remote sensing

MDPI

Article

Two-Step Matching Method Based on Co-Occurrence Scale Space Combined with Second-Order Gaussian Steerable Filter

Genyi Wan [1,2,3], Ruofei Zhong [1,2,3,*], Chaohong Wu [1,2,3], Yusheng Xu [4], Zhen Ye [4] and Ke Yu [1,2,3]

[1] Key Laboratory of 3D Information Acquisition and Application, MOE, Capital Normal University, Beijing 100048, China
[2] Base of the State Key Laboratory of Urban Environmental Process and Digital Modeling, Capital Normal University, Beijing 100048, China
[3] College of Resource Environment and Tourism, Capital Normal University, Beijing 100048, China
[4] College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China
* Correspondence: zrf@cnu.edu.cn

Abstract: Multimodal images refer to images obtained by different sensors, and there are serious nonlinear radiation differences (NRDs) between multimodal images for photos of the same object. Traditional multimodal image matching methods cannot achieve satisfactory results in most cases. In order to better solve the NRD in multimodal image matching, as well as the rotation and scale problems, we propose a two-step matching method based on co-occurrence scale space combined with the second-order Gaussian steerable filter (G-CoFTM). We first use the second-order Gaussian steerable filter and co-occurrence filter to construct the image's scale space to preserve the image's edge and detail features. Secondly, we use the second-order gradient direction to calculate the images' principal direction, and describe the images' feature points through improved GLOH descriptors. Finally, after obtaining the rough matching results, the optimized 3DPC descriptors are used for template matching to complete the fine matching of the images. We validate our proposed G-CoFTM method on eight different types of multimodal datasets and compare it with five state-of-the-art methods: PSO-SIFT, CoFSM, RIFT, HAPCG, and LPSO. Experimental results show that our proposed method has obvious advantages in matching success rate (SR) and the number of correct matches (NCM). On eight different types of datasets, compared with CoFSM, RIFT, HAPCG, and LPSO, the mean SRs of G-CoFSM are 17.5%, 6.187%, 30.462%, and 32.21%, respectively, and the mean NCMs are 5.322, 11.503, 8.607, and 16.429 times those of the above four methods.

Keywords: multimodal image matching; nonlinear radiation distortions; co-occurrence filter; second-order Gaussian steerable filter

Citation: Wan, G.; Zhong, R.; Wu, C.; Xu, Y.; Ye, Z.; Yu, K. Two-Step Matching Method Based on Co-Occurrence Scale Space Combined with Second-Order Gaussian Steerable Filter. *Remote Sens.* **2022**, *14*, 5976. https://doi.org/10.3390/rs14235976

Academic Editor: Riccardo Roncella

Received: 25 October 2022
Accepted: 22 November 2022
Published: 25 November 2022

1. Introduction

Image matching refers to the detection of reliable corresponding feature relationships from images of the same scene collected at different times from different sensors or perspectives [1]. At present, image matching is widely used in medical image analysis [2,3], intelligent transportation [4], visual navigation and positioning [5,6], change monitoring [7–10], and other fields. Image matching has made great progress in recent years. However, most of the matching methods, such as scale-invariant feature transform (SIFT) [11], speeded-up robust features (SURF) [12], histogram of oriented gradient (HOG) [13], and others can only be applied to images collected by the same sensor device, and how to accurately align images of the same scene collected by different sensors is still a big challenge.

Multimodal images refer to images acquired by sensors with different imaging mechanisms, showing significant differences on the same ground object, usually with severe NRD, such as synthetic aperture radar (SAR) optical, infrared–optical, depth–optical, map–optical, etc. [1]. In order to solve the matching problem of multimodal images, scholars have carried

out a lot of research on it. Multimodal image matching methods are generally divided into area-based, feature-based, and deep learning. Area-based methods include mutual information (MI) [14] and histogram of orientated phase congruency (HOPC) [15,16], etc. HOPC uses phase congruency to highlight the edge features of the images, and robustly completes the matching work using template matching. Such methods generally use template matching, and it is difficult to obtain a good matching effect when the image is zoomed, rotated, and changed in perspective.

In contrast, the feature-based approach can better handle the scaling, rotation, and viewing angle differences between images. Dellinger et al. [17] proposed to use the ratio of pixel points instead of the difference to generate new gradient features, and to use log-polar coordinate descriptors to describe feature points. This method improves the adaptability of SIFT but still cannot deal with images with large speckle noise. Li et al. [18] used phase information to create a maximum index map (MIM), and the proposed radiation-invariant feature transform (RIFT) initially solves the problem of image rotation and perspective change. The method based on deep learning [19–21] has better matching speed and effect than the above two methods, but its matching effect depends on the training samples. It is not easy to obtain comprehensive and extensive training samples.

Due to the geometric distortion and NRD between multimodal images, obtaining the corresponding feature points in traditional matching methods is difficult. In contrast, the edge information of the image can better reflect the characteristics of the image and improve the similarity between the descriptors. For example, HOPC, RIFT, and NCFT [22] use phase information to extract the edge features of images to complete matching. A co-occurrence filter (CoF) [23] is an edge-preserving filter that aims to detect the boundaries between image textures and smooth edges within textured regions. CoF does not cross the boundaries between surfaces, and avoids smoothing across texture boundaries by using a normalized co-occurrence matrix to assign higher weights to frequently occurring pixel values. CoFSM [24] uses CoF to reduce the NRD between images, extract edge information, and perform robust matching. In this paper, in order to better solve the problems of NRD, rotation, and scale in multimodal image matching, we use the second-order Gaussian steerable filter combined with the co-occurrence filter to generate scale space to improve the edge information and detail features of the image. At the same time, for the problem where the distribution of matching points is not wide enough, a two-step matching strategy is introduced to improve the reliability and accuracy of image matching.

The main contributions of this paper are as follows:

1. A multimodal feature matching algorithm called G-CoFTM is developed, which is superior to the current state-of-the-art matching algorithms in terms of success rate, efficiency, and the number of correct matches.
2. We design a co-occurrence scale space combined with second-order Gaussian steerable filtering, which can improve the image similarity while better retaining the edge and detailed features of the image.
3. A two-step matching strategy is adopted, and the 3DPC descriptor is optimized to increase the number of correct matches and to reduce registration errors.

The rest of this article is organized as follows. Section 2 provides a review of existing multimodal image registration algorithms. Section 3 introduces our proposed G-CoFSM algorithm. Section 4 analyzes the experimental results of our algorithm and five other state-of-the-art algorithms on eight types of multimodal datasets. Section 5 presents the performance analysis of our algorithm and the comparison algorithms, and discusses the experimental results of our algorithm in the coarse matching stage. Finally, Section 6 provides our conclusions.

2. Related Works

In this section, we review the existing multimodal image matching algorithms in detail. According to the classification, multimodal image matching algorithms can be divided into area-based, feature-based, and deep-learning methods.

Area-based methods. The key to area-based methods is to establish an effective similarity measure by treating a predefined window on the image as a local template (or treating the entire image as a local template), and using this local template as a feature for matching. Traditional area-based matching methods include MI, the normalized cross-correlation (NCC) [25], and the sum of squared differences (SSD) [26]. These methods search for each region feature in the entire search space and complete the matching by comparing the similarity between the two selected region features. However, among the above algorithms, it is difficult to handle NRD for NCC and SSD between multimodal images [27]. Although MI can solve the impact of NRD to a certain extent, the calculation of MI is very time-consuming [28]. At the same time, MI also has the problem of easily falling into a local optimum. Aiming at the problem of NRD between multimodal images that cannot be handled in the spatial domain, scholars solve this problem by transforming the images into the frequency domain. Ye et al. proposed a novel pixel-level feature based on image-oriented gradients called the channel feature of the orientated gradients (CFOG) [29], which achieved good results. Xiang et al. combined robust features from optical and SAR images with 3D phase congruency (OS-PC) [30] to improve the robustness and accuracy of matching. Although the use of image frequency information can better solve the problem that traditional methods cannot handle NRD between multimodal images, region-based methods still fail when dealing with rotation, scaling, and viewing angle differences between images.

