applied sciences

MDPI

Article

Remote Sensing Multimodal Image Matching Based on Structure Feature and Learnable Matching Network

Songlai Han [1], Xuesong Liu [1], Jing Dong [1,*] and Haiqiao Liu [2]

[1] Research Institute of Aerospace Technology, Central South University, Changsha 410017, China
[2] School of Electrical and Information, Hunan Institute of Engineering, Xiangtan 411228, China
* Correspondence: dongjing@csu.edu.cn

Abstract: Matching remotely sensed multimodal images is a crucial process that poses significant challenges due to nonlinear radiometric differences and substantial image noise. To overcome these difficulties, this study presents a novel and practical template-matching algorithm specifically designed for this purpose. Unlike traditional approaches that rely on image intensity, the proposed algorithm focuses on matching multimodal images based on their geometric structure information. This approach enables the method to effectively adapt to variations in grayscale caused by radiometric differences. To enhance the matching performance, principal component analysis calculation based on the log-Gabor filter is proposed to estimate the structural feature of the image. The proposed method can estimate the structure feature accurately even under severe noise distortion. In addition, a learnable matching network is proposed for similarity measuring to adapt to the gradient reversal caused by the radiometric difference among remotely sensed multimodal images. Infrared, visible light, and synthetic aperture radar images are adopted for the evaluation, to verify the performance of the proposed algorithm. Based on the results, the proposed algorithm has a distinct advantage over other state-of-the-art template-matching algorithms.

Keywords: remotely sensed image; multimodal image; template matching; principal component analysis; structure feature

check for updates

Citation: Han, S.; Liu, X.; Dong, J.; Liu, H. Remote Sensing Multimodal Image Matching Based on Structure Feature and Learnable Matching Network. *Appl. Sci.* **2023**, *13*, 7701. https://doi.org/10.3390/app13137701

Academic Editors: Weitao Chen, Ailong Ma and Guohua Wu

Received: 26 May 2023
Revised: 19 June 2023
Accepted: 24 June 2023
Published: 29 June 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Multimodal image matching is the process of overlaying two or more images of the same scene captured by different sensors [1]. Since the imaging methods are based on different physical effects, remotely sensed multimodal images acquired by different sensors can capture different object characteristics, which provide complementary information. Multimodal image matching can integrate the complementary information by registering different multimodal images into one identical map, which is an important step for many remote sensing image processing tasks, such as image fusion [2], vision-based satellite attitude determination [3], and image-to-map rectification [4].

Automatic high-performance matching for remotely sensed multimodal images remains a problematic task because of the severe radiometric deformation produced by different types of sensors.

Traditional image-matching methods can be classified into two categories: feature-based [5–8] and area-based methods [9–11]. The scale-invariant feature transform (SIFT), which is invariant to scale and rotation changes, is the most representative feature-based method [12]. The SIFT-based method has been widely employed in the registration of remotely sensed multimodal images [6–8,13,14]. However, these SIFT-based descriptors were developed to handle the geometric affine variation of images with linear intensity changes. Therefore, these methods cannot resolve complicated nonlinear intensity changes caused by radiometric variations among remotely sensed multimodal images [14,15]. Ye et al. proposed a descriptor based on the local histogram of phase congruency to adapt

to nonlinear intensity variation [16]. This method can accurately register remotely sensed multimodal images if the overlapping areas in the images are sufficiently large.

Feature correspondence-based methods usually fail to match remotely sensed multimodal images because repeatable and reliable feature detection is considerably difficult for these images. To address this problem, locality-preserving matching [17] and matching based on local linear transformation [18] are proposed to achieve reliable feature correspondence.

The area-based method also called the correlation-like or template-matching method [1], compares the template to each candidate window on the base image; the corresponding window with maximum similarity is then selected as the matching result. Compared with the feature-based matching method, the area-based method exhibits better performance in matching images with few features or noise distortion [11].

There are two problems in using the area-based method for matching remotely sensed multimodal images. One is that some remotely sensed images contain intense noise. For example, synthetic aperture radar (SAR) images have randomly distributed speckle noise introduced by the interference of ground objects or surfaces to the backward reflection of electromagnetic waves.

Another problem is the significant nonlinear radiometric difference among remotely sensed multimodal images. This difference introduces nonlinear intensity variation, indicating that the same part of an object may be represented by different intensities in the images captured by different modalities [19]. It is impossible to calculate (even roughly) the intensity variation among multimodal images through a single mapping function [20]. Therefore, it is not ideal to match two remotely sensed multimodal images directly based on grayscale via commonly used similarity measurements techniques, such as the sum of squared differences (SSD), the sum of absolute differences (SAD), normalized cross-correlation (NCC), and matching by tone mapping (MTM) [21].

In addition, radiometric differences can cause gradient reversal, i.e., the gradients of corresponding parts of images change their orientation in the opposite direction [19], as shown in Figure 1. In this case, if the image feature is represented by a descriptor based on the texture gradient or structure orientation, then the two images of the same object display opposite geometric information. However, gradient reversal does not always occur, and the location may be unknown, making this problem more intractable.



(a) an infrared image (b) a visible light image

Figure 1. Gradient reversal between infrared and visible light images. Arrows show gradient orientation from brighter areas to darker areas.

In recent years, deep learning-based methods are proposed for addressing the challenge problems of image matching [22–26]. Their testing results show that the deep learning based matching methods can achieve significant improvement compared with the traditional matching methods.

The main advantage of deep learn based methods is that they employ convolutional neural networks to learn powerful feature descriptors which are more robust to appearance changes than the classical descriptors. However, the feature learning networks are usually pre-trained in large datasets such as ImageNet [27] which consists of visible light images with rich and clear features. Therefore, the performance of these deep learning-based

methods could drop significantly for matching SAR or infrared images, and the retraining or the fine tuning is also not idea ways if the application dataset is small.

In order to address the aforementioned challenges, this study introduces a template-matching algorithm that aims to achieve accurate matching of remotely sensed multimodal images. The algorithm proposed in this research enhances the estimation of structural features by incorporating principal component analysis (PCA) and employs a learnable matching network (LMN) to measure the similarity between two images. The primary contributions of this study can be summarized as follows:

1. Novel descriptor based on PCA-enhanced structure feature. As introduced, nonlinear intensity variation can significantly decrease the gray-level correlation among images. Instead of directly matching images based on the image grayscale, a descriptor based on the structure feature for capturing the geometric information of the image is introduced. Since the structure feature may be distorted by noise affection, a PCA-enhancing method is proposed to reduce the noise component in signals and estimate the local dominant orientation. The structure feature can be accurately calculated by the proposed descriptor even in images with severe noise distortion.
2. Improved similarity measurement based on LMN. Severe miscalculations can result if SAD or SSD is used to measure the similarity of the structure feature because the complicated radiometric variation causes gradient reversal. To solve this problem, a similarity measurement based on the LMN is proposed. The correlation layer and the regression network of the LMN can handle the gradient reversal and significantly improve the matching of remotely sensed multimodal images, as described in the experimental section.
3. A Novel combined matching method for application with a small dataset. It is very hard to train a deep convolutional network to extract robust cross-modal features with a small dataset. Therefore, the PCA-enhanced structure feature is adopted, which is a handcraft stable cross-modal feature. For addressing the complicated gradient reversal and radiometric variation between multi-modal images, we developed a light learnable matching network to learn the similarity measurement and regress the transformation parameters.

