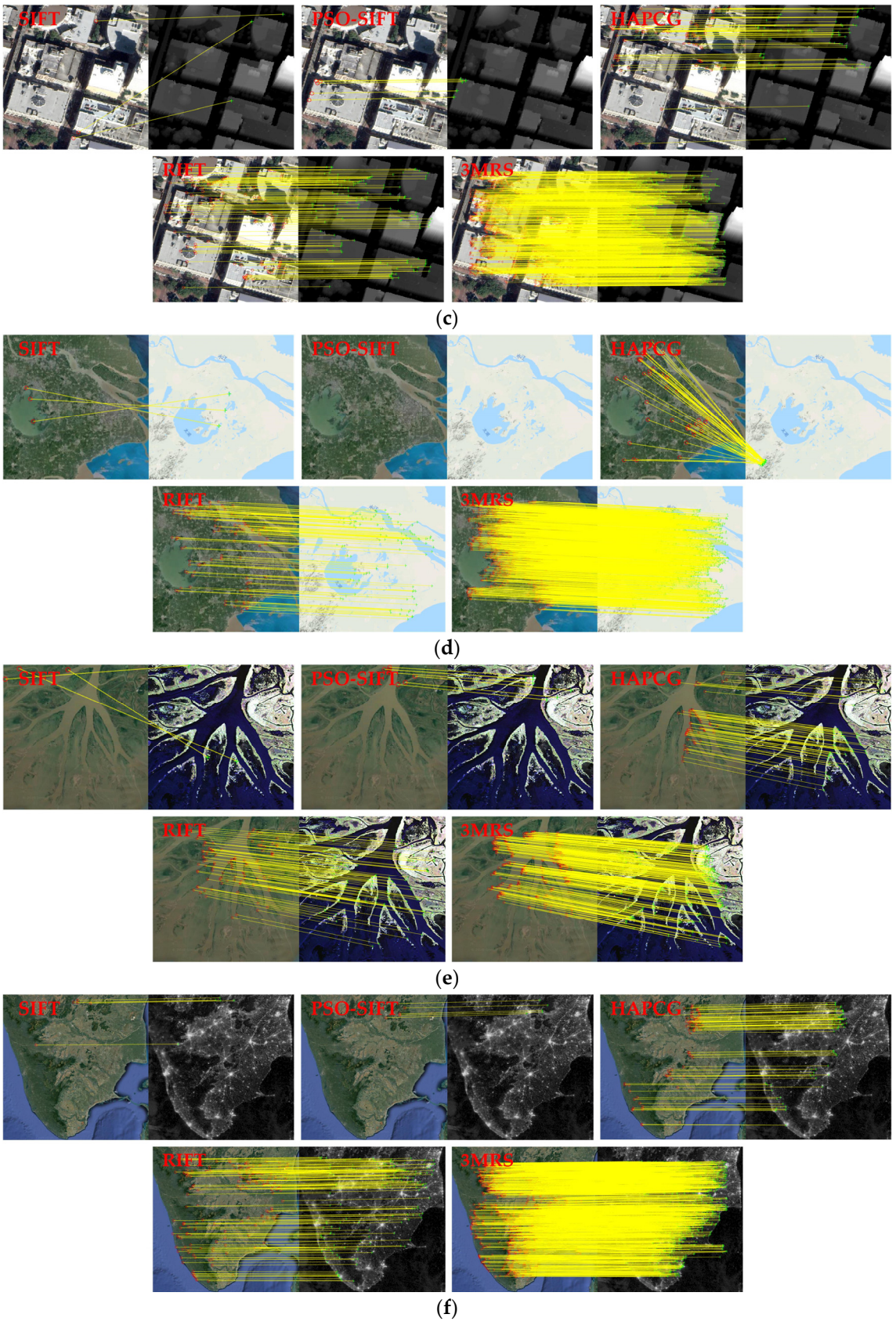


Figure 7. Comparison of visualization results of SIFT, PSO-SIFT, HAPCG, RIFT, and 3MRS on six types of multimodal images: (**a**–**f**) correspond to optical–optical, optical–infrared, optical–depth, optical–map, optical–SAR, and day–night image pairs.

From the results of Figure 7, we can see that SIFT almost failed in all image pairs except for a few matches of image pairs of optical–infrared images. The reason can be that there are no large NID between the optical and infrared images, while the NID is significant between the other types of multimodal image pairs. HAPCG matched five of the six types of multimodal image pairs and failed to match the optical and map images. PSO-SIFT successfully matched the image pairs of optical–optical, optical–infrared, optical–depth, and day–night image pairs; a few matches for the image pair of optical–SAR ultimately failed to match the optical–map pair. However, we can see that the matches of the optical–SAR are incorrect. These results indicate that HAPCG and PSO-SIFT have resistance to NID to some extent but cannot handle the matching of all types of multimodal remote sensing image pairs. On the contrary, RIFT effectively matched all image pairs and obtained considerable matches, demonstrating that RIFT has good robustness against various types of NID. However, the successfully matched points still took a relatively small percentage of all extracted feature points, and the other feature points were wasted. Based on the correct matches obtained from coarse matching, 3MRS calculates the transformation model between the image pair and further matches the unmatched feature points with a template matching strategy, significantly improving the matching rate of extracted feature points and thus obtaining more corrected matches than all the competed methods. Besides, the matches of 3MRS have the best distribution among all methods.

Apart from the visualization results of image matching, we also displayed the registration and fusion results of the image pairs with the obtained matches of our proposed 3MRS algorithm in Figure 8. Accurate registration and fusion can be achieved when the matches have high precision and even distribution. From the results, all the image pairs are well registered, proving that the obtained matches of 3MRS have excellent quality in accuracy and distribution.