Feature-based methods. Feature-based methods extract salient structural features from images, including point, line, and surface features, and match them according to the similarity between the descriptors of each element. The current classic feature-based methods include SIFT, SURF, ORB [31], and so on. SIFT finds feature points in the scale space by constructing a Gaussian scale space, and uses the gradient histogram to describe the features. SIFT is widely used in the matching of optical images due to its robustness to illumination, rotation scale, and noise. Since SIFT was proposed, many derivative algorithms of SIFT have been developed. PAC-SIFT [32] utilizes principal component analysis to reduce the dimensionality of descriptors and reduce the space and time complexity. SAR-SIFT uses the ratio of pixel points instead of difference to generate new gradient features, and uses log-polar coordinate descriptors to describe feature points, which improves the adaptability of SIFT. Adaptive binning scale-invariant feature transform (AB-SIFT) uses adaptive column histograms to generate feature descriptors, making it better able to cope with radial geometric distortions [33]. Unlike SIFT and its derived algorithms, SURF uses a box filter with very little computation to replace the second-order Gaussian partial derivative, which speeds up the matching. However, these methods are generally not well suited for matching between multimodal images, which suffer from severe NRD [34]. Recently, the RIFT proposed by Li [18] can better solve the NRD between images of different modalities. Still, this method cannot handle the scaling problem between images because it does not construct scale space. At the same time, the performance of RIFT is not satisfactory for large rotation angles. To solve the scale problem between multimodal images, Yao et al. [35] proposed the histogram of absolute phase consistency gradients (HAPCG), which uses the phase consistency directions and Gaussian scale space to complete the matching problem of multimodal images. Yang et al. [36] used the modified phase sharpness direction as the main direction of the feature descriptor, and established a local phase sharpness orientation (LPSO) descriptor using log-polar coordinates. On this basis, to better solve the problem of rotation between images, Yao et al. [24] used CoF to reduce the NRD between images, extract edge information, and perform robust matching. However, although this method can handle scale and rotation differences to a certain extent, there is still the possibility of matching failure in some scenarios. Although the above feature-based methods have different degrees of robustness to translation, scaling, and rotation differences between multimodal images, the methods' performance decreases significantly when multiple problems exist simultaneously.

Deep-learning methods. With the rapid development of deep learning, deep learning has also been applied to the matching of multimodal images [37–39]. Ye et al. [40] used a deep convolutional neural network (CNN) combined with SIFT to generate combined features from images, and incorporated them into the PSO-SIFT [41] for matching. Ma et al. adopted a two-step matching strategy, first using CNN to calculate the spatial approximation between images. Then, combined with hand-crafted local features, the matching relation and transformation matrix are further adjusted [42]. Deep learning methods need to label and train many images, thus consuming many computing resources. At the same time, the lack of training sample scene types will also affect the final results. Therefore, deep-learning methods still require further research.

In a word, multimodal image matching has made great progress in template matching, classical feature methods, and deep learning, but there are still great challenges in multimodal image matching, mainly focusing on two aspects: (1) traditional feature descriptors cannot accurately describe feature points, leading to matching failure; and (2) since the NRD between multimodal images is significant, the main direction of images cannot be well calculated.

In order to solve the above two problems, this paper introduces a new multimodal image matching method, which uses the $G^{\theta}_{2,\sigma} - CoF$ construct the scale space of the image, better retains the contour information of the image and calculates the main direction, and uses optimized feature descriptors to enhance the description of feature points to achieve effective matching.

3. Methodology

This paper proposes a two-step matching algorithm based on co-occurrence scale space combined with the second-order Gaussian steerable filter called G-CoFTM. We first combine the second-order Gaussian steerable filter and co-occurrence filter to construct the scale space of the image. Then, the preliminary matching of the image is completed by using the second-order gradient of the image and the improved GLOH descriptor to describe the features. Finally, based on preliminary matching, a more accurate matching is performed by using the optimized 3DPC descriptor combined with the preliminary matching results. Figure 1 shows the flow of our entire method.

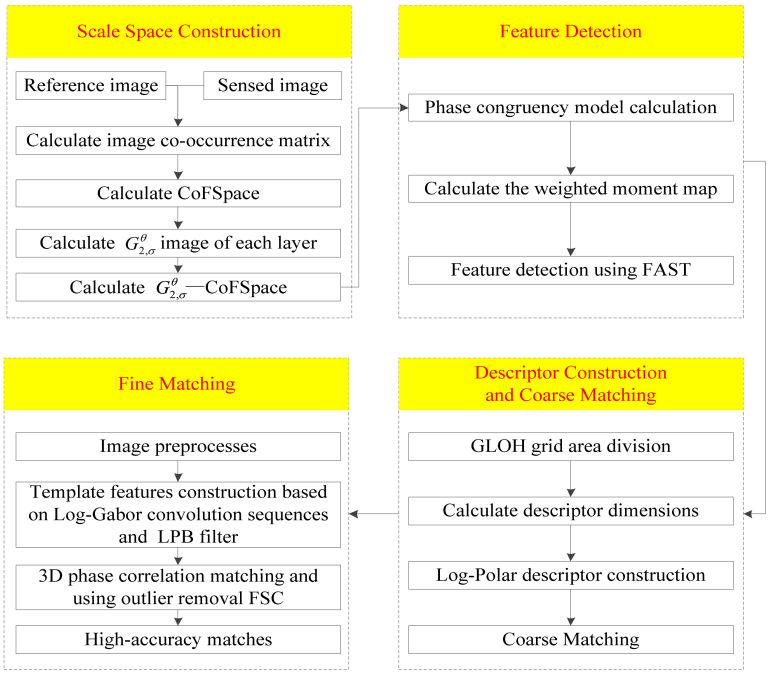

Figure 1. The framework of the matching process.

3.1. Co-Occurrence Scale Space Construction Combined with Second-Order Gaussian Steerable Filter

The scale of the image refers to the thickness of the image content. The image is constructed via convolution with continuous Gaussian kernels [43] to construct scale space by extracting the contour information of the image. Noise suppression and edge preservation are two critical metrics for scale space construction. To effectively preserve image edge information and to remove image noise, we propose a co-occurrence scale space combined with a second-order Gaussian steerable filter. The co-occurrence filter is an edge filter that preserves the texture boundary information of the image. The second-order Gaussian steerable filter is a linear steerable filter used to preserve the local spatial information of the image and to smooth the image to improve the anti-interference ability to noise.

When we construct the co-occurrence scale space combined with the second-order Gaussian steerable filter, we first calculate the co-occurrence scale space and obtain scale images. Then, we use the second-order Gaussian steerable filter to perform continuous convolution operation on the original image to obtain a smoothed Gaussian image. Finally, the two images are subtracted to obtain the final co-occurrence scale space, combined with the second-order Gaussian steerable filter. Since we use the co-occurrence filter to preserve the image's edge information and remove the local information using the second-order Gaussian steerable filter, our scale space keeps only the details and edge parts, which are crucial for multimodal image matching.

3.1.1. Co-Occurrence Scale Space Construction

The definition of CoF is shown in Equation (1):

$$J_p = \frac{\sum_{q \in N(p)} G_{\sigma_s}(p,q) \cdot M(I_p, I_q) \cdot I_q}{\sum_{q \in N(p)} G_{\sigma_s}(p,q) \cdot M(I_p, I_q)} \tag{1}$$

where J_p and I_q are the pixel values of the image output and input, p, q are the index positions of the pixel in the image, $G_{\sigma_s}(p,q) \cdot M(I_p, I_q)$ is the weight of the co-occurrence of pixel q to the output of pixel p, $G_{\sigma_s}(p,q)$ is the Gaussian filter with scale σ_s, $M(I_p, I_q)$ represents the calculation result of the normalized co-occurrence matrix, and M is a 256×256 matrix. The equation is shown in Equation (2):

$$\begin{cases} M(a,b) = \frac{C(a,b)}{h(a)h(b)} \\ C(a,b) = \sum\limits_{p,q} \exp\left(-\frac{d(a,b)^2}{2{\sigma_o}^2}\right)[I_p = a][I_q = b] \\ h(a) = \sum\limits_{p}[I_p = a], h(b) = \sum\limits_{q}[I_q = b] \end{cases} \tag{2}$$

In Equation (2), the calculation of $M(a,b)$ relies on the co-occurrence matrix $C(a,b)$, the number of co-occurrences of the calculated values a and b divided by their frequencies $h(a)$, $h(b)$. σ_0 is a fixed value in the calculation of this paper; $\sigma_0 = 2\sqrt{5} + 1$; $[\cdot]$ means 1 if the expression in parentheses is true and 0 otherwise.