The remainder of this paper is organized as follows. Section 2 introduces related works. The proposed PCA–LMN template-matching algorithm is described in Section 3. In Section 4, the performance of the proposed algorithm is evaluated. Conclusions are presented in Section 5.

2. Related Work

Complex grayscale variation is a major problem for area-based multimodal image matching. Some template-matching algorithms have attempted to solve this problem by improving the similarity measurement [21,28,29]. These algorithms usually assume that the gray distortion caused by different imaging conditions, or the spectral sensitivity of sensors can satisfy a mapping model [1]. Therefore, gray distortion can be resolved by developing a similarity measurement that ignores the grayscale variation, conforming to the mapping model. The NCC, which is invariant to linear gray changes, is the most commonly used similarity measurement approach for adapting gray distortion among images [29]. Even under conditions with monotonic nonlinear gray variation, the NCC usually performs well, because these variations can be typically assumed as locally linear. However, the NCC cannot handle complex gray distortions, such as non-monotonic nonlinear gray differences, or situations where the gray mapping between two images is not function mapping [21]. Visual examples of gray mapping between remote sensing multimodal images can be found in [30].

Hel-Or proposed a fast-matching measurement called MTM [21], which is invariant to nonlinear gray variation. It can be regarded as a generalization of the NCC for non-linear mappings and reduces to the NCC when the mappings are linear. Although the computational time of MTM is the same as that of the NCC, it exhibits better matching

performance. However, the MTM also assumes that the grayscale mapping between two images is function mapping.

The mutual information (MI) technique is a similarity measurement approach commonly used in multimodal image matching [28]. This technique measures the gray statistical dependency between two images, without requiring their grayscale mapping to be function mapping. Moreover, compared with the NCC and MTM, MI affords advantages in adapting the nonlinear gray variation among multimodal images [10]. However, it requires the construction of a local histogram for each candidate window during the search process, thereby leading to high computational costs. Additionally, the MI technique is sensitive to the size of histogram bins for joint density estimation [21].

Measurement improvement is not the only approach for resolving multimodal image matching. Some area-based methods match remotely sensed multimodal images with dense feature descriptors based on structural information.

The histogram of oriented gradient (HOG) is a commonly used descriptor that employs the orientation and amplitude of gradients to capture the structural features of an image [31]. This descriptor was successfully applied to many image-matching methods. Sibiryakov proposed a template-matching algorithm based on the projected and quantizing histograms of oriented gradients (PQ-HOG). It transforms the images into dense binary codes to improve their computational efficiency [32]. The HOG is considerably resistant to illumination change or contrast variation; however, it cannot adapt to the complex nonlinear grayscale distortion among remotely sensed multimodal images. In addition, the gradient-based descriptor is usually sensitive to image noise.

Schechtman and Irani proposed the local self-similarity (LSS) descriptor [33], which had been previously applied to various template-matching methods [14,34]. However, the LSS cannot effectively capture informative features for multimodal matching in textureless areas [35], and its discriminative power is considerably limited [36].

The phase congruency model proposed by Kovesi [37] can capture the structure magnitude of the image, which is invariant to the complex nonlinear grayscale distortion among multimodal images. However, this model cannot capture the structure orientation of an image which is crucial for multimodal image matching. To solve this problem, Ye et al. extended the phase congruency model to build a dense descriptor called histogram of oriented phase congruency (HOPC) [9]. They used the log-Gabor odd-symmetric wavelets to calculate the orientation of phase congruency and construct a descriptor using the orientation and amplitude of phase congruency.

Compared with the HOG, the HOPC is more robust for matching remotely sensed multimodal images. Ye et al. demonstrated that the performance of the template-matching algorithm based on the HOPC is superior to those based on the NCC, MTM, or MI for remotely sensed multimodal images [9]. Recently, a novel template-matching method based on the channel features of oriented gradients (CFOGs) was proposed. This novel feature is an extension of the pixel-wise HOG descriptor [35]. Compared with the HOPC, the CFOG is more robust and efficient in matching multimodal images [35]. However, both the HOPC and CFOG handle the gradient reversal in a problematic manner, as described in part 2 of Section 3. Furthermore, they are sensitive to noise distortion, as discussed in Section 4.

In recent years, deep learning-based methods are proposed for matching multimodal images or aerial images. X. Han et al. [23] proposed a unified approach for feature and metric learning, dubbed Match-Net. They developed a deep convolutional network to extract features from images and a network of three fully connected layers to measure the similarity. Match-Net can achieve better performance compared with the state-of-the-art handcraft methods according to their testing results. I. Rocco et al. [24] proposed a trainable end-to-end matching network, which is not just for learning the feature and the similarity, but also estimating the transformation parameters with a regression network. This method is further developed for aerial image matching in [25,26], and the testing results confirmed

that deep learning-based matching methods can achieve significant improvement compared with the traditional matching methods.

3. Template Matching Based on PCA–LMN

The proposed matching algorithm is a full-search template-matching method that compares the template with a candidate window of the same size on the base image to identify the position of the target window (Figure 2). Since the method only searches in translation, a preliminary correction must be performed before the matching, so the direction and scale of the template and the direction and scale of the base image are approximately the same. Usually, the preliminary correction can be automatically performed according to the altitude and attitude information provided by the onboard navigation system [38].

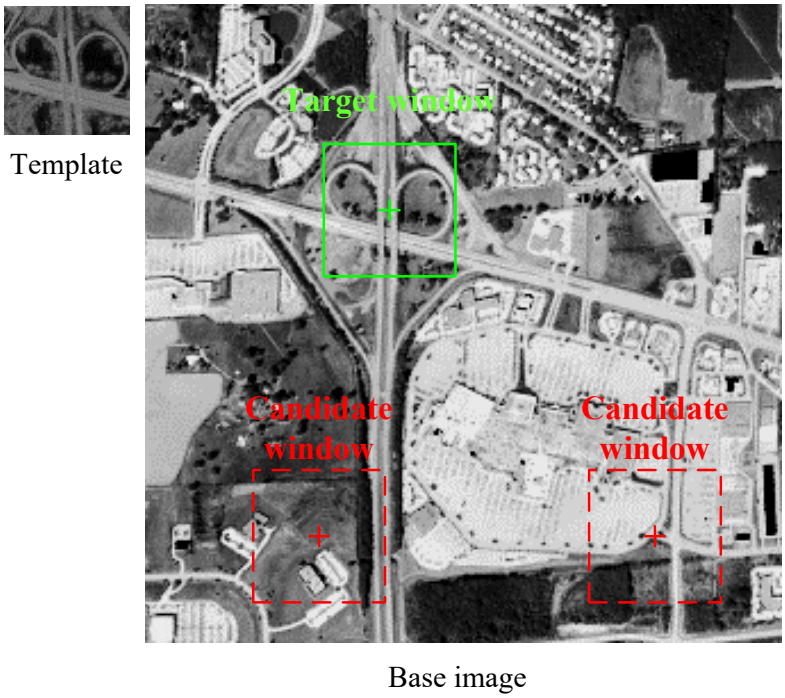


Figure 2. Example of template matching with the multimodal image.