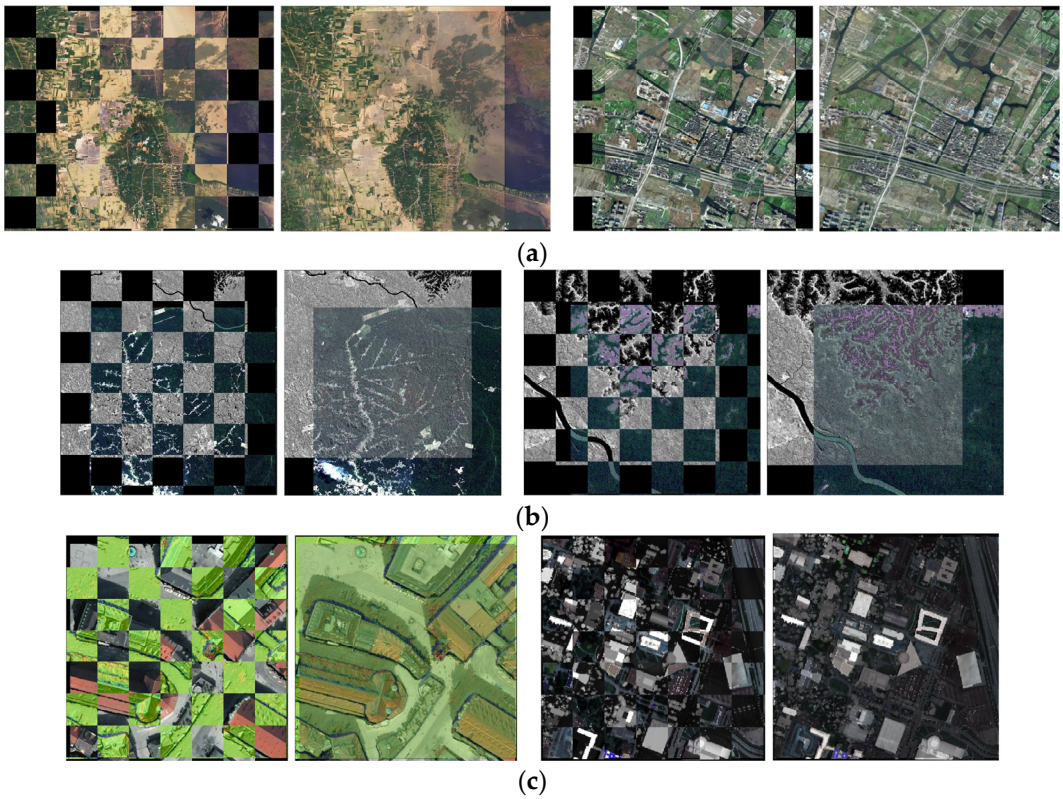


Figure 8. *Cont.*

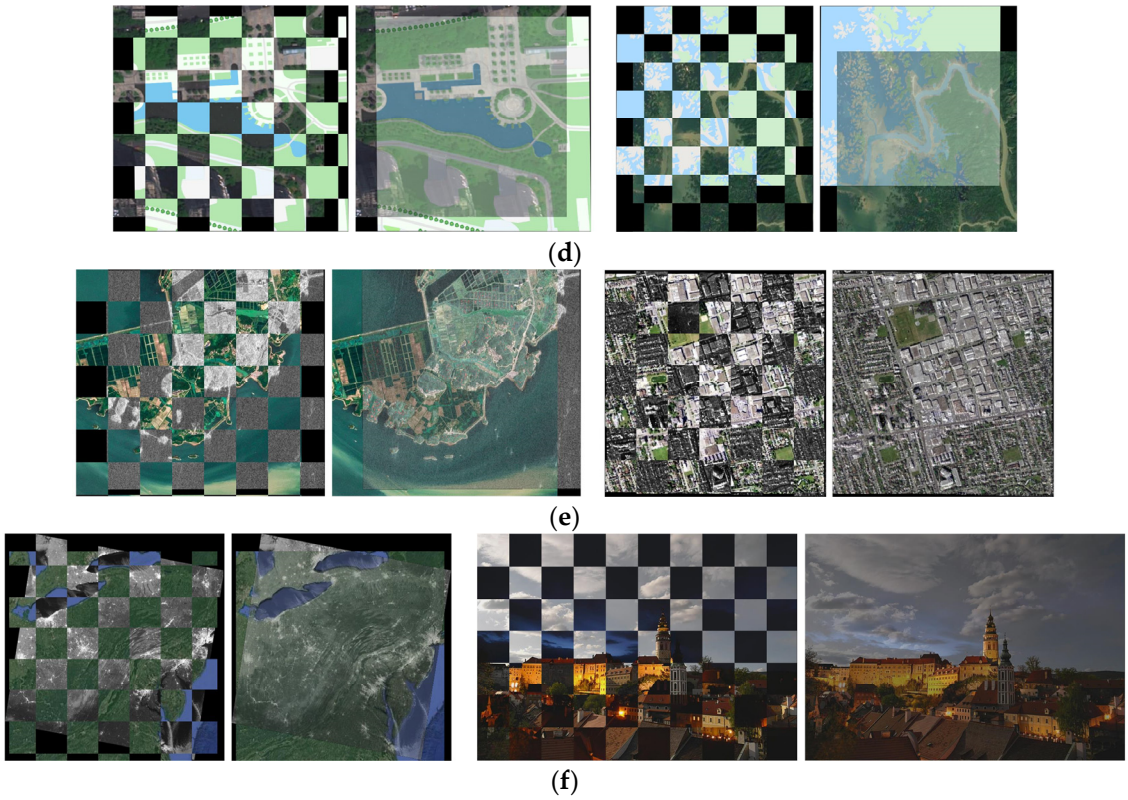


Figure 8. The registration (**left**) and fusion (**right**) results with the matches obtained by 3MRS, where the first two columns are a pair, and the last two columns are another pair: (**a**–**f**) correspond to optical–optical, optical–infrared, optical–depth, optical–map, optical–SAR, and day–night image pairs.

3.4. Quantitative Results

The quantitative results on all 60 image pairs in terms of the four indices are given in this section. Table 1 presents the results of SR for all methods. As we can see, SIFT had the worst SR on all types of image pairs and ultimately failed on the optical–depth and optical–SAR image pairs. PSO-SIFT obtained a better SR on the optical–infrared image pair with an SR of 90%. However, it had similar performance with SIFT for the other types of image pairs. HAPCG successfully matched the optical–infrared data set with an SR of 100%, but it performed slightly worse on the optical–map, optical–SAR, and day–night data sets, all at 70%. RIFT performed much better on all types of image pairs, with an SR of 90% for the optical–SAR image pairs and 100% for all the other types of image pairs, while 3MRS successfully matched all image pairs, demonstrating excellent robustness in multimodal remote sensing image matching.

Table 1. Comparisons on SR metric.

Method	SR/%					
	Optical–Optical	Optical–Infrared	Optical–Depth	Optical–Map	Optical–SAR	Day–Night
SIFT	80	30	0	40	0	50
PSO-SIFT	60	90	10	40	0	40
HAPCG	90	100	90	70	70	70
RIFT	100	100	100	100	90	100
3MRS	100	100	100	100	100	100

Figure 9 presents the results of NCM. We can see that SIFT and PSO-SIFT performed the worst, obtaining very few matches on all image pair categories. HAPCG was similar to RIFT, where RIFT performed more stable and had larger NCM than HAPCG in most cases. Notably, HAPCG failed to match the 4th pair of optical–optical, the 10th pair of optical–depth image pairs, the 1st, 4th, and 8th pairs of optical–map image sets, the 1st, 2nd, and 7th pairs of optical–SAR image sets, and the 5th, 6th, and 9th pairs of the day–night image sets, while RIFT successfully matched all these image pairs. Moreover, the NCM of 3MRS had noticeable improvement compared with all the comparative methods. Particularly, the NCM of 3MRS was 164.47, 123.91, 4.88, and 4.33 times that of SIFT, PSO-SIFT, HAPCG, and RIFT for the successfully matched image pairs of each method.