The scale image of each layer in the co-occurrence scale space can be obtained using Equations (1) and (2). Divide the scale space into S layers, then the scale of each layer scale image is defined as follows: $\sigma_{s_n} = \sigma_{s_0} \cdot \sqrt[3]{2^n}$, $(n = 0, 1, 2 \ldots n)$; σ_{s_0} represents the scale of the first layer image in scale space, and S represents the scale-space layers of the multimodal image. Because we need to define the size of each local window when using co-occurrence filtering to process images, the local window size of each layer scale image

can be calculated from the window of the first layer scale image and the scale of each layer image. The definition is shown in Equation (3):

$$\begin{cases} OC_n = \frac{{\sigma_{s_n}}^2 \cdot NO}{2} \\ CoFSpace = \left\{ OC_n \cdot J_p^n \right\}_{n=0}^{N} \end{cases} \tag{3}$$

where $CoFSpace$ represents the co-occurring scale space; OC_n represents the local window size of the first-level scale image, NO represents the initial window size of the co-occurrence filter (set as 5 in this paper), N represents the number of layers of the scale image in the scale space, and J_p^n represents the nth multimodal image after co-occurrence filtering.

3.1.2. Co-Occurrence Scale Space Combined with Second-Order Steerable Filter

The steerable filter can realize the adaptive control of the filter by adjusting different angles, and has the characteristics of linearity, multi-direction, and multi-scale, thus providing more details in the image information of direction and edge [44]. Compared with the first-order gradient, the second-order gradient can better describe the local information of the image. The second-order Gaussian steerable filter is defined as follows:

$$\begin{cases} G_{2,\sigma}^{0^\circ} = G_{xx} = \left(-\frac{1}{2\pi\sigma^4} \right) \left(1 - \frac{x^2}{\sigma^2} \right) e^{\frac{-(x^2+y^2)}{2\sigma^2}} \\ G_{2,\sigma}^{90^\circ} = G_{yy} = \left(-\frac{1}{2\pi\sigma^4} \right) \left(1 - \frac{y^2}{\sigma^2} \right) e^{\frac{-(x^2+y^2)}{2\sigma^2}} \\ G_{xy} = \frac{xy}{2\pi\sigma^6} e^{\frac{-(x^2+y^2)}{2\sigma^2}}, G_{2,\sigma}^{60^\circ} = G_{yy} - G_{xy}, G_{2,\sigma}^{120^\circ} = G_{yy} + G_{xy} \\ G_{2,\sigma}^{\theta} = cos^2(\theta) G_{2,\sigma}^{0^\circ} + sin^2(\theta) G_{2,\sigma}^{60^\circ} - 2cos(\theta) sin(\theta) G_{2,\sigma}^{120^\circ} \end{cases} \tag{4}$$

where $G_{2,\sigma}^{0^\circ}$, $G_{2,\sigma}^{60^\circ}$ and $G_{2,\sigma}^{120^\circ}$ are used as the basic filter of $G_{2,\sigma}^{\theta}$ filter, and the $G_{2,\sigma}^{\theta}$ filter in all directions can be composed of these three filters. σ represents the scale of the second-order Gaussian steerable filter. To better collect the local spatial information of the image, we sum the convolutions in six directions $\left(0, \frac{\pi}{6}, \frac{2\pi}{6}, \frac{3\pi}{6}, \frac{4\pi}{6}, \frac{5\pi}{6}\right)$ of the image.

The co-occurrence scale space combined with the second-order Gaussian steerable filter is constructed using a second-order Gaussian steerable filter and co-occurrence scale images. The co-occurrence scale space definition combined with the second-order Gaussian steerable filter is defined in Equation (5):

$$\begin{cases} G_{2,\sigma}^{\theta} - CoFSpace = \left\{ 6 \cdot OC_n \cdot J_p^n - \sum_{\theta} G_{2,\sigma_{G_n}}^{\theta} * I^n \right\}_{n=0}^{N} \\ I^n = G_{2,\sigma_{G_{n-1}}}^{\theta} * I^{n-1}, n = (1, 2, \ldots, N) \end{cases} \tag{5}$$

In Equation (5), $G_{2,\sigma}^{\theta} - CoFSpace$ represents the image set of co-occurrence scale space combined with the second-order Gaussian steerable filter, $OC_n \cdot J_p^n$ represents the co-occurrence scale image of the n-th layer, θ is the convolution direction of the second-order Gaussian steerable filter, $*$ represents the convolution operation, and σ_{G_n} represents the scale of the second-order Gaussian steerable filter of each layer of images. In this paper, $\sigma_{G_n} = \sigma_{s_n}$; I^n represents the image that each layer participates in the second-order Gaussian steerable filter convolution. When $n = 1$, I^0 is the original input image.

In theory, since we use $G_{2,\sigma}^{\theta} - CoF$ to reduce the NRD between images, the processed image to be registered should be closer to the reference image than the unprocessed image to be registered. That is to say, they have a better similarity. To test our conjecture, we selected 18 pairs of images from six types of multimodal image pairs: optical–infrared, optical–depth, optical–map, optical–SAR, day–night, and optical–optical to test. Each image pair is pre-registered. Figure 2 shows a pair of original images (Figure 2a), $G_{2,\sigma}^{\theta} - CoF$ filtered images (Figure 2b,c), and the normalized mutual information (NMI) [45] and structural similarity (SSIM) [46] scores of the selected 18 pairs of images. As shown in Figure 2d,e, the

images after $G^{\theta}_{2,\sigma} - CoF$ filtering and CoF filtering have better NMI and SSIM scores than the original images. That is to say, after filtering, the resulting images will become more similar. From Figure 2, we can find that although the image using $G^{\theta}_{2,\sigma} - CoF$ filtering has a slightly lower score on NMI than the image using CoF filtering. The score on SSIM is much higher than that using the CoF filtering image, which also shows that our method can better preserve the structural features of the image. At the same time, combining Figure 2d,e, we found that compared with images in other modes, NMI and SSIM in the optical–map mode (7–9) have the best scores, because the image structure information in the optical–map mode is more suited for highlighting. With the day–night mode (13–15), although the original images have the worst NMI and SSIM scores, they also significantly improved after filtering with $G^{\theta}_{2,\sigma} - CoF$. Therefore, it is necessary to construct the scale space using $G^{\theta}_{2,\sigma} - CoF$.

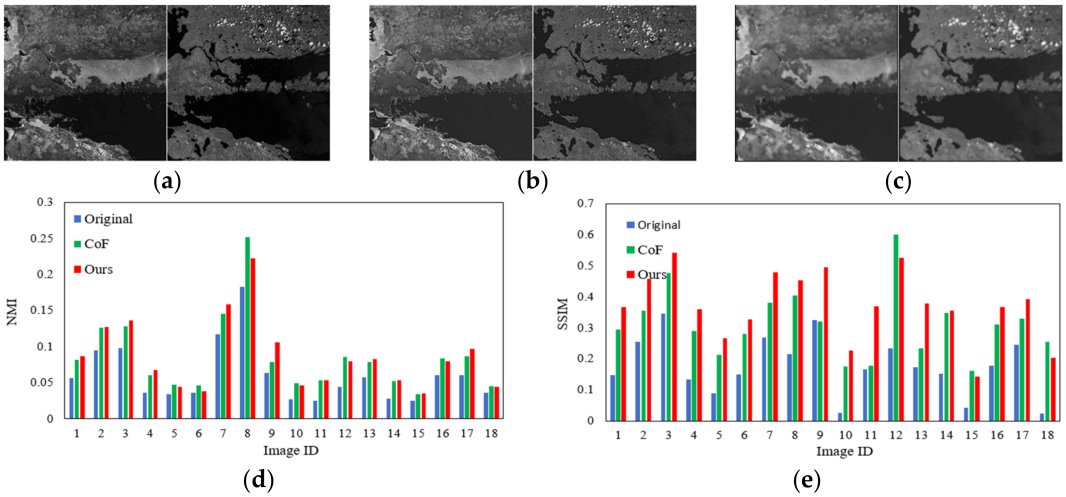

Figure 2. The effect diagram of $G^{\theta}_{2,\sigma} - CoF$ filtering, and the comparison results of $G^{\theta}_{2,\sigma} - CoF$ filtering and CoF filtering: (**a**) Original image pair; (**b**) The image after $G^{\theta}_{2,\sigma} - CoF$ filtering; (**c**) Image after $G^{\theta}_{2,\sigma} - CoF$ filtering twice; (**d**) NMI scores for $G^{\theta}_{2,\sigma} - CoF$ filtering and CoF filtering; (**e**) SSIM scores for $G^{\theta}_{2,\sigma} - CoF$ filtering and CoF filtering.