The template-matching algorithm based on the PCA–LMN consists of three main steps: log-Gabor orientation estimation, orientation enhancement using the PCA, and similarity measurement and translation estimation based on the LMN; the algorithm flowchart is shown in Figure 3.

3.1. Orientation Estimation and Enhancing

A significant challenge in matching remotely sensed multimodal images lies in the severe distortion of grayscale relationships among the images. As shown in Figure 4, the grayscale change between infrared image (a) and visible light image (b) is considerably significant; however, the structural orientation, which is present in images (c) and (d), is considerably more stable than the gray part of the image. Accordingly, the structure orientation based on log-Gabor is employed by some methods [7,9] for matching multimodal images. These methods exhibit a significant advantage in terms of matching performance over template-matching algorithms directly based on the image grayscale.

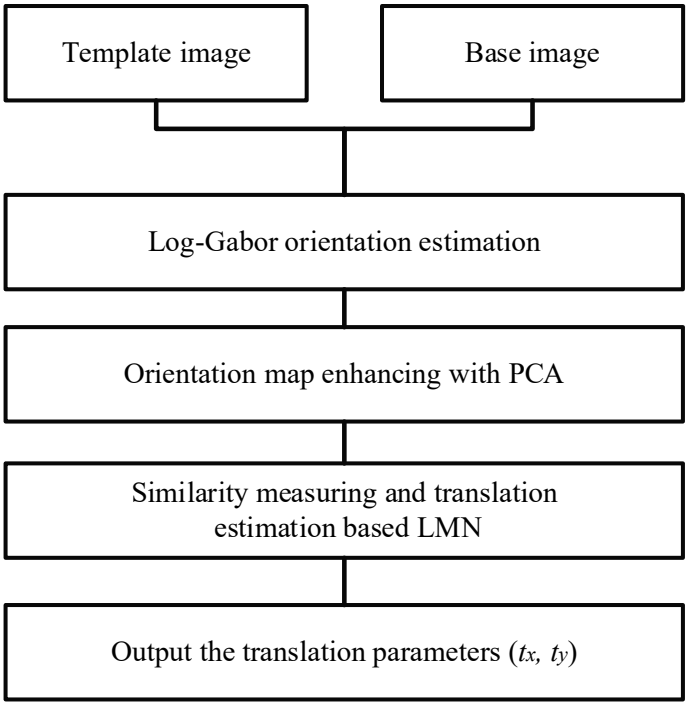


Figure 3. Flowchart of PCA–LMN template-matching algorithm.

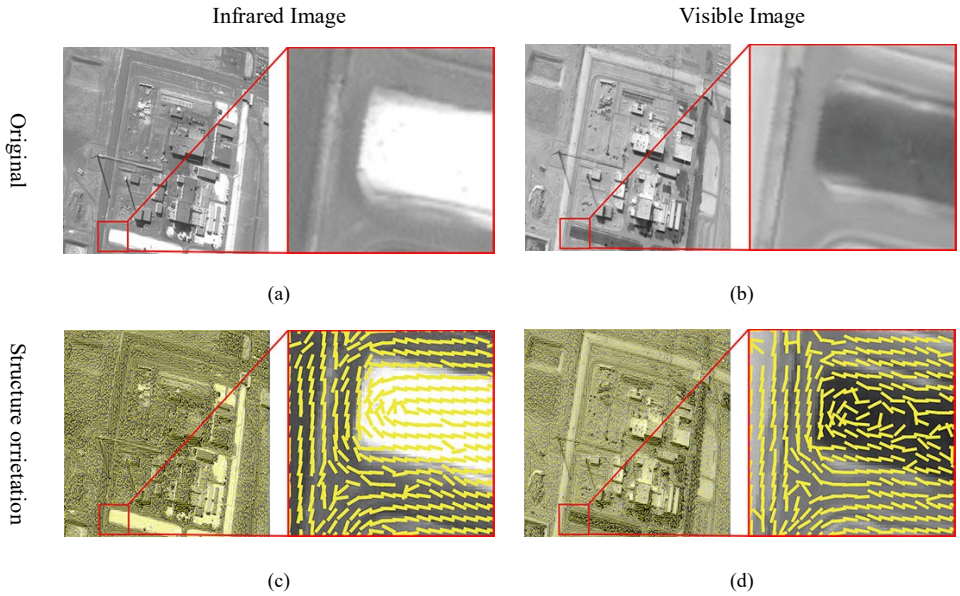


Figure 4. Multimodal images and their orientation maps: (**a**) infrared image; (**b**) visible light image; (**c**) orientation map of infrared image; and (**d**) orientation map of visible light image. Short yellow lines indicate the structure orientation direction, which is also the image edge direction.

However, the structure orientation estimated with the log-Gabor filter is sensitive to noise distortion. The orientation map is estimated via log-Gabor filters. Note that the orientation map is disturbed by noise distortion.

To improve the noise adaptiveness, the PCA is employed to enhance the structure orientation estimated using log-Gabor. The PCA is typically used to calculate the dominant vectors of a given dataset that can reduce the noise component in signals and estimate the local dominant orientation.

For each pixel, the PCA can be applied to the local gradient vectors to obtain their local dominant direction. In general, the PCA can be implemented in two ways: eigenvalue decomposition (EVD) of the data covariance matrix and singular value decomposition

(SVD) of the data matrix. In this work, because of the superiorities in flexibility and robustness [39], SVD is employed to calculate the PCA.

Given the original image (I) with its horizontal derivative image (a) and vertical derivative image (b), an $N \times 2$ local gradient matrix is constructed for each pixel:

$$G = \begin{bmatrix} \mathrm{A}^T & \mathrm{B}^T \end{bmatrix} \tag{1}$$

where N is determined by the size of the local window of the PCA calculation. For example, if the size of the window for each pixel is 3×3, then the value of N is 9. The vectors of the local derivatives, i.e., A and B, can be calculated using the following expression.

$$\mathrm{A} = \{a_1 a_2, \ldots, a_N\},\ \mathrm{B} = \{b_1 b_2, \ldots, b_N\} \tag{2}$$

where $a_{1,2,\ldots,N}$ can be calculated according to Equation (1) and $b_{1,2,\ldots,N}$ can be calculated according to Equation (2).