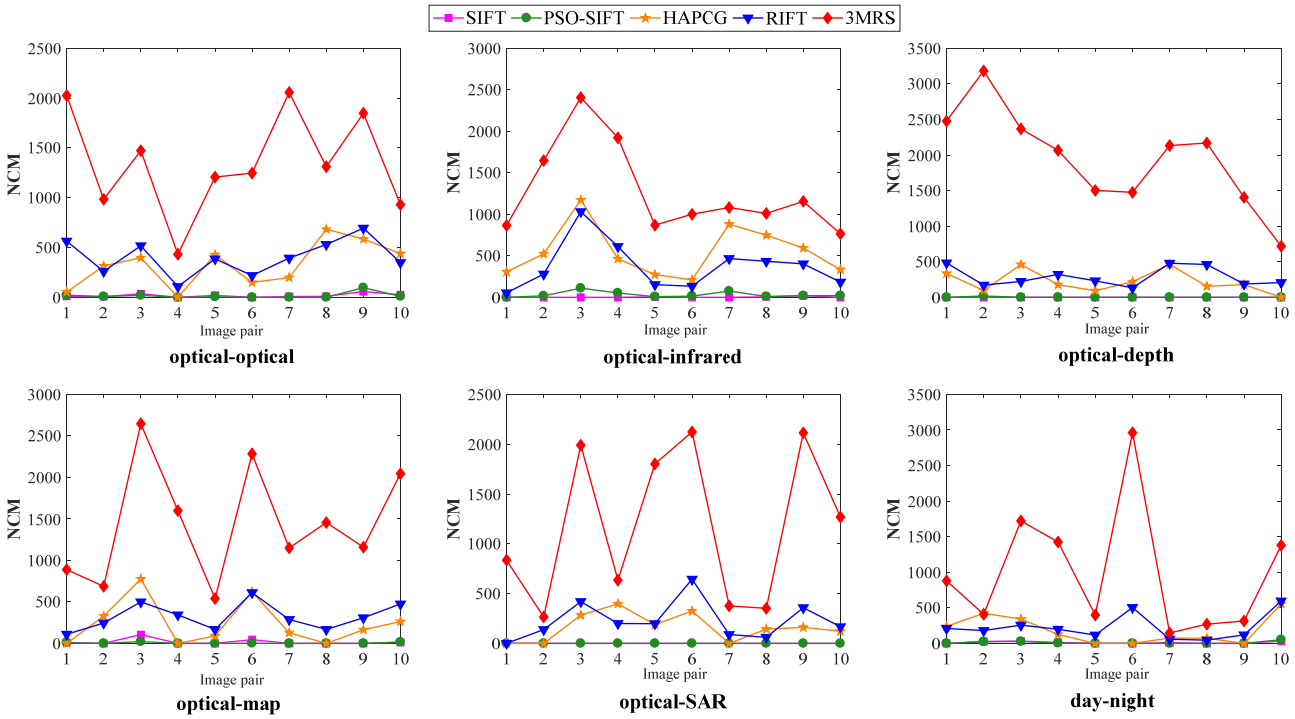


Figure 9. Comparisons on the NCM metrics.

Considering the accuracy, the root-mean-square error of NCM for 3MRS, RIFT, HAPCG, PSO-SIFT, and SIFT are 1.47, 2.45, 2.83, 1.79, and 1.98 pixels, respectively. Specifically, Figure 10 gives the detailed comparative results of RMSE. For better visualization, we added a failed line in the subfigures; the value of RMSE would be on the line when the situation of matching failure appears. The results show that SIFT still performed the worst, failing to match in many image pairs and having the largest RMSE on the successfully matched image pairs. PSO-SIFT had high matching accuracy on some image pairs, such as the 2nd and 10th pairs of the optical–optical image sets, even higher than 3MRS. However, it failed a lot, and its performance declined dramatically on the other types of image pairs, especially the optical–depth and optical–SAR image sets. The performance of HAPCG was quite unstable, while that of RIFT was much more stable and better than HAPCG. The RMSE of 3MRS is always the smallest, demonstrating high accuracy with a value between 0.5 and 2 pixels. Even though most of the RMSEs of the image pairs are smaller than 2, there are several unexpected cases. Significantly, the RMSE of the 10th image pair of the optical–infrared is 3.37 pixels, those of the 1st and 3rd image pairs of the optical–SAR image pairs are 4.35 pixels and 2.67 pixels, and those of the 1st and 7th pairs of the day–night are 2.91 pixels and 3.32 pixels, respectively. After checking with these image pairs, we found that these images cover many areas with less texture or structural information, including the water or woodland areas, significantly increasing the matching difficulty.