3.2. Feature Point Extraction Based on Phase Congruency

In this part, we use the phase information of images to detect edge and corner features [18]. The calculation equation of phase congruency is as follows:

$$PC(x,y) = \frac{\sum_o \sum_n W_o(x,y) \lfloor A_{no}(x,y)\Delta\Phi_{no}(x,y) - T \rfloor}{\sum_o \sum_n A_{no}(x,y) + \varepsilon} \tag{6}$$

where $PC(x,y)$ is the magnitude of phase congruency, $(x,\ y)$ is the pixel index of any point on the image, $A_{no}(x,y)$ is the amplitude of the image at $(x,\ y)$ after wavelet transformation with scale n and direction o, and $\Delta\Phi_{no}(x,y)$ is a more sensitive phase deviation. $W_o(x,y)$ is the weighting factor used to measure the two-dimensional frequency spread, T is a noise threshold, and ε is a small constant to avoid zero denominators in the calculations. $\lfloor\ \rfloor$ denotes that the enclosed quantity is equal to itself when its value is positive, and zero otherwise.

It can be known from the moment analysis method that the minimum moment diagram can display the direction information of the image features; the maximum moment diagram can reflect the saliency of the image features. We construct a weighted moment map, taking local anisotropy features into account according to the maximum moment and the minimum moment, and the calculation results are as follows:

$$M_{max} = \frac{1}{2}\left(c + a + \sqrt{b^2 + (a-c)^2}\right) \tag{7}$$

$$M_{min} = \frac{1}{2}\left(c + a - \sqrt{b^2 + (a-c)^2}\right) \tag{8}$$

$$M_{pc} = \frac{1}{2}(M_{max} + M_{min}) + \frac{k}{2}(M_{max} - M_{min}) \tag{9}$$

where

$$a = \sum_{\theta}(pc(\theta)cos(\theta))^2 \tag{10}$$

$$b = 2\sum_{\theta}(pc(\theta)cos(\theta))(pc(\theta)sin(\theta)) \tag{11}$$

$$c = \sum_{\theta}(pc(\theta)sin(\theta))^2 \tag{12}$$

where M_{max} represents the edge map of the image, M_{min} represents the corner map, M_{pc} is the final weighted moment graph, and k is the weight coefficient, which is used to measure the value ratio of the maximum moment graph and the minimum moment graph, and the value of k is between -1 and 1; a, b, c are three intermediate quantities. Moreover, we use a FAST detector to detect the interest points.

3.3. Improved Log-Polar Descriptor

3.3.1. Improved Gradient Feature and Feature Direction

After the feature points in the image are detected, the feature points need to be described to construct a feature vector for each feature point. We use the second-order gradient information of the $G^{\theta}_{2,\sigma} - CoFSpace$ scale images to generate new gradient features to weaken the NRD of images further. Since $G^{\theta}_{2,\sigma} - CoFSpace$ preserves the edge and detailed features of the images, so the constructed gradient can better describe the image features, thereby increasing the number of corresponding feature points and improving the final matching effect. The Sobel operator calculates the gradient of the image. Correspondingly, the horizontal and vertical gradient operators are as follows:

$$\Gamma_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 0 \end{bmatrix}, \Gamma_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{13}$$

In Equation (13), Γ_x represents the Sobel operator template in the X direction, and Γ_y represents the Sobel operator template in the Y direction.

Using Γ_x and Γ_y to calculate new first-order gradient magnitude and second-order gradient magnitude, defined as shown in Equations (14) and (15):

$$\begin{cases} G^1 = \sqrt{(I(x,y) * \Gamma_x)^2 + (I(x,y) * \Gamma_y)^2} \\ Angle^1 = arctan\left(\frac{I(x,y)*\Gamma_y}{I(x,y)*\Gamma_x}\right) \end{cases} \tag{14}$$

$$\begin{cases} G^2 = \sqrt{(G^1 * \Gamma_x)^2 + (G^1 * \Gamma_y)^2} \\ Angle^2 = \arctan\left(\frac{G^1*\Gamma_y}{G^1*\Gamma_x}\right) \end{cases} \tag{15}$$

where $I(x,y)$ represents the input scale image, G^1 represents the new first-order gradient magnitude, $*$ represents the convolution operation, $Angle^1$ represents the new first-order gradient direction, G^2 represents the new second-order gradient amplitude, $Angle^2$ represents the new second-order gradient direction, Γ_x represents the Sobel operator template in the X direction, and Γ_y represents the Sobel operator template in the Y direction.

In our study, new gradient images are generated using co-occurrence-scale images combined with the second-order Gaussian steerable filter, and more structured gradient features are obtained. To highlight the new gradient effect, we visualized the gradient amplitude and gradient direction, and the results are shown in Figures 3 and 4.

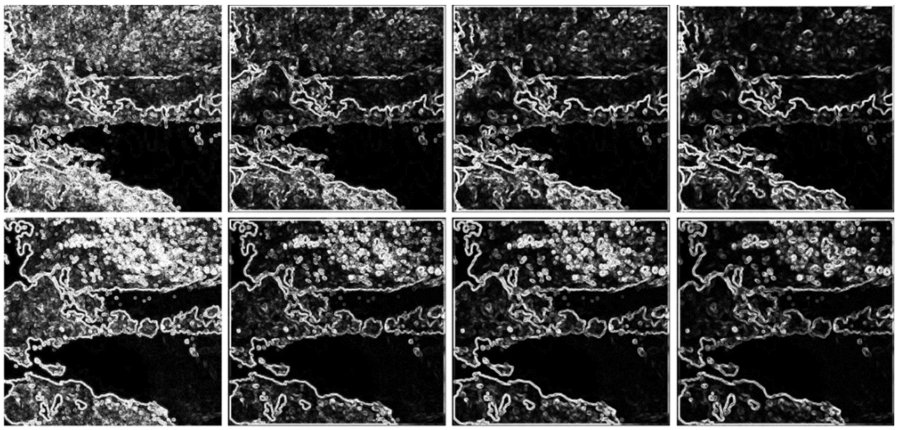

Figure 3. Results of new gradient magnitude: (**a**) The gradient amplitude of the original image pair; (**b**) The gradient amplitude of the image pair in first layer scale space; (**c**) The gradient amplitude of the image pair in second layer scale space; (**d**) The gradient amplitude of the image pair in third layer scale space.

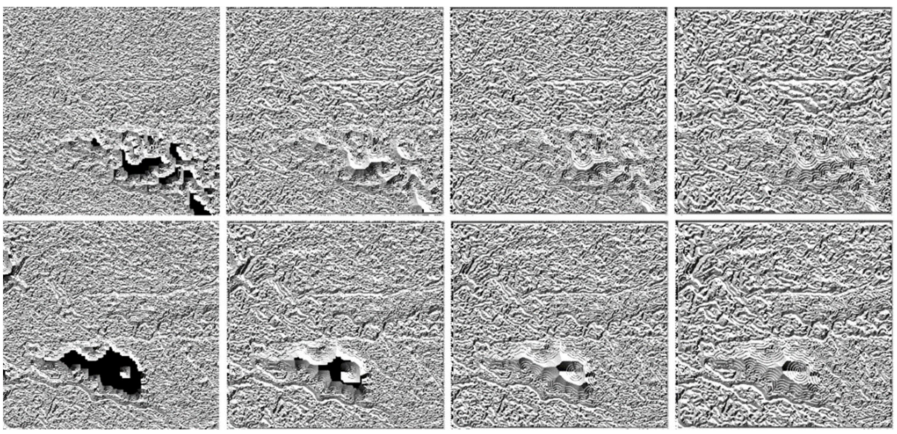

Figure 4. Results of new gradient direction: (**a**) The gradient direction of the original image pair; (**b**) The gradient direction of the image pair in first layer scale space; (**c**) The gradient direction of the image pair in second layer scale space; (**d**) The gradient direction of the image pair in third layer scale space.