The dominant orientation can be estimated by determining a unit vector, u, perpendicular to the local gradient vectors (Figure 5). This can be formulated as the following minimization problem.

$$\left\| u^T G \right\| = \sum_{i=1}^{N} [a_i, b_i]^T u \tag{3}$$

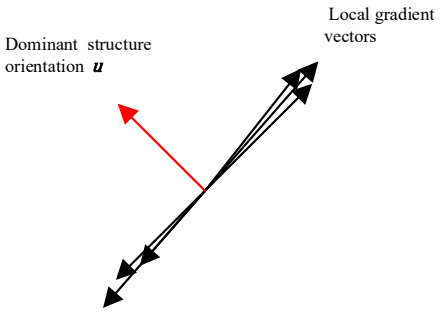


Figure 5. Dominant structure orientation and local gradient vectors.

This can be solved by applying SVD to the local gradient matrix, G.

$$G = USV^T \tag{4}$$

where U is an $N \times N$ orthogonal matrix; S is an $N \times N$ matrix; $V = \begin{bmatrix} v & u \end{bmatrix}$ is a 2×2 orthogonal matrix; and v indicates the local dominant orientation of the local gradient vectors. The PCA-enhanced structure orientation can be calculated according to the following equation.

$$\varphi = \frac{1}{2} \angle (u_1, u_2) \tag{5}$$

where $\varphi \in [0°, 180°)$, u_1 is the element in the first row of u and u_1 is the element in the second row of u. Note that the structure orientation is orthogonal to the orientation of the gradient vector.

The structure orientation enhanced by PCA is more robust against noise than the structure orientation directly estimated with log-Gabor filters. The enhanced structure orientation, which can be represented by (u_1, u_2), is estimated for each pixel. Suppose the size of an image is $w \times h$, and the size of its feature map is $w \times h \times 2$.

3.2. Learnable Matching Network

The gradient reversal among remotely sensed multimodal images is considerably common. This causes orientation reversal (Figure 1), which is a critical problem for the similarity measurement of structure orientation.

To solve this problem, some methods remap the structure orientation to range [0–180°] by adding 180° to the negative value [9,10,35]. However, miscalculations can result if the orientation is not appropriately reversed. For example, suppose that a structure orientation vector in one image is 5°. The orientation of the corresponding structure in the other image should be changed to −5° because of the noise effect. According to the remapping rule, the orientation difference between them is 170°, which is evidently unreasonable. The SAR and infrared images sometimes contain intense noise; hence, these methods may encounter problems in matching the images.

Instead of changing the orientation mapping, a learnable matching network (LMN) is proposed, which consists of a correlation layer and a regression network.

3.2.1. Correlation Layer

Suppose the feature map of a base image is $f_B \in R^{w_b \times h_b \times 2}$, and the feature map of a template is $f_T \in R^{w_t \times h_t \times 2}$, the correlation layer between them is shown in Figure 6.

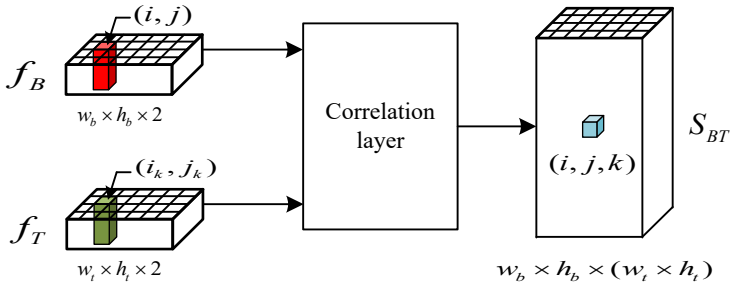


Figure 6. Simulation map computation by correlation layer.

The similarity between the feature map of the base image and the feature map of the template is calculated with the following equation,

$$S_{BT}(i,j,k) = \mathrm{f_B}(i,j)^T \mathrm{f_T}(i_k, j_k) \tag{6}$$

where (i,j) indicates the individual feature position in the feature map of the base image, and (i_k, j_k) indicates the individual feature position in the feature map of the image. The correlation layer output S_{BT} contains all pairs of similarities between individual features of $\mathrm{f_B} \in f_B$ and $\mathrm{f_T} \in f_T$.

3.2.2. Regression Network

The similarity map is passed through a regression network for translation estimation, which can be represented by the function F:

$$F : R^{w_b \times h_b \times w_t \times h_t} \rightarrow R^{\mathrm{n}} \tag{7}$$

where n is the number of parameters to regress, $n = 2$ for translation.

As shown in Figure 7, The regression network consists of two blocks, each block contains a convolutional layer, followed by batch normalization and ReLU. The last layer is a fully connected layer that regresses to the parameters of the transformation.

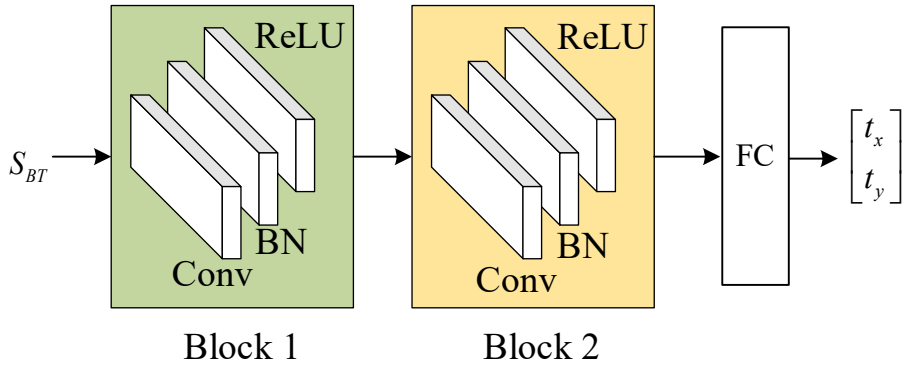


Figure 7. Transformation parameters estimation based on regression network.

3.2.3. Loss Function

Each training pair includes a template and a base image. Suppose the ground truth transformation between them is H_{gt}, and the transformation estimated between the sub-areas of this training pair is $\hat{H}_k$, the loss function is calculated by the following:

$$loss = \sum_{k=1}^{4} \sum_{i=1}^{N} ||H_{gt}(x_i, y_{i'}) - (\hat{H}_k(x_i - t_x, y_i - t_y)) + (t_x, t_y)||^2 \quad (8)$$

where N is the number of grid points, and (t_x, t_y) is the translation between the sub-area and the template, as shown in Figure 7. The loss function is based on the transformed grid loss [24], which minimizes the discrepancy between the estimating transformation and the ground truth transformation.