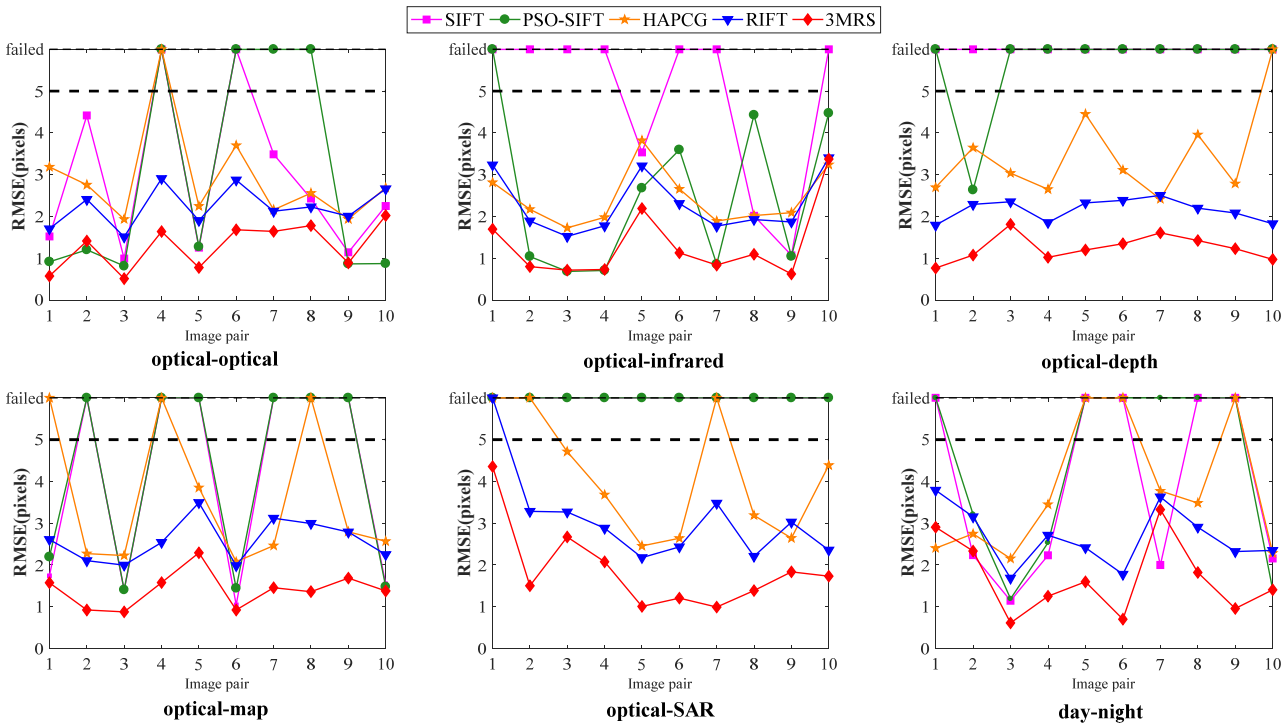


Figure 10. Comparisons of the RMSE metrics. Note that the matching failure cases are drawn onto the failed line.

Table 2 gives the comparative results of the running time t. As we can see, 3MRS costs the longest time, which is about three times of RIFT. However, if we distribute the whole time T_{total} onto the time T_{one} used for obtaining one match, 3MRS has the highest efficiency, which only needs 12.54ms for one match. RIFT is on the same level as 3MRS, slightly faster than HAPCG. PSO-SIFT had the lowest efficiency and took 163.46ms for one match, about 13 times of 3MRS.

Table 2. Comparisons of the time metric.

Method	SIFT	PSO-SIFT	HAPCG	RIFT	3MRS
T_{total} (s)	48.44	108.05	509.83	355.56	1027.32
T_{one} (ms)	97.27	163.46	30.39	19.25	12.54

4. Discussion

4.1. Performance Analysis

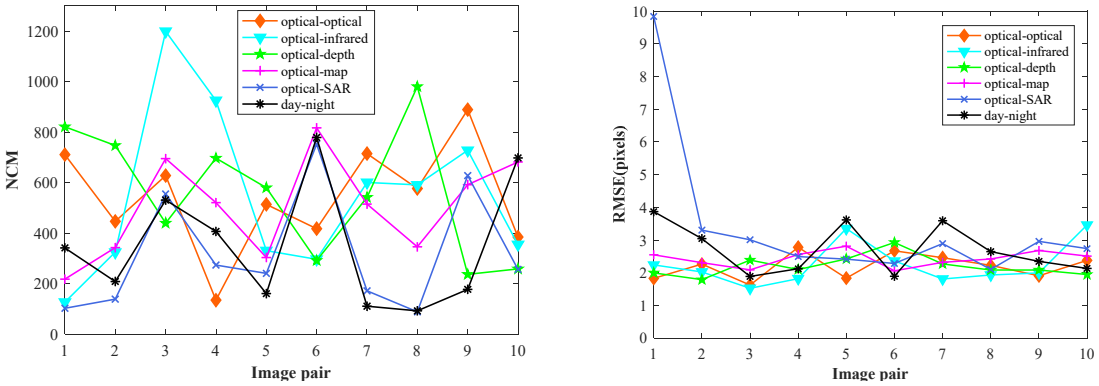
From the qualitative and quantitative results, we can see that SIFT can match the optical images with temporal differences, but it can hardly be applied to other types of multimodal datasets. Compared with SIFT, PSO-SIFT performs well on optical–infrared image pairs, but its performance decreases dramatically on the image pairs with severe NID such as the optical–depth and optical–SAR image pairs. HAPCG handles the NID much better than SIFT and PSO-SIFT, but its performance is quite unstable. Even though RIFT shows a high success matching rate, it wastes lots of the extracted feature points, and its accuracy is relatively low. In contrast, 3MRS demonstrates high robustness to various multimodal remote sensing image pairs with different NID, obtaining the most significant number of correct matches with high accuracy.

In terms of SIFT, the main reason why it performs poorly on multimodal remote sensing image datasets is that it uses image gradients as the basis of feature detection and description, considering that the gradient information is sensitive to NID and noises. Therefore, the repetition rate of detected feature points on different images is relatively

low, the estimated primary orientation during feature description can be wrong, leading to inaccurate feature descriptor construction. For PSO-SIFT, it uses the second derivatives, improving the orientation estimation accuracy and performing well on optical–infrared image pairs. However, it cannot handle large NID, so its performance dramatically decreases when applied to optical–SAR and optical–depth image pairs. The thought of HAPCG is similar to SIFT, and the most significant difference is that HAPCG uses phase congruency rather than gradient information for image matching. It detects feature points on an anisotropic weighted moment graph using the Harris algorithm and constructs feature descriptors based on the phase congruency model. Even though it performs better than SIFT, its robustness to different kinds of NID and matching accuracy is still relatively low. In addition, the main orientation estimated by HAPCG is inaccurate in many situations due to the large NID, further decreasing its performance. Unlike HAPCG, RIFT does not calculate the main orientation directly. It constructs a feature descriptor based on a MIM feature map which is calculated using the phase congruency information and achieves orientation invariance by analyzing the influence of different orientation angles on the MIM feature. Even though RIFT has a highly successful matching rate, its accuracy is similar to HAPCG. Different from the above methods, 3MRS employs a coarse-to-fine matching strategy, where the coarse and fine matching stages have high robustness against NID. Based on phase congruency, 3MRS first utilizes the index information of the maximum value in all orientations of convolved images in the stage of coarse matching for fast matching, which is similar to RIFT. The rough transformation between the images can be estimated with the obtained few matches, which can be used to estimate the initial correspondences for all extracted feature points. Then, in the fine matching stage, we construct a template feature using the whole convolution image sequence rather than the index information, significantly improving the feature description ability. Under this framework, many feature points that are not matched during coarse matching are matched in the fine matching process, significantly improving the feature utilization rate and obtaining more correct matches. Moreover, the matches have high accuracy benefiting from the well-structured template features and 3D phase correlation similarity measure.