3.3.2. Improved Log-Polar Descriptor

For the feature point set that has been extracted from the image pair, it is necessary to construct the feature descriptor of each feature point to increase the discrimination between these features. A lot of research has been performed on how to construct feature descriptors, including SIFT feature descriptors, HOG feature descriptors, and GLOH feature descriptors. The GLOH descriptor is an extension of the SIFT descriptor, which uses a logarithmic grid to divide local image blocks instead of regular grid division, improving the descriptor's robustness and saliency. In this paper, to address the geometric distortion problem in multimodal image matching, a feature descriptor similar to GLOH is designed.

For each local image block, we specify that it is a local feature block of $R \times R$ pixel size, selected from the feature point as the center. When we calculate the main direction of each local image block, we divide it into 24 bins according to the directions of all pixels in the local image block, calculate the sum of the gradient magnitudes corresponding to each bin, and select the highest score bin as the main direction. We not only use the bin with the highest score as the main direction, but also record the bin with a score greater than 0.8 times the highest bin as the auxiliary direction. This is because Lowe mentioned in

the original paper that although only about 15% of the feature points will have multiple orientations, and it can effectively improve the robustness of matching [11]. The direction of the improved log-polar descriptor is:

$$LPA = [A_1, A_2, \cdots, A_N]^T \tag{16}$$

where LPA represents the direction set of all feature points, ${A_i}^T$ represents the direction set of each feature point, including the main direction and auxiliary directions; T represents the matrix transpose character, N represents the number of feature points, ${A_i}^T = [Angle_{main}, Angle_{assist1}, Angle_{assist2}, \cdots, Angle_{assistn}]$; $Angle_{main}$ is the main direction of the i-th feature point, and $[Angle_{assist1}, Angle_{assist2}, \cdots, Angle_{assistn}]$ is all auxiliary directions of the feature point.

A feature descriptor is constructed for each feature point's direction. Generally speaking, for the GLOH descriptors, the entire local image block is first rotated to the main direction. Then, the whole description region is divided into a circular region and two annular regions from the inside to the outside. The size of the three regions can be determined by R_1, R_2, and R_3 (seen as in Figure 5). Each annular region is equally divided into eight fan-shaped regions, and all directions in the region are divided into 16 column-shaped parts. The inner circle area has only divided all directions into 16 directions. Finally, a $(2 \times 8 + 1) \times 16 = 272$ dimensional feature vector can be obtained. However, the feature vectors constructed in this manner have complexities that are too high, which makes the time and space complexity of the algorithm large. Therefore, considering the stability and time complexity of the descriptor, we divide each ring into 12 fan-shaped regions equally, and divide the directions within the fan-shaped region into eight equal divisions. The directions' division in the inner circle area is consistent with each sector area. Finally, as shown in Figure 5, we generated a $(2 \times 12 + 1) \times 8 = 200$-dimensional feature vector. Such a division not only improves the instability of the description caused by fewer grids, but also reduces the time complexity of the descriptor.

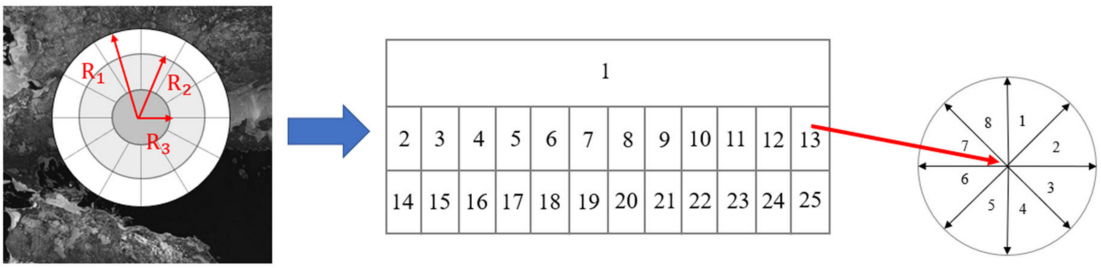

Figure 5. A log-polar descriptor for multimodal image grid optimization.

3.4. Extended 3D Phase Correlation Similarity Metrics

After the feature descriptors are constructed, we perform initial matching and obtain preliminary matching results. Due to various problems in multimodal image matching, such as rotation, translation, scaling, and viewing angle differences, it is difficult to obtain uniformly distributed and numerous matching points in all scenes. However, as we all know, if the matching points are distributed more evenly and the number of matching points is greater, the matching accuracy is also higher. Therefore, after completing the preliminary matching, we use the obtained transformation matrix H1 to project the image to be registered onto the reference image, and perform secondary matching to obtain a more reliable spatial transformation relationship.

Phase correlation (PC) is an efficient frequency-domain matching method that has been widely used in images due to its sub-pixel accuracy and robustness to image contrast, noise, and occlusion [30]. Since we eliminated the rotation and scaling differences between the two image patches in the preliminary matching, there is mainly a translational offset between them. Therefore, we adopt a template matching scheme for image matching. The

matching process based on 3D phase correlation mainly includes the following steps: (1) image preprocessing, (2) optimized 3DPC [47] descriptor construction, and (3) template matching.

3.4.1. Image Preprocessing

After completing the initial matching, we obtain the preliminary transformation matrix H1, and we use H1 to project all the pixels (x, y) on the to-be-registered image $I_{Sen\_Before}(x, y)$ to the reference image $I_{Ref}(x, y)$. Above, we generate a new image to be registered $I_{Sen\_After}(x, y)$. The calculation equation is shown in Equation (17):

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H1 \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{17}$$

where (x, y) is the pixel coordinate point on the image to be registered $I_{Sen\_Before}(x, y)$, H1 is the homography transformation matrix, and (x', y') is the pixel coordinate of the point after transformation.

Due to the problems of rotation, translation, and scaling of the two images, the transformed pixel coordinates (x', y') may exceed the boundary range of the reference image. Thus, the final generated $I_{Sen\_After}(x, y)$ can be represented using the following code:

$$\textbf{If } \left(x' > 0 \;\&\&\; y' > 0 \;\&\&\; x' <= I_{Ref\_width} \;\&\&\; y' <= I_{Ref\_height}\right), \textbf{ Then } \text{delete } \left(x', y'\right) \tag{18}$$

$$I_{Sen\_After}(x, y) = \text{padarray}\left(I_{Sen_{After}}(x, y), [\text{marg, marg}], 0, '\text{both}'\right) \tag{19}$$

$$I_{Ref}(x, y) = padarray\left(I_{Ref}(x, y), [marg, marg], 0, 'both'\right) \tag{20}$$

where I_{Ref_width} is the length of the reference image, I_{Ref_height} is the width of the reference image, padarray represents the filling operation on the image, $[marg, marg]$ represents the number of rows and columns to fill, marg represents the search boundary for template matching, 0 represents the padding value, and $'both'$ means padding before the first element and after the last element of each dimension of the image. The operation steps are shown in Figure 6.

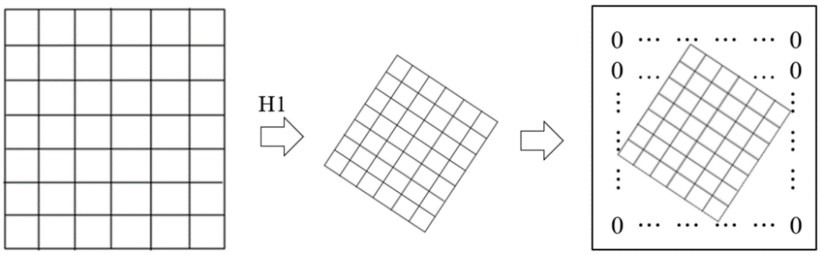

Figure 6. Schematic diagram of image preprocessing: H1 is the transformation matrix calculated via preliminary matching.