The partition approach increases the resolution of the input image and the resolution of the feature extraction, which helps improve the matching precision. In the partition approach, the four subarea pairs share the regression network, which means the number of the parameters of the network is not increased. This facilitates the retraining processing and the deployment of the network, which will eventually enhance the matching performance with a small training dataset.

The other way of enhancing the precision is inputting image with high resolution, but that may introduce an enormous increase of parameters to the network, which makes the retraining process (for cross-modal images) difficult and eventually lower the performance of the inference.

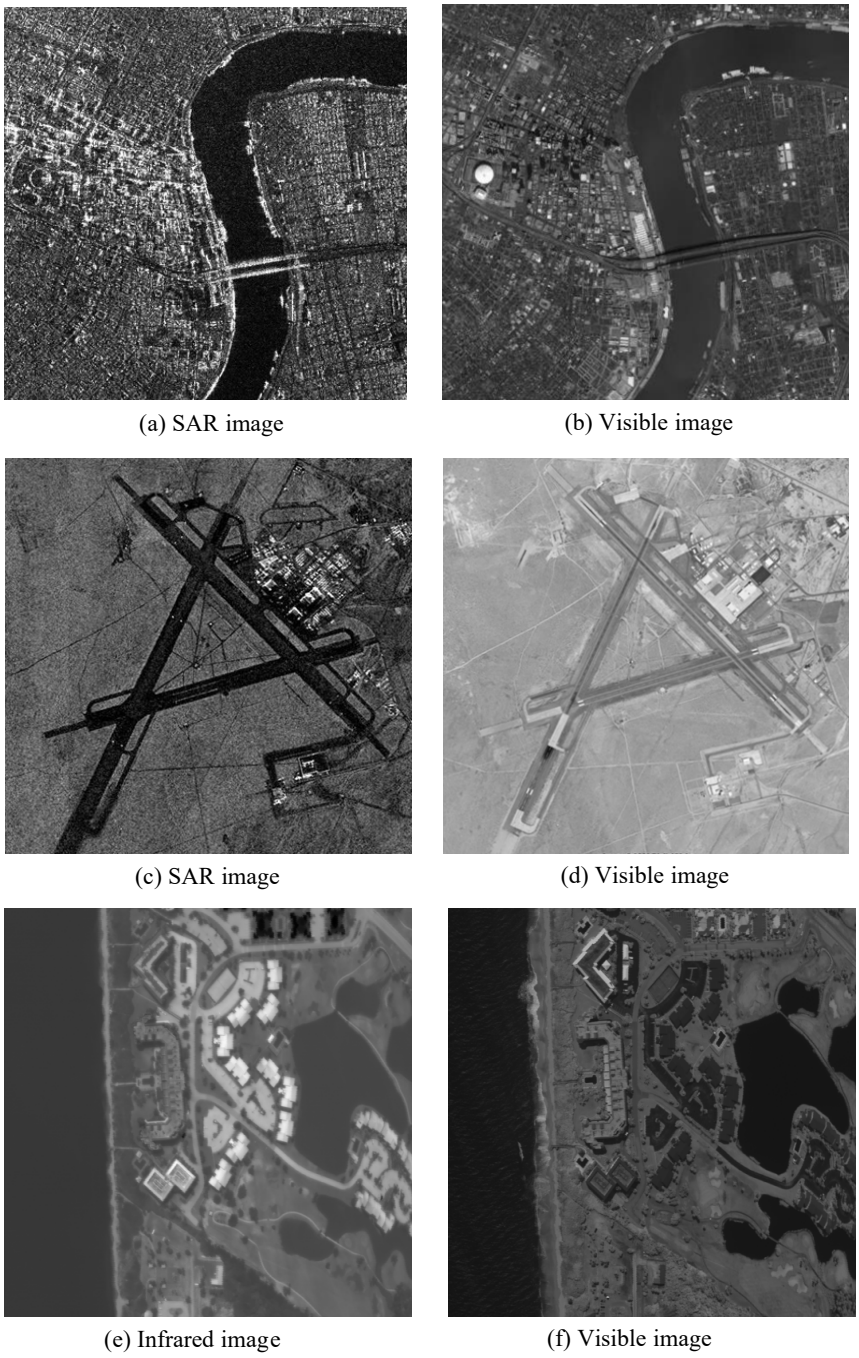
4. Experiment

The evaluation and comparison of the matching performance of PCA–LMN with the MTM [21], MI [28], HOPC [9], CFOG [35], and deep homography estimation (DHE) [26] are presented in this section. MTM, MI, HOPC, and CFOG are traditional image-matching methods. MTM and MI match images directly based on image gray, while HOPC and CFOG match images based on handcraft features. PCA–LMN is also based on the handcrafted feature, but it employs a learnable matching network to measure the similarity and regress the transformation parameters. DHE is a deep learning-based end to end-trainable method.

For the ablation study, PCA–LMN is also compared with the PCA–SAD, which uses the similarity measurement based on the SAD, and LGO–LMN, which directly estimates the structure orientation based on the log-Gabor filters.

4.1. Dataset and Training

To test the proposed algorithms, 200 remotely sensed multimodal image pairs were used, which were taken from areas such as urban airports, plantations, harbors, and hilly terrain. Some examples of multimodal image pairs are shown in Figure 8. Significant radiometric differences among these remotely sensed multimodal images are observed. In addition, the SAR image contains severe noise distortions and lacks details.



(a) SAR image (b) Visible image

(c) SAR image (d) Visible image

(e) Infrared image (f) Visible image

Figure 8. Examples of multimodal image pairs.

For each remotely sensed image, 100 templates of different sizes were randomly selected and matched to the base image. Since the dataset was small, 60% of the samples were employed for training and 20% of the samples were employed for validation and 20% were employed for testing. The final result was generated with the testing set. Data augmentation techniques such as grayscale variation, noise injection, and random erasing were adapted to the training set. Since our dataset was small, DHE used the pre-trained model provided by [25] and fine-tuned it with our dataset. The regression network of the LMN was totally trained with our dataset.

To evaluate the algorithms, 16 tests were performed. Table 1 summarizes the test information. Before the testing, the sensed image was manually corrected to the same coordinates as the base image. After the correction, the true position of the template in the based image and the position of the template selected from the sensed image were the

same. In addition, Gaussian noise with different variances was added to test the noise adaptiveness of the algorithms.

Table 1. Testing information.