4.2. The Influence of Coarse Matching on the Final Result

As described before, the fine matching process is conducted on the basis of coarse matching, and this section discusses the effectiveness of coarse matching and the influence of the result of coarse matching on the subsequent fine matching process. Generally, the area-based fine matching requires a relatively good initial predicted matching location, and the initial correspondences are computed based on the transformation model calculated from the matches obtained from coarse matching. Theoretically, three high-accuracy matches are enough for calculating the affine transformation between an image pair, and more accurate transformation parameters are supposed to be achieved using large amounts of high-accuracy matches with the least square method. Nevertheless, when less than three matches are obtained during coarse matching, we cannot calculate the transformation between the image pair and carry out the subsequent fine matching process. Figure 11 shows the detailed experimental results of coarse matching on all experimental image pairs, and Table 3 gives the average statistical results of NCMs and RMSEs for each type of multimodal image pair. Results show that hundreds of NCM can be obtained for each image pair, with the average of NCM of each type of image pair larger than 300, the least number of NCM more than 100, and the most significant number of NCM larger than 1200. Besides, the RMSE of correct matches of most image pairs lies between 1.5 and 3.5 pixels, demonstrating high accuracy. These large amounts of high-precision matches can be reliably applied to estimate the transformation and provide initial correspondences for the following fine matching process.



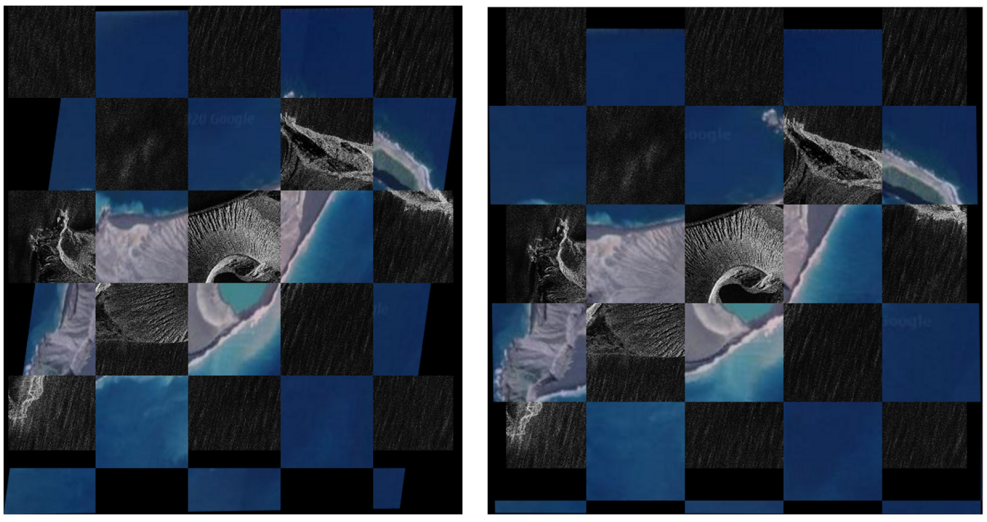
(**a**) (**b**)

Figure 11. Coarse matching results of 3MRS on the NCM (**a**) and RMSE metrics (**b**).

Table 3. Average results of NCM and RMSE in Figure 11 for each type of image pair.

Criteria	Optical–Optical	Optical–Infrared	Optical–Depth	Optical–Map	Optical–SAR	Day–Night
NCM_{ave}	542	548	560	503	321	351
$RMSE_{ave}$	2.19	2.25	2.20	2.43	3.40	2.71

There may be situations where enough matches are obtained, but significant errors involve the matches, leading to inaccurate estimation of image transformation. For example, the RMSE of the matches obtained from the coarse matching for the first image pair of optical and SAR images is 9.84 pixels; we deemed it a matching failure considering its value is larger than five. In this case, the initial predictions would severely violate the accurate locations. However, our method can still find correct correspondences. After the fine matching process, the RMSE improves to within four pixels, an acceptable range. Figure 12 demonstrates the registration checkboard images before and after fine matching. We can see that the wrong registration result is corrected using the proposed fine matching strategy. Therefore, our fine matching method has good robustness against the relatively bad coarse matching result.



(**a**) (**b**)

Figure 12. Registration checkboard overlays before (**a**) and after (**b**) fine matching.

4.3. Performance of 3MRS with Respect to Rotation and Scale Change

3MRS is designed for handling various types of nonlinear intensity differences, and the experimental datasets have no obvious geometric deformations. In terms of rotation, it will affect the performance of both the stages of coarse matching and fine matching. During the process of coarse matching, feature description is conducted using an index map constructed by finding the index of the maximum value in all orientations of convolved images obtained using a set of log-Gabor filters. For each orientation of the convolved image, the employed log-Gabor filters have a specific direction. Specifically, we apply six directions (0°, 30°, 60°, 90°, 120°, and 150°) of log-Gabor filters to convolve the original image. If rotation change exists between an image pair, the relative orientations between the two images and the same direction of the log-Gabor filter would be different. Even the convolved images of the same images with different rotation angles would be different using the exact orientation of the log-Gabor filter. As a result, the constructed feature map (as displayed in Figure 13), which uses the information of convolved images explicitly, will also be different.

Moreover, the rotation will affect the feature construction and matching process of fine matching. The temple feature cannot handle rotation, considering that the feature is also built based on the log-Gabor convolution image sequence. Besides, the 3D PC matching strategy is also sensitive to rotation: the larger the rotation angle, the less distinctive the peak value of the 3D phase correlation function. The peak value will not be detected if the rotation is significant and the matching process fails.