3.4.2. Optimized 3DPC Descriptor Construction Image Preprocessing

For the preprocessed image to be registered $I_{Sen\_After}(x, y)$ and the reference image $I_{Ref}(x, y)$, we use the co-occurrence filter and the Low-Pass Butterworth (LPB) filter further to enhance the edge feature of the two images and to reduce NRD. The LPB filter can reduce the energy of the high-frequency part of the image to achieve the effect of smoothing the image and reducing noise [48,49]. We do not use the second-order Gaussian steerable filter mentioned above for image processing, because when we construct the optimized 3DPC descriptor, we need to use the Log-Gabor filter to calculate the amplitude map in the six directions of the image and arrange them on the *Z*-axis. The second-order Gaussian steerable filter collects convolutional information in multiple directions of the image to improve the local features of the image. However, this may lead to a decrease in the

difference of amplitude maps in different directions. Therefore, we choose LPB filter to reduce the NRD between images.

After obtaining the filtering $I_{Sen\_After}(x,y)$ and $I_{Ref}(x,y)$, we do not directly calculate its phase information to generate 3DPC descriptors. To better retain the structure of the image, the second-order gradient amplitude of the image is generated according to the steps in Section 3.3 as the final operation image. Similar to other methods of using templates, we stipulate that the size of each template is $R_m \times R_m$. Then, we use the even-symmetric and odd-symmetric log-Gabor wavelets at scale n and orientation $\bar{o}$ to convolute the image to obtain the corresponding response component $e_{n\bar{o}}(x,y)$ and $o_{n\bar{o}}(x,y)$. The calculation equation is shown in Equation (21):

$$[e_{no}(x,y), o_{no}(x,y)] = [I(x,y) * M^e_{no}, I(x,y) * M^o_{no}] \tag{21}$$

where M^e_{no} and M^o_{no} represent the even-symmetric and odd-symmetric log-Gabor wavelets at scale n and orientation $\bar{o}$, $I(x,y)$ represents an input image, and $e_{no}(x,y)$ and $o_{n\bar{o}}(x,y)$ represent the response components obtained with two small wave functions through convolution.

So, the amplitude after the wavelet transforms at scale n and orientation $\bar{o}$ is:

$$A_{n\bar{o}}(x,y) = \sqrt{e_{n\bar{o}}(x,y)^2 + o_{n\bar{o}}(x,y)^2} \tag{22}$$

For each direction $\bar{o}$, find the amplitude value of all n scales and obtain the direction amplitude diagram $A_{\bar{o}}(x,y)$. Arrange the direction amplitude maps on the Z-axis according to the size of the direction to obtain an optimized 3DPC descriptor. Figure 7 shows the processing process of the optimized 3DPC descriptor.

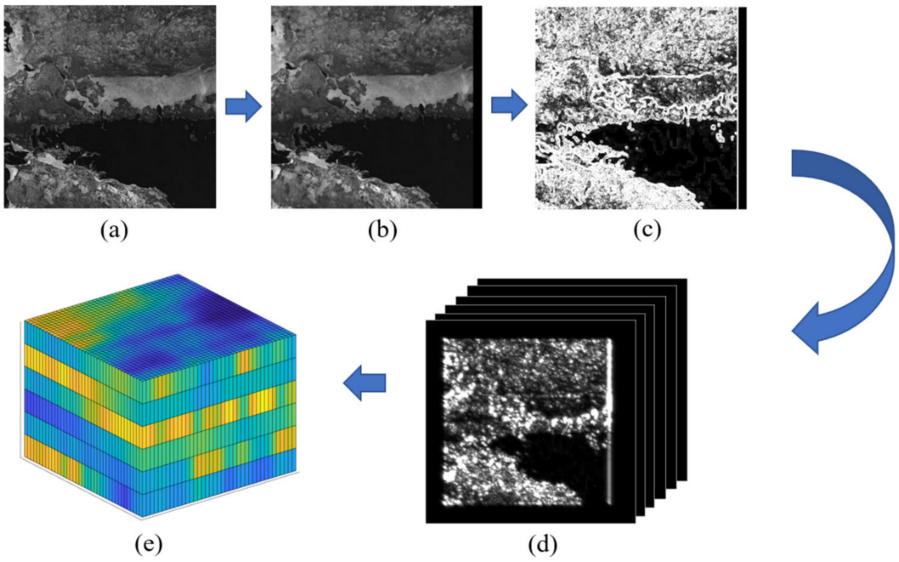

Figure 7. Optimized 3DPC description process: (**a**) The original image; (**b**) Preprocessed image; (**c**) Second-order gradient amplitude of the image; (**d**) Phase direction magnitude map of the image; (**e**) Optimized 3DPC descriptor.

3.4.3. Template Matching

The basic theory of image matching models based on phase correlation is the Fourier shift property, which states that the relative displacement of a pair of similar images in the spatial domain can be transformed into a linear phase difference in the frequency domain [50]. Let $I_1(x,y)$ and $I_2(x,y)$ be a pair of corresponding feature points in the image

to be registered and the reference image, respectively, then $I_1(x,y)$ and $I_2(x,y)$ should satisfy the following relationship:

$$I_1(x,y) = I_2(x - x_0, y - y_0) \tag{23}$$

$$f_1(u,v) = f_2(u,v)\exp\{-2\pi i(ux_0, vy_0)\} \tag{24}$$

where (x_0, y_0) is the displacement difference between points $I_1(x,y)$ and (x,y), $f_1(u,v)$ is the coordinate of $I_1(x,y)$ in frequency after the Fourier transform (FT) of the image, and $f_2(u,v)$ is the coordinates of $I_2(x,y)$ in frequency after the reference image undergoes FT.

Because we have expanded the original image and added its description on the Z axis, Equations (23) and (24) can be written as follows:

$$I_1(x,y,z) = I_2(x - x_0, y - y_0, z) \tag{25}$$

$$f_1(u,v,w) = f_2(u,v,w)\exp\left\{-2\pi i(ux_0, vy_0)\overleftarrow{\gamma}\right\} \tag{26}$$

where z is the dimension of the reference image and the image to be registered on the Z axis; $\overleftarrow{\gamma}$ is a 3D unit vector. The normalized cross-power spectrum matrix is calculated as an image-matching similarity measure. Then, the 3D inverse Fast Fourier transform (IFFT) is performed to obtain the 3D Dirac function:

$$Q(u,v,w) = \frac{f_1(u,v,w)f_2(u,v,w)^*}{\left|f_1(u,v,w)f_2(u,v,w)^*\right|} = exp\left\{-2\pi i(ux_0, vy_0)\overleftarrow{\gamma}\right\} \tag{27}$$

$$Q(u,v,w) = \vartheta^{-1}\{Q(u,v,w)\} = \delta(u - x_0, v - y_0)\overleftarrow{\gamma} \tag{28}$$

where $*$ denotes a complex conjugate and ϑ^{-1} denotes IFFT; we can use the Dirac function to determine the offset between two images. Usually, the peak of the Dirac function will appear at (x_0, y_0). Thus, we can complete image matching work by searching for local maxima.

4. Experiment and Analysis

In this section, we compare our proposed G-CoFTM method with five state-of-the-art methods: PSO-SIFT, HAPCG, RIFT, LPSO, and COFSM. For a fair comparison, the codes of the comparative methods provided by the authors are applied. At the same time, phase information is also used in RIFT, HAPCG, and LPSO. Therefore, we set the parameters related to this paper in these three algorithms to be consistent, to ensure the fairness of the experiment.

4.1. Data Description

We use the datasets provided by Yao et al. [24] and Yang et al. [36] for testing. The datasets contain six types of multimodal image pairs (optical–optical, optical–infrared, optical–depth, optical–map, optical–SAR, and day–night). In the datasets Yao provided, each image type contains 10 image pairs. There are mainly translation relationships between these image pairs, and about 10 to 30 corresponding homonymous feature points were manually selected for each image pair by the provider. In the datasets provided by Yang, there are image pairs with rotation and scale changes. There are significant NRD and slight geometric distortion between the two images in image pairs. In order to better prove the processing effect of our proposed method for rotation and scale changes, the image pairs are randomly selected from the above six types of image pairs for rotation and scale transformation. Finally, for each method, we validated a total of 8×10 pairs of images (optical–optical, optical–infrared, optical–depth, optical–map, optical–SAR, day–night, rotation, and scale).