Test	Variance of Gaussian Noise	Size of Base Image	Size of Template
T_1	Without noise	512 × 512	64 × 64
T_2	Without noise	512 × 512	96 × 96
T_3	Without noise	512 × 512	128 × 128
T_4	Without noise	512 × 512	160 × 160
T_5	0.01	512 × 512	64 × 64
T_6	0.01	512 × 512	96 × 96
T_7	0.01	512 × 512	128 × 128
T_8	0.01	512 × 512	160 × 160
T_9	0.03	512 × 512	64 × 64
T_{10}	0.03	512 × 512	96 × 96
T_{11}	0.03	512 × 512	128 × 128
T_{12}	0.03	512 × 512	160 × 160
T_{13}	0.05	512 × 512	64 × 64
T_{14}	0.05	512 × 512	96 × 96
T_{15}	0.05	512 × 512	128 × 128
T_{16}	0.05	512 × 512	160 × 160

4.2. Evaluation Criteria

The correct matching rate (CMR) was selected as the evaluation criterion; CMR = CM/M; where M is the number of total matched point pairs, and CM is the number of correctly matched results. A matching result is correct if the overlapping area ratio (OAR) between the matching and true positions reaches 90%. The OAR is calculated according to the following equation:

$$OAR = \frac{f(TW - \triangle x) f(TW - \triangle y)}{TW^2} \tag{9}$$

where TW is the template length; Δx and Δy denote the errors between the matching and true positions, respectively; and $f(\cdot)$ is a truncate function, which is defined as follows:

$$f(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{10}$$

4.3. Results and Analysis

The matching results of the tests (i.e., T_{1-4}) are shown in Figure 9. Notably, the CMRs of the investigated algorithms increase with the size of the templates. This is because a larger template aids in avoiding repetitive patterns in the base images.

Overall, the PCA–LMN achieved the best matching performance in the tests. The average CMR of PCA–LMN was 91.69%, which is 10.63% higher than that of the CFOG (the best traditional method in this test). It is assumed that this is because the PCA–LMN benefits from the PCA orientation enhancement and LMN. The PCA-enhanced orientation can accurately capture the structural features, even in the presence of significant noise distortions. This makes PCA–LMN more reliable in matching multimodal images with noise distortion, such as the optical SAR matching pairs in Figure 10c. In addition, LMN can be trained to adapt to the gradient reversal caused by the radiometric difference among remotely sensed multimodal images. The measurement function trained from LMN can be more sophisticated and accurate than the remapping function employed by the CFOG and HOPC.

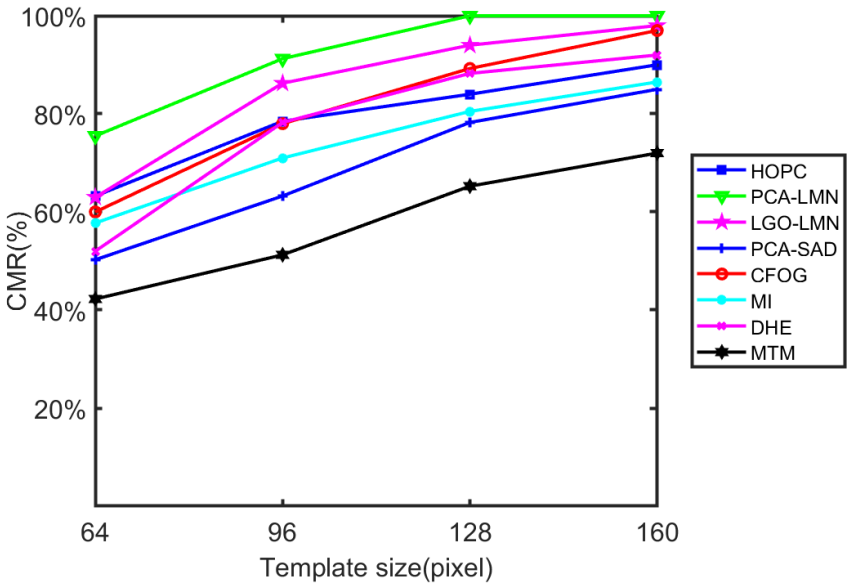


Figure 9. Matching results of test T1–T4.

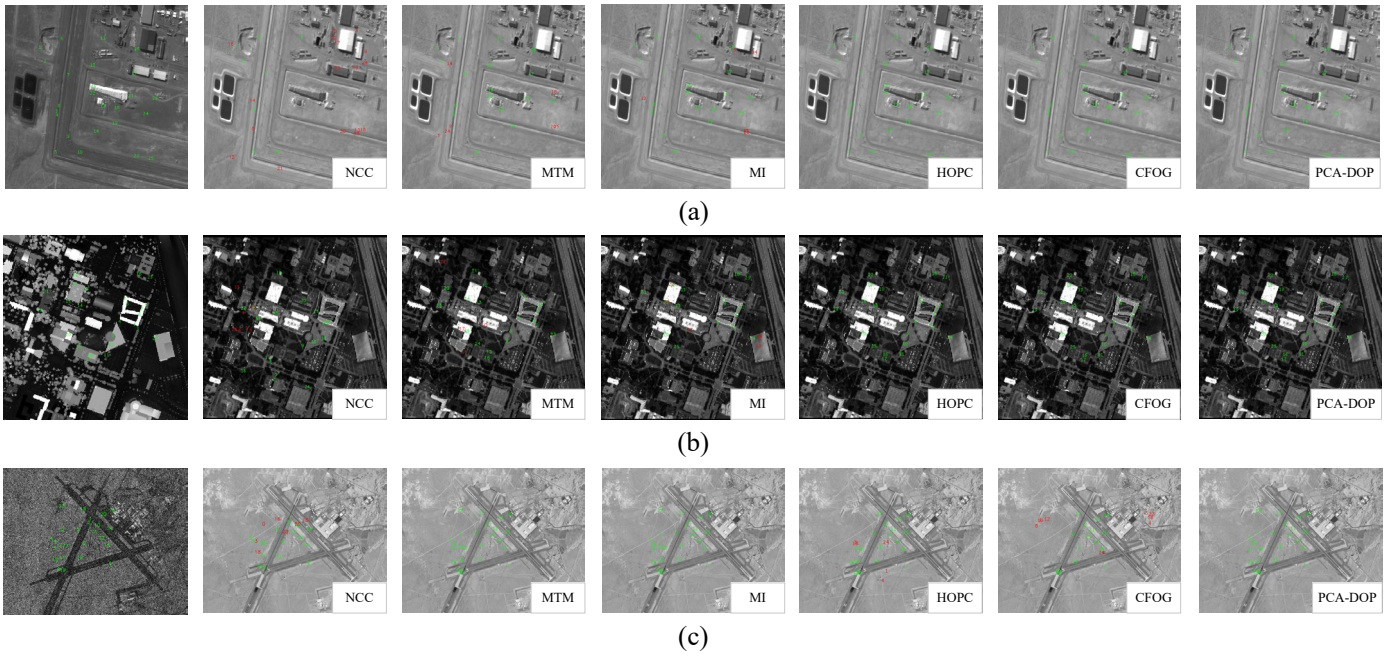


Figure 10. (**a**–**c**) Matching results of investigated algorithms on multimodal images. The first column shows true positions on the base image; columns 2–7 show matching positions of investigated algorithms on the sensed image; red and green numbers indicate erroneous and correct matching results, respectively.