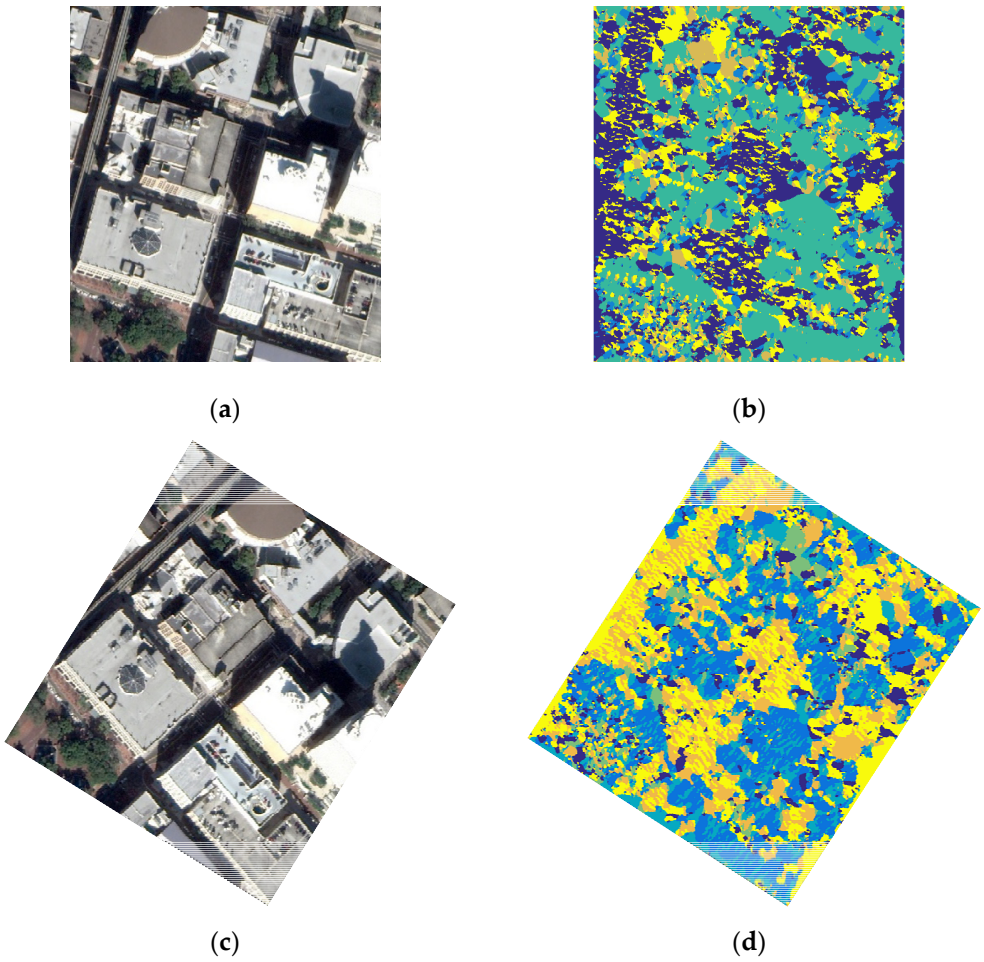


Figure 13. Feature description on an image pair with rotation: (**c**) is obtained from rotating (**a**) by 30 degrees; and (**b**,**d**) are the feature maps extracted from (**a**,**c**), respectively.

When the rotation angle is not large, there is still considerable overlap between the image pair and our feature descriptor is built based on statistics, increasing the image matching success rate. Therefore, we tested the robustness of 3MRS against rotation. A large number of experiments show that the performance of 3MRS will not decline significantly when the rotation is less than 20 degrees between the image pair, even though 3MRS does not consider the rotation explicitly. Figure 14 gives the matching and registration results under 10 and 17 degrees rotation angles, respectively. We can see that the number of matches decreases with the increase in rotation, but lots of correct matches are obtained. Moreover, we can still obtain a good registration result. Thus, 3MRS has good robustness against slight rotation.

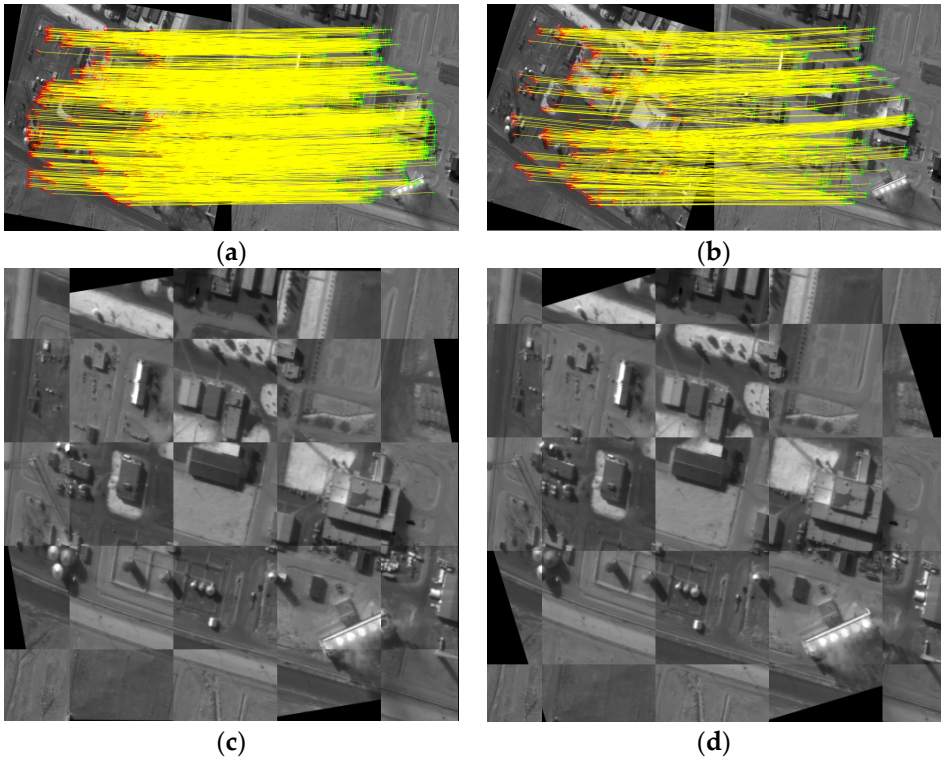


Figure 14. Comparison performance under rotation angles of 10 and 17 degrees: (**a**,**c**) are the matching and registration checkboard overlay results under a rotation angle of 10 degrees, while those under a rotation angle of 17 degrees are given in (**b**,**d**).

In terms of scale change, since 3MRS does not apply any scale-resistance strategies such as scale-space construction, it shows less robustness against scale change. The features constructed reflect distinctive information under different image contents at different scales. Besides, the 3D phase correlation matching strategy cannot handle scale change either. However, 3MRS still shows some robustness against scale change. As displayed in Figure 15, when the scale is slightly different, such as no more than 30 percent, relatively good matching results can still be achieved. The reason can be that the sizes of the window image used for constructing the feature descriptor in the coarse matching process and building temple feature in the fine matching process are relatively large. There is still extensive overlap between the image pair, enabling the success of image matching.