4.2. Evaluation Indices

In the quantitative evaluation, we select NCM, SR, and root mean square error (RMSE) as the evaluation index to measure the matching effect of each algorithm. Among them, SR represents the ratio of the correct matching in all successful matching, and NCM represents the number of homonymous feature points correctly matched. If the number of NCM is less than 5, it is considered to be a matching failure. RMSE represents the pixel error between the coordinates of the feature points after using the transformation matrix and the corresponding homonymous feature points. Suppose that there is a point coordinate (x, y) on the original image, and its position is (x', y') after matrix transformation. Its corresponding homonymous feature point is (x_2, y_2), so that RMSE can be calculated using Formula (29). RMSE reflects the accuracy of matching; the smaller the value of RMSE, the higher the matching accuracy. We use the high-precision correspondence manually selected by the provider to estimate the real transformation model H of the image pair, and use it as the actual value of the following evaluation process.

$$MSE = \sqrt{\frac{1}{NCM}\sum_{1}^{NCM}\left[(x' - x_2)^2 + (y' - y_2)^2\right]} \tag{29}$$

4.3. Parameter Study

Our proposed G-CoFTM mainly consists of three parameters, namely S, R, and R_m. S is the number of layers of the scale image in the scale space. The larger the value of S, the more information the scale space contains, and the computational complexity will also increase. Parameter R is used to describe the size of the local image block in preliminary matching. If the local image block is too small, it contains insufficient local information and cannot fully reflect the uniqueness of the feature. Conversely, if the local image block is too large, then the time required for calculation is higher; the longer the time that is required for matching. The parameter R_m, like the parameter R, is used to describe the size of the local image patch, and the parameter R_m is used to construct the search template in precise matching. The larger the parameter R, the more information the template contains and the higher the computational complexity. Therefore, the correct parameter design will have an important impact on the matching results and the time required for matching. We randomly selected a pair of optical–map images for the test. To obtain the appropriate parameters, we designed three independent experiments to learn the parameters S, R, and R_m. In each experiment, we only allow one parameter to exist as a variable, and the other parameters will be fixed as a constant. The overall setting details of the experiment are shown in Table 1. The results of the experiment are shown in Tables 2–4. The image pairs used in the experiment are shown in Figure 8.

Table 1. Details of parameter settings.

Experiments	Variable	Fixed Parameters
parameter S	$S = [1, 2, 3, 4, 5]$	$R = 48$, $R_m = 108$
parameter R	$R = [32, 40, 48, 56, 64]$	$S = 3$, $R_m = 108$
parameter R_m	$R_m = [48, 72, 96, 120, 144]$	$S = 3$, $R = 48$

Table 2. The results of the parameter S.

Metric	S, R = 48, R_m = 108				
	1	**2**	**3**	**4**	**5**
NCM	719	1540	1586	1068	1226
SR/%	100	100	100	100	100

Table 3. The results of the parameter R.

Metric	R, S = 3, R_m = 108				
	32	**40**	**48**	**56**	**64**
NCM	534	1071	1586	1960	1629
SR/%	97.8	100	100	100	100

Table 4. The results of the parameter R_m.

Metric	R_m, S = 3, R = 48				
	84	**96**	**108**	**120**	**132**
NCM	947	1059	1586	1457	1381
SR/%	100	100	100	100	100

Figure 8. Image pairs for parameter study experiments.

From the results of the above three experiments, we can draw the following conclusions: (1) The larger S is, the more layers of scale space that we construct for G-CoFTM, and the richer the information extracted from the image is. Additionally, the larger S is, the more NCM will be obtained. However, too large a value of S will cause unnecessary computational overhead. It can be seen from Table 2 that our method has achieved good results in the scale space of any number of layers. In order to balance the accuracy and calculation time of the matching results, we set S to 3. (2) It can be seen from Table 3 that the mismatch of the constructed G-CoFTM decreases with an increase in the value of R. When $R \geq 40$, the mismatch disappears, and the number of NCM increases. However, when $R = 64$, the number of NCM is reduced to 1629, which may be because too large a number of local image blocks contain some unnecessary noise while increasing the description information. Therefore, according to Table 3, we set $R = 48$. (3) In all assumed values of the parameter R_m, SR reaches the highest value. The obtained NCM begins to increase when $R_m \geq 40$, and it reaches the highest point when $R_m = 108$ and begins to decrease. The reason for this is similar to R_m, because raising the template size increases the amount of description information and includes noise. We choose $R_m = 108$ as the parameter of our experiment, which is not only because the number of correct matches is the highest when $R_m = 108$; it is also to reduce our computing time. In the following experiments, these parameters are fixed as $S = 3$, $R = 48$, $R_m = 108$.

4.4. Performance Evaluation

4.4.1. Qualitative Comparisons

We randomly selected an image pair from each of the eight types of multimodal datasets collected for qualitative comparison, and we visualized the matching results of the various methods. Due to the different temporal phases or the imaging mechanisms of these images, these images contain significant NRD. The matching results of G-CoFTM, CoFSM, RIFT, HAPCG, LPSO, and PSO-SIFT are shown in Figure 9.