The performance of DHE is not ideal in the tests, the average CMR of DHE is 3.44% lower than that of the CFOG and 14.07% lower than PCA–LMN. We believe the reason for these results is that the testing dataset is too small to train the deep feature extraction network of DHM, while the pre-trained dataset is not including multi-modal image pairs, which is very different from the testing dataset.

A clear contribution of the PCA can be observed by comparing the PCA–LMN and LGO–LMN. The average CMR of PCA–LMN is 6.38% higher than the average CMR of LGO–LMN in $T_{1–4}$, as shown in Figure 9. This is because the proposed PCA-enhanced method is more stable and can accurately capture the structure direction of remotely sensed multimodal images when compared with the log-Gabor orientation method, as described

in parts 1 and 2 of Section 3. A clear contribution of the LMN can be observed by comparing the PCA–LMN and PCA–SAD. The average CMR of PCA–LMN is 22.50% higher than that of the PCA–SAD in $T_{1–4}$, as shown in Figure 9. This indicates that gradient reversal is a significant problem in matching remotely sensed images, as emphasized by numerous other works in this field of research [9,10,35]. Clearly, the matching performance considerably can be improved by resolving the gradient reversal problem with the proposed LMN.

In tests $T_{5–16}$, Gaussian noise is added to the test images to evaluate the noise adaptability of the investigated algorithms; Figure 11 shows the matching results. Notably, the CMRs of the algorithms decrease as the noise level increases. The CMRs of the structure feature-based algorithms (i.e., HOPC, CFOG, and PCA–LMN) decrease faster than the three algorithms directly based on the image grayscale (MTM, and MI). This indicates that the structure feature-based methods are more sensitive to noise distortion than the methods directly based on the image grayscale. This trend is assumed to occur because structural features can be easily distorted by image noise. For example, the log-Gabor orientation used in the HOPC, and the gradient orientation employed by the CFOG, are considerably sensitive to noise distortion. However, the structure orientation of the proposed method is enhanced by the PCA, which has considerably better noise adaptiveness than the log-Gabor and gradient orientations. Therefore, compared with the CFOG and HOPC, the PCA–LMN shows a significant advantage in tests $T_{5–16}$. In these tests, the average CMR of PCA–LMN is 80.27%, which is 6.54% and 11.29% higher than that of the CFOG and HOPC, respectively.

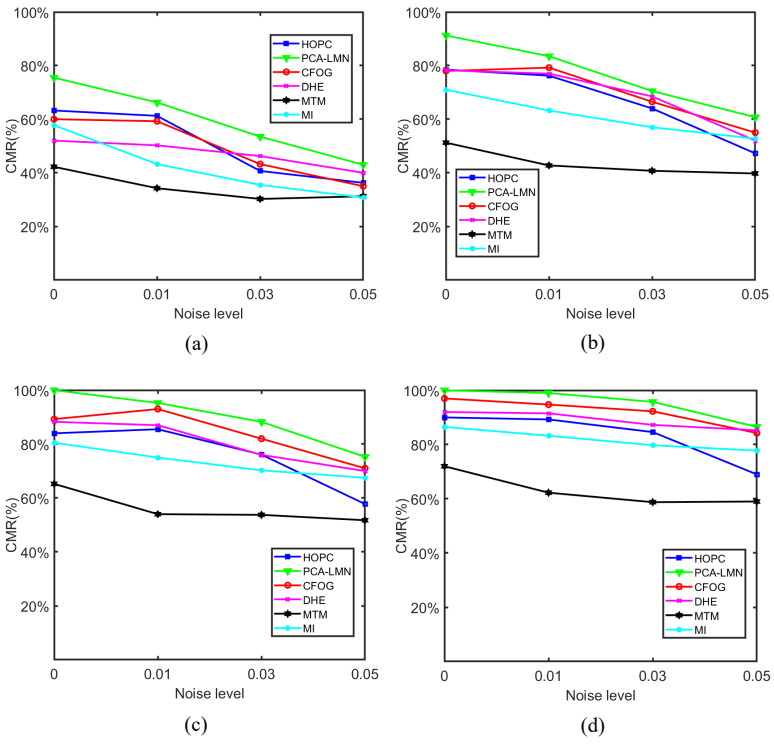


Figure 11. Matching results of T5–T16 template sizes: (**a**) 64 × 64; (**b**) 96 × 96; (**c**) 128 × 128; (**d**) 160 × 160.

5. Conclusions

In this paper, we propose a novel approach that combines PCA-based noise adaptiveness with a learnable matching network to address the challenge of matching remotely sensed multimodal images. Our method focuses on enhancing the noise adaptiveness of structural features and providing a robust trainable measurement of similarity while regressing transform parameters.

By integrating the learnable matching network with PCA-enhanced structure features, our proposed method effectively handles the complex radiometric variations that exist

among remotely sensed multimodal images. This adaptability is crucial in achieving robust image matching. As demonstrated in the experiments, the PCA–LMN approach achieves the best matching performance among all the methods evaluated. The average CMR achieved by PCA–LMN is 91.69%, which surpasses the CMR of CFOG, the best traditional method, by 10.63%.

The ablation study in Section 4.3 further revealed that the improved performance of PCA–LMN can be attributed to two main factors. Firstly, PCA–LMN benefits from the PCA orientation enhancement, which enables accurate capture of structural features even in the presence of significant noise distortions. This enhancement plays a vital role in matching multimodal images with challenging noise distortions. Secondly, LMN is trained to adapt to gradient reversal caused by radiometric differences among remotely sensed multimodal images. This adaptability allows for more precise measurements and better handling of radiometric variations compared to traditional methods.

Furthermore, our method offers the advantage of not requiring the training of a deep convolutional neural network for feature extraction. This makes it easy to train and deploy, and it can achieve stable performance even with small training datasets. The testing results in Section 4.3 show that PCA–LMN does have a significant advantage over DHE (which employs deep convolutional neural network for feature extraction) when the training dataset is small.

While our proposed method demonstrates superior performance, it is important to note that it lacks the ability to handle rotation and scale variations between images. As a prerequisite for accurate matching, it is necessary to correct the template's direction and scale to align them with the base image approximately. Failure to do so can result in changes to directional features due to rotation and alterations to the scale of extracted features caused by image scaling.

Author Contributions: Conceptualization, J.D. and S.H.; methodology, S.H. and J.D.; software, J.D., S.H. and H.L.; validation, S.H., J.D. and H.L.; formal analysis, X.L.; investigation, X.L.; resources, J.D.; data curation, X.L.; writing—original draft preparation, J.D., S.H. and X.L.; writing—review and editing, J.D. and X.L.; visualization, H.L.; supervision, J.D. and S.H.; project administration, H.L.; funding acquisition, S.H. and H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62203163; and in part by the Scientific Research Foundation of Hunan Provincial Department of Education under Grant 21B0661.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: There are no new data is created.