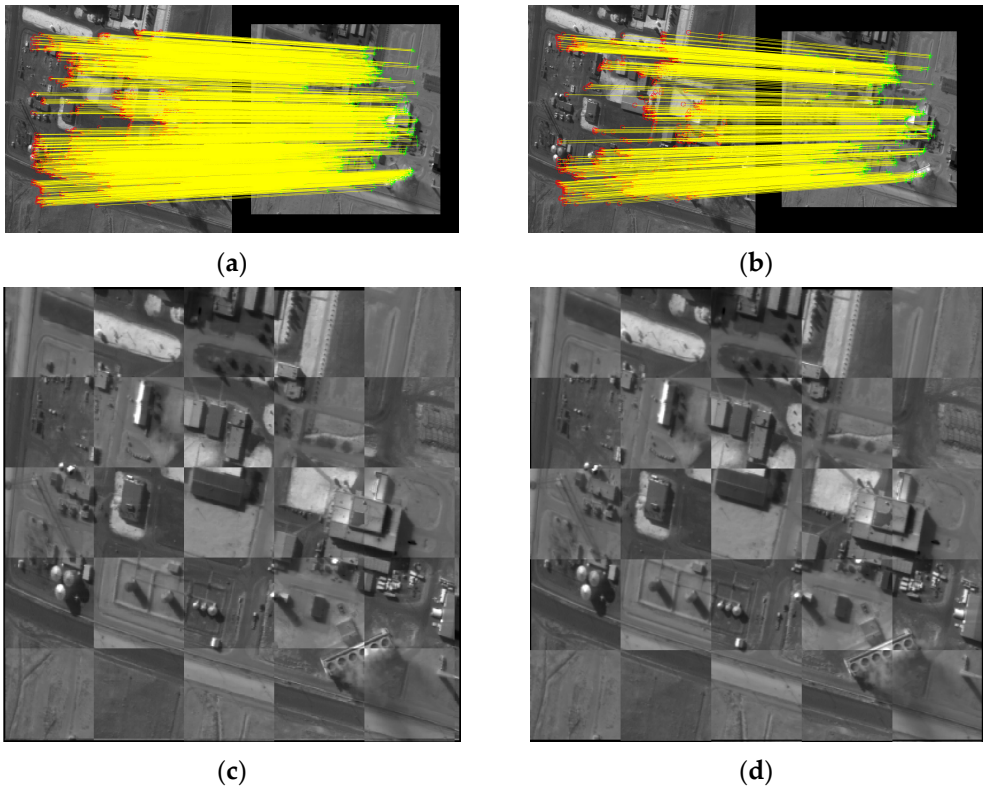


Figure 15. The matching and registration results on two image pairs with scale changes by 20 and 30 percent, respectively: (**a**,**c**) are the results of scale change by 20 percent, and (**b**,**d**) are those by 30 percent.

In real applications, relatively accurate prior information is available such as orbital position and attitude or rational polynomial coefficients (RPC) parameters provided with the remote sensing images. With this information, the geometric differences in scale and rotation angle between images can be calculated and roughly eliminated in advance. Therefore, 3MRS can be applied to real engineering applications.

5. Conclusions

Aiming to tackle the severe NID between multimodal remote sensing images, this paper proposed a novel and effective coarse-to-fine method, 3MRS, to match multimodal remote-sensed images. Firstly, feature extraction is conducted by calculating the maximum moment map from the 2D phase congruency model, and feature description is performed on an index map of all orientations of log-Gabor filter-convolved images. Then, feature match and outlier removal are conducted to complete image matching. Moreover, the obtained high accuracy matches are used to estimate a reliable transformation between the image pair. After that, the estimated transformation is used to predict the initial correspondences, which are further optimized using a newly developed template matching method that builds features from the log-Gabor convolution image sequence and employs a 3D phase correlation matching method.

According to the results of a large number of different types of multimodal images, the superiority of 3MRS on robustness, number of correct matches, and matching accuracy is proved through comparing it with four state-of-the-art matching methods, SIFT, PSO-SIFT, HAPGC, and RIFT. In detail, SIFT obtains the worst performance on all indices, revealing that it can hardly match multimodal remote sensing images. This is because the feature descriptors of SIFT are constructed based on the gradients, and there are significant differences in gradient direction and gradient magnitude between the two images due to NID. PSO-SIFT was initially designed to match optical–optical and optical–infrared image

pairs, and its resistance to NID is also limited. HAPCG has improved considerably more than SIFT and PSO-SIFT, but its performance is quite unstable. RIFT shows good robustness to different conditions of multimodal images and obtains many matches with relatively high accuracy. However, the matching rate of the extracted feature points is still low, and most of the feature points are unmatched.

On the contrary, 3MRS successfully matches all experimental image pairs and obtains the most significant correct matches with the highest accuracy. Moreover, 3MRS has the highest matching efficiency for obtaining a correct match. Therefore, it can be well applied to the task of multimodal remote sensing image matching. However, 3MRS cannot handle large geometric deformations such as rotation, and we will focus on improving the robustness of geometric changes in the future.

Author Contributions: Conceptualization, Z.F., Y.L. (Yuxuan Liu), and L.Z.; methodology, Z.F., Y.L. (Yuxian Liu), Y.L. (Yuxuan Liu), and L.Z.; software, Z.F., H.A., and J.Z.; validation, Y.L. (Yuxian Liu), J.Z., Y.S., and H.A.; formal analysis, Z.F., Y.L. (Yuxian Liu), and J.Z.; resources, Y.S., H.A., and L.Z.; writing—original draft preparation, Z.F., Y.L. (Yuxuan Liu), and J.Z.; writing—review and editing, Z.F., Y.L. (Yuxuan Liu), J.Z., and L.Z.; supervision, L.Z., Y.L. (Yuxuan Liu), and Y.L. (Yuxian Liu); project administration, Y.L. (Yuxuan Liu), and L.Z.; and funding acquisition, Y.L. (Yuxuan Liu), and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shenzhen Special Project for Innovation and Entrepreneurship, grant number: JSGG20191129103003903, and the Basic scientific research project of the Chinese Academy of Surveying and Mapping, grant number: AR2106.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Our implementation of the proposed method and the detailed experimental results is available on the website https://github.com/Zhongli-Fan/3MRS, accessed on 17 January 2022.

Acknowledgments: We would like to thank Yongxiang Yao at Wuhan University for making the CoFSM datasets available on https://skyearth.org/publication/project/HAPCG/, accessed on 17 January 2022. Moreover, we owe great appreciation to the anonymous reviewers for their critical, helpful, and constructive comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhou, K.; Lindenbergh, R.; Gorte, B.; Zlatanova, S. LiDAR-guided dense matching for detecting changes and updating of buildings in Airborne LiDAR data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 200–213. [CrossRef]
2. De Alban, J.D.T.; Connette, G.M.; Oswald, P.; Webb, E.L. Combined Landsat and L-band SAR data improves land cover classification and change detection in dynamic tropical landscapes. *Remote Sens.* **2018**, *10*, 306. [CrossRef]
3. Niu, X.; Gong, M.; Zhan, T.; Yang, Y. A conditional adversarial network for change detection in heterogeneous images. *IEEE Geosci. Remote Sens.* **2018**, *16*, 45–49. [CrossRef]
4. Touati, R.; Mignotte, M.; Dahmane, M. Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based Markov random field model. *IEEE Trans. Image Process.* **2019**, *29*, 757–767. [CrossRef] [PubMed]
5. Chen, H.; Li, Y.; Su, D. Multimodal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. *Pattern Recognit.* **2019**, *86*, 376–385. [CrossRef]
6. Sharma, M.; Dhanaraj, M.; Karnam, S.; Chachlakis, D.G.; Ptucha, R.; Markopoulos, P.P.; Saber, E. YOLOrs: Object Detection in Multimodal Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 1497–1508. [CrossRef]
7. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
8. Liu, S.; Qi, Z.; Li, X.; Yeh, A.G.-O. Integration of convolutional neural networks and object-based post-classification refinement for land use and land cover mapping with optical and SAR data. *Remote Sens.* **2019**, *11*, 690. [CrossRef]
9. Shao, Z.; Zhang, L.; Wang, L. Stacked sparse autoencoder modeling using the synergy of airborne LiDAR and satellite optical and SAR data to map forest above-ground biomass. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 5569–5582. [CrossRef]
10. Zhang, H.; Xu, R. Exploring the optimal integration levels between SAR and optical data for better urban land cover mapping in the Pearl River Delta. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *64*, 87–95. [CrossRef]