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

References

1. Zitová, B.; Flusser, J. Image registration methods: A survey. *Image Vis. Comput.* **2003**, *21*, 977–1000. [CrossRef]
2. Simone, G.; Farina, A.; Morabito, F. Image fusion techniques for remote sensing applications. *Inf. Fusion* **2002**, *3*, 3–15. [CrossRef]
3. Kouyama, T.; Kanemura, A.; Kato, S.; Imamoglu, N.; Fukuhara, T.; Nakamura, R. Satellite Attitude Determination and Map Projection Based on Robust Image Matching. *Remote Sens.* **2017**, *9*, 90. [CrossRef]
4. Dave, C.P.; Joshi, R.; Srivastava, S.S. A Survey on Geometric Correction of Satellite Imagery. *Int. J. Comput. Appl.* **2015**, *116*, 24–27.
5. Fan, J.; Wu, Y.; Li, M.; Liang, W.; Zhang, Q. SAR image registration using multiscale image patch features with sparse representation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *10*, 1483–1493. [CrossRef]
6. Yi, Z.; Cao, Z.; Yang, X. Multi-spectral remote image registration based on SIFT. *Electron. Lett.* **2008**, *44*, 107–108. [CrossRef]
7. Hossain, M.; Lv, G.; Teng, S.; Lu, G.; Lackmann, M. Improved symmetric-sift for multi-modal image registration. In Proceedings of the International Conference on Digital Image Computing Techniques and Applications (DICTA), Noosa, QLD, Australia, 6–8 December 2011; pp. 197–202.
8. Wang, L.; Niu, Z.; Wu, C. A robust multisource image automatic registration system based on the SIFT descriptor. *Int. J. Remote Sens.* **2012**, *33*, 3850–3869. [CrossRef]

9. Ye, Y.; Shen, L. HOPC: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. In Proceedings of the 23rd ISPRS Congress, Prague, Czech Republic, 16–19 July 2016; pp. 9–16.
10. Ye, Y.; Shan, I.; 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]
11. Korman, S.; Reichman, D.; Tsur, G.; Avidan, S. Fast-match: Fast affine template matching. *Int. J. Comput. Vis.* **2017**, *121*, 2331–2338. [CrossRef]
12. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
13. Sedaghat, A.; Ebadi, H. Distinctive order based self-similarity descriptor for multi-sensor remote sensing image matching. *ISPRS J. Photogramm. Remote Sens.* **2015**, *108*, 62–71. [CrossRef]
14. Ye, Y.; Shan, J. A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences. *ISPRS J. Photogramm. Remote Sens.* **2014**, *90*, 83–95. [CrossRef]
15. Kelman, A.; Sofka, M.; Stewart, C. Keypoint descriptors for matching across multiple image modalities and non-linear intensity variations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 3257–3264.
16. Ye, Y.; Shan, J.; Hao, S.; Bruzzone, L.; Qin, Y. A local phase based invariant feature for remote sensing image matching. *ISPRS J. Photogramm. Remote Sens.* **2018**, *142*, 205–221. [CrossRef]
17. Ma, J.; Zhao, J.; Jiang, J.; Zhou, H.; Guo, X. Locality preserving matching. *Int. J. Comput. Vis.* **2019**, *127*, 12–31. [CrossRef]
18. Ma, J.; Zhou, H.; Zhao, J.; Gao, Y.; Jiang, J.; Tian, J. Robust feature matching for remote sensing image registration via locally linear transforming. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6469–6481. [CrossRef]
19. Pluim, J.P.; Maintz, J.A.; Viergever, M.A. Image registration by maximization of combined mutual information and gradient information. *IEEE Trans. Med. Imaging* **2000**, *19*, 809–814. [CrossRef]
20. Kagarlitsky, S.; Moses, Y.; Helor, Y. Piecewise-consistent color mappings of images acquired under various conditions. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 2311–2318.
21. Hel-Or, Y.; Hel-Or, H.; David, E. Matching by tone mapping: Photometric invariant template matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 317–330. [CrossRef]
22. Ma, J.; Jiang, X.; Fan, A.; Jiang, J.; Yan, J. Image matching from handcrafted to deep features: A survey. *Int. J. Comput. Vis.* **2021**, *12*, 23–79. [CrossRef]
23. Han, X.; Leung, T.; Jia, Y.; Sukthankar, R.; Berg, A.C. Matchnet: Unifying feature and metric learning for patch-based matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
24. Rocco, I.; Arandjelovic, R.; Sivic, J. Convolutional neural network architecture for geometric matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
25. Park, J.H.; Nam, W.J.; Lee, S.W. A two-stream symmetric network with bidirectional ensemble for aerial image matching. *Remote Sens.* **2020**, *12*, 465. [CrossRef]
26. Oh, M.S.; Lee, Y.J.; Lee, S.W. Precise Aerial Image Matching based on Deep Homography Estimation. *arXiv* **2021**, arXiv:2107.08768.
27. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis. (IJCV)* **2015**, *115*, 211–252. [CrossRef]
28. Viola, P.; Wells, W. Alignment by maximization of mutual information. In Proceedings of the IEEE International Conference on Computer Vision, Cambridge, MA, USA, 20–23 June 1995; pp. 16–23.
29. Buntinga, P.; Labrosseb, F.; Lucasa, R. A multi-resolution area-based technique for automatic multi-modal image registration. *Image Vis. Comput.* **2010**, *28*, 1203–1219. [CrossRef]
30. Inglada, J.; Giros, A. On the possibility of automatic multisensor image registration. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 2104–2120. [CrossRef]
31. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2005, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
32. Sibiryakov, A. Fast and high-performance template matching method. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1417–1424.
33. Shechtman, E.; Irani, M. Matching local self-similarities across images and videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1744–1752.
34. Kim, S.; Min, D.; Ham, B.; Ryu, S.; Do, M.N.; Sohn, K. DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
35. Ye, Y.; Shan, I.; Bruzzone, L. Fast and robust matching for multimodal remote sensing image registration. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9059–9070. [CrossRef]
36. Kim, S.; Min, D.; Lin, S.; Sohn, K. Deep self-correlation descriptor for dense cross-modal correspondence. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part VIII 14*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 679–695.
37. Kovesi, P. Image features from phase congruency. *J. Comput. Vis. Res.* **1999**, *1*, 1–26.

38. Geometric Corrections in Remote Sensed Image [OL]. Available online: https://geolearn.in/geometric-corrections-in-remote-sensing-images/ (accessed on 25 May 2023).
39. Alexandris, N.; Gupta, S.; Koutsias, N. Remote sensing of burned areas via PCA, Part 1; centering, scaling and EVD vs. SVD. *Open Geospat. Data Softw. Stand.* **2017**, *2*, 17. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.