11. Campos-Taberner, M.; García-Haro, F.J.; Camps-Valls, G.; Grau-Muedra, G.; Nutini, F.; Busetto, L.; Katsantonis, D.; Stavrakoudis, D.; Minakou, C.; Gatti, L. Exploitation of SAR and optical sentinel data to detect rice crop and estimate seasonal dynamics of leaf area index. *Remote Sens.* **2017**, *9*, 248. [CrossRef]
12. Hong, D.; Hu, J.; Yao, J.; Chanussot, J.; Zhu, X.X. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 68–80. [CrossRef] [PubMed]
13. Jiang, X.; Ma, J.; Xiao, G.; Shao, Z.; Guo, X. A review of multimodal image matching: Methods and applications. *Inf. Fusion* **2021**, *73*, 22–71. [CrossRef]
14. Ma, W.; Zhang, J.; Wu, Y.; Jiao, L.; Zhu, H.; Zhao, W. A novel two-step registration method for remote sensing images based on deep and local features. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4834–4843. [CrossRef]
15. Li, Z.; Zhang, H.; Huang, Y. A Rotation-Invariant Optical and SAR Image Registration Algorithm Based on Deep and Gaussian Features. *Remote Sens.* **2021**, *13*, 2628. [CrossRef]
16. Merkle, N.; Luo, W.; Auer, S.; Müller, R.; Urtasun, R. Exploiting deep matching and SAR data for the geo-localization accuracy improvement of optical satellite images. *Remote Sens.* **2017**, *9*, 586. [CrossRef]
17. Hughes, L.H.; Marcos, D.; Lobry, S.; Tuia, D.; Schmitt, M. A deep learning framework for matching of SAR and optical imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 166–179. [CrossRef]
18. Ma, J.; Chan, J.C.-W.; Canters, F. Fully automatic subpixel image registration of multiangle CHRIS/Proba data. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 2829–2839.
19. Cole-Rhodes, A.A.; Johnson, K.L.; LeMoigne, J.; Zavorin, I. Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient. *IEEE Trans. Image Process.* **2003**, *12*, 1495–1511. [CrossRef]
20. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, USA, 25 June 2005; Volume 1, pp. 886–893.
21. Ye, Y.; Shan, J.; Bruzzone, L.; Shen, L. Robust Registration of Multimodal Remote Sensing Images Based on Structural Similarity. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2941–2958. [CrossRef]
22. Kovesi, P. Phase congruency detects corners and edges. In Proceedings of the Digital Image Computing: Techniques and Applications 2003, Sydney, Australia, 10–12 December 2003.
23. Ye, Y.; Bruzzone, L.; Shan, J.; Bovolo, F.; Zhu, Q. Fast and robust matching for multimodal remote sensing image registration. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9059–9070. [CrossRef]
24. Fan, Z.; Zhang, L.; Liu, Y.; Wang, Q.; Zlatanova, S. Exploiting High Geopositioning Accuracy of SAR Data to Obtain Accurate Geometric Orientation of Optical Satellite Images. *Remote Sens.* **2021**, *13*, 3535. [CrossRef]
25. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
26. Ma, W.; Wen, Z.; Wu, Y.; Jiao, L.; Gong, M.; Zheng, Y.; Liu, L. Remote sensing image registration with modified SIFT and enhanced feature matching. *IEEE Geosci. Remote Sens.* **2016**, *14*, 3–7. [CrossRef]
27. Yao, Y.; Zhang, Y.; Wan, Y.; Liu, X.; Guo, H. Heterologous Images Matching Considering Anisotropic Weighted Moment and Absolute Phase Orientation. *Geomat. Inf. Sci. Wuhan Univ.* **2021**, *46*, 1727–1736.
28. Li, J.; Hu, Q.; Ai, M. RIFT: Multimodal image matching based on radiation-variation insensitive feature transform. *IEEE Trans. Image Process.* **2019**, *29*, 3296–3310. [CrossRef]
29. Wu, Y.; Ma, W.; Gong, M.; Su, L.; Jiao, L. A novel point-matching algorithm based on fast sample consensus for image registration. *IEEE Geosci. Remote Sens.* **2014**, *12*, 43–47. [CrossRef]
30. Liu, Y.; Mo, F.; Tao, P. Matching multi-source optical satellite imagery exploiting a multi-stage approach. *Remote Sens.* **2017**, *9*, 1249. [CrossRef]
31. Li, Y.; Hu, Y.; Song, R.; Rao, P.; Wang, Y. Coarse-to-fine PatchMatch for dense correspondence. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 2233–2245. [CrossRef]
32. Lai, J.; Lei, L.; Deng, K.; Yan, R.; Ruan, Y.; Jinyun, Z. Fast and robust template matching with majority neighbour similarity and annulus projection transformation. *Pattern Recognit.* **2020**, *98*, 107029. [CrossRef]
33. Fischer, S.; Šroubek, F.; Perrinet, L.; Redondo, R.; Cristóbal, G. Self-invertible 2D log-Gabor wavelets. *Int. J. Comput. Vis.* **2007**, *75*, 231–246. [CrossRef]
34. Horn, B.; Klaus, B.; Horn, P. *Robot Vision*; MIT Press: Cambridge, CA, USA, 1986.
35. Rosten, E.; Porter, R.; Drummond, T. Faster and better: A machine learning approach to corner detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *32*, 105–119. [CrossRef] [PubMed]
36. Xiang, Y.; Tao, R.; Wan, L.; Wang, F.; You, H. OS-PC: Combining feature representation and 3-D phase correlation for subpixel optical and SAR image registration. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 6451–6466. [CrossRef]
37. Zhu, B.; Ye, Y.; Zhou, L.; Li, Z.; Yin, G. Robust registration of aerial images and LiDAR data using spatial constraints and Gabor structural features. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 129–147. [CrossRef]