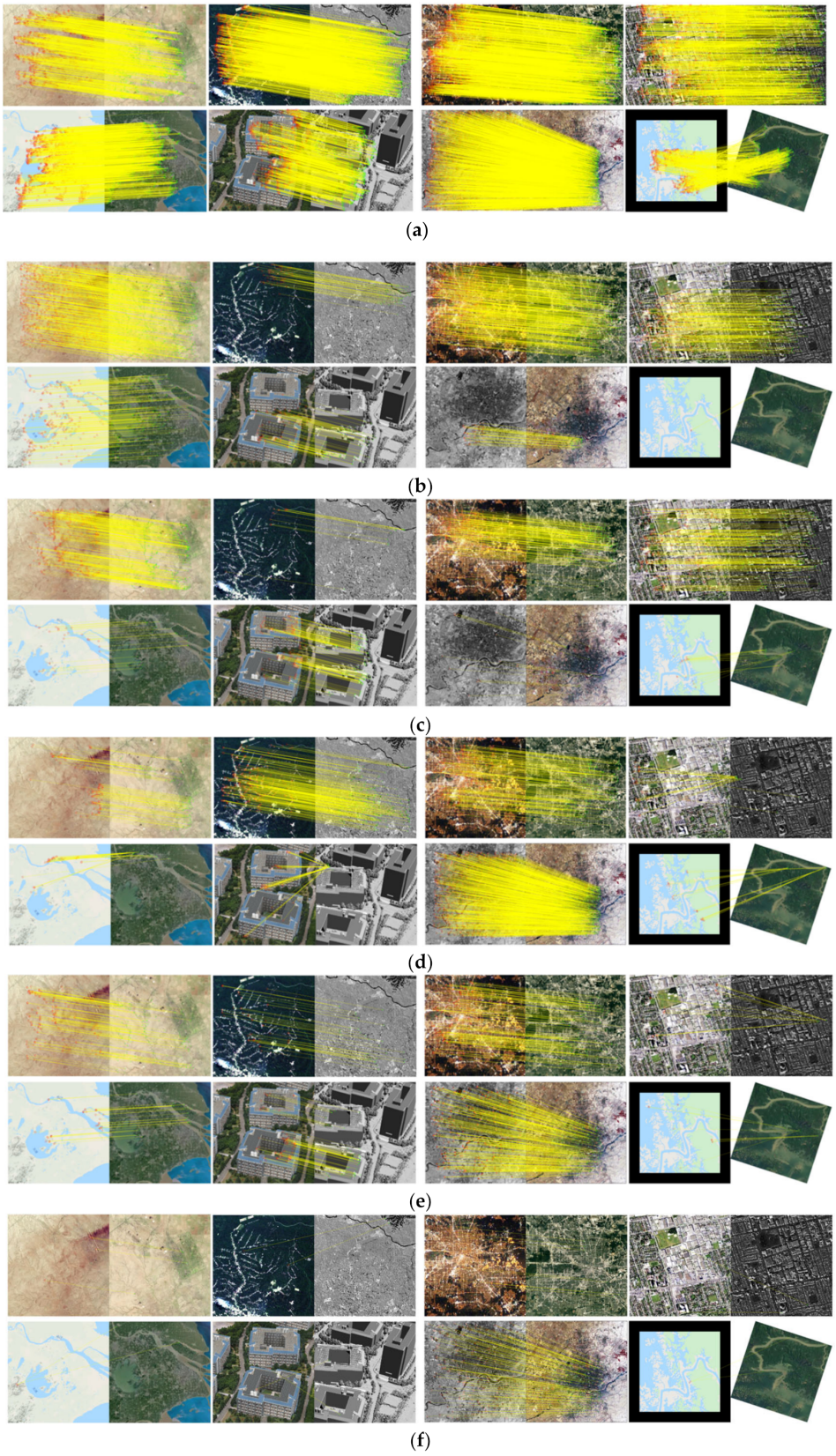

Figure 9. Qualitative comparison results of the sample data: (**a**) Results of G-CoFTM; (**b**) Results of CoFSM; (**c**) Results of RIFT; (**d**) Results of HAPCG; (**e**) Results of LPSO; (**f**) Results of PSO-SIFT.

It can be seen from Figure 9 that our proposed G-CoFTM has achieved good results on eight types of data, and the number of extracted homonymous feature points is the

largest. CoFSM, RIFT, HAPCG, LPSO, and PSO-SIFT involved in the comparison have various problems with eight types of data: they cannot extract the correct corresponding feature points on all experimental images, or the number of corresponding feature points extracted on certain types of images is sparse and unevenly distributed. Below, we will analyze the experimental results in detail. From Figure 9b, we can see that CoFSM has obtained many corresponding feature points with uniform distribution on optical–optical, day–night, optical–SAR, and optical–map. However, for optical–infrared, optical–depth, scale, and rotation images, the number of corresponding feature points extracted by CoFSM is scarce, unevenly distributed, or wrongly extracted. This is because CoFSM uses CoFSM filtering to extract the edge features of the image to increase the reliability of the matching. However, for each descriptor, if the image information contained is too little, the reliability of the matching will also decrease. For multimodal image matching, it is necessary to contain enough local information.

Compared with CoFSM, RIFT uses second-order phase gradient features instead of gradient features to calculate the main direction, and uses the maximum index map (MIM) to construct local descriptors. At the same time, RIFT also discusses the influence of the size of each descriptor on matching. This means that RIFT can extract the correct corresponding feature points on each data type, as shown in Figure 9c. However, due to the characteristics of the phase information itself, the scale changes between the images cannot be well processed. This is the reason for the poor extraction effect of RIFT on scale images. At the same time, since each descriptor in RIFT has only six description latitudes, it cannot achieve good results for images with insufficient texture features (such as optical–infrared).

HAPCG and LPSO are improved based on RIFT, and we can regard it as a RIFT upgrade. HAPCG uses phase congruency direction to solve the rotation problem in the images. However, due to the large NRD between multimodal images, it is sometimes inaccurate to use the phase consistency feature direction to calculate the main direction of the image. LPSO further proposes to use the odd symmetric filtering direction instead of the phase consistency direction on HAPCG, which makes its matching effect on multimodal images better than HAPCG. However, LPSO still cannot solve the rotation problem in multimodal image matching, as shown in Figure 9d,e.

PSO-SIFT uses the second-order gradient feature of the image to construct the descriptor and to calculate the main direction of the descriptor. Due to the existence of NRD, the gradient features of the images are very unreliable for multimodal image matching; PSO-SFT is the worst in all test methods, such as in Figure 9f.

4.4.2. Quantitative Comparisons

In order to better compare the methods mentioned above, we conducted a more detailed quantitative evaluation. Since PSO-SIFT only extracts a few wrong corresponding feature points on multiple data types, we do not analyze PSO-SIFT. At the same time, we set the RMSE to 10 when the image matching fails. For an image, if the matching time is more than 10 min and there is no response, it is also considered a failed matching. We set NCM to 0 and RMSE to 10 in this case.

Table 5 shows the SR of each method on different datasets. We can see that the SR of our method reaches 100 on all types of image pairs. HAPCG and LPSO have the worst SRs of all methods. The SRs of CoFSM on optical–optical, optical–infrared and optical–map data reach 100, and CoFSM also achieves good results on optical–SAR images, with SR reaching 90. However, the effects on optical–depth, day–night, scale, and rotation types of images are poor, with SRs of 70, 70, 70, and 60, respectively. RIFT is the best matching method among the five methods, except for G-CoFTM. Except for the scale test, the SR of RIFT reached 100. However, the SR of RIFT on the scale images is only 50. This is because RIFT does not construct scale space, and cannot deal with the scale change of the image well. Unfortunately, we find that the matching time of RIFT is the longest among the five methods. If the scale space is constructed to solve the scale problem, its matching time will

reach an unacceptable level (about 6–7 min). HAPCG and LPSO received the worst SRs in all comparison methods, but the matching effect of LPSO is slightly better than HAPCG.

Table 5. Comparison on SR metric.

Method	SR/%							
	Optical–Optical	Optical–Infrared	Optical–Depth	Optical–Map	Optical–SAR	Day–Night	Scale	Rotation
G-CoFTM	100	100	100	100	100	100	100	100
CoFSM	100	100	70	100	90	70	70	60
RIFT	100	100	100	100	100	100	50.5	100
HAPCG	100	100	79.6	59.9	100	70	36.8	10
LPSO	90	90	80	80	60	70	49.8	22.5

Figure 10 shows the NCM results, and Table 6 shows the average NCM of all methods on each data type. It can be seen from Figure 10 that the proposed G-CoFTM shows excellent performance on all types of data and obtains the most matching numbers. The NCM_{ave} obtained by G-CoFTM on all types of data are 1917.5, 3003.8, 1755.9, 2061, 1631.2, 1150.1, 1890.4, and 1479.3, respectively, as shown in Table 6. Compared with G-CoFTM, although the average extraction number of CoFSM ranked second among the five methods, the eight types of NCM_{ave} were 556, 647.4,223, 368.1, 172, 335.5, 177.5, and 318.1, respectively. However, from Figure 10, we can see that CoFSM has matching failures on optical–depth, optical–SAR, day–night, scale, and rotation images. Especially in the day–night and scale types of images, there are three instances where CoFSM cannot match. RIFT uses the phase index to construct the descriptor. Although the description latitude of the descriptor is reduced, the stability is greatly improved. As shown in Table 5, RIFT only has mismatches on the scale images. Table 6 and Figure 10 show the specifics of RIFT matching. We can see that RIFT has only 14.9 NCM_{ave} on the scale images, and failed matching occurs for almost every pair of scale images. Therefore, for RIFT, it is necessary to improve its adaptability to scale images while reducing the computational complexity. HAPCG is similar to RIFT. It can be found from Table 6 that the NCM of HAPCG on all types of data is almost improved compared with RIFT. Figure 10 more intuitively reflects the lifting effect of HAPCG. However, although HAPCG has been greatly improved on NCM compared with RIFT, HAPCG is significantly lower than RIFT in the success rate of matching. RIFT has better stability than HAPCG, as shown in Table 5. LPSO and HAPCG are similar, although they are slightly lower than HAPCG in NCM, but there is a certain degree of stability improvement; see Tables 5 and 6.

Table 6. Comparison on the NCM_{ave} metric.

Method	NCM_{ave}							
	Optical–Optical	Optical–Infrared	Optical–Depth	Optical–Map	Optical–SAR	Day–Night	Scale	Rotation
G-CoFTM	1917.5	3003.8	1755.9	2061	1631.2	1150.1	1890.4	1479.3
CoFSM	556	647.4	223	368.1	172	335.5	177.5	318.1
RIFT	267.7	324.6	172.2	94.9	151.3	79.7	14.9	189.1
HAPCG	317.9	468.3	195.7	242.3	178.2	153.8	116.1	57.6
LPSO	113.16	189.1	74.6	171.7	68.6	74.7	155.1	59.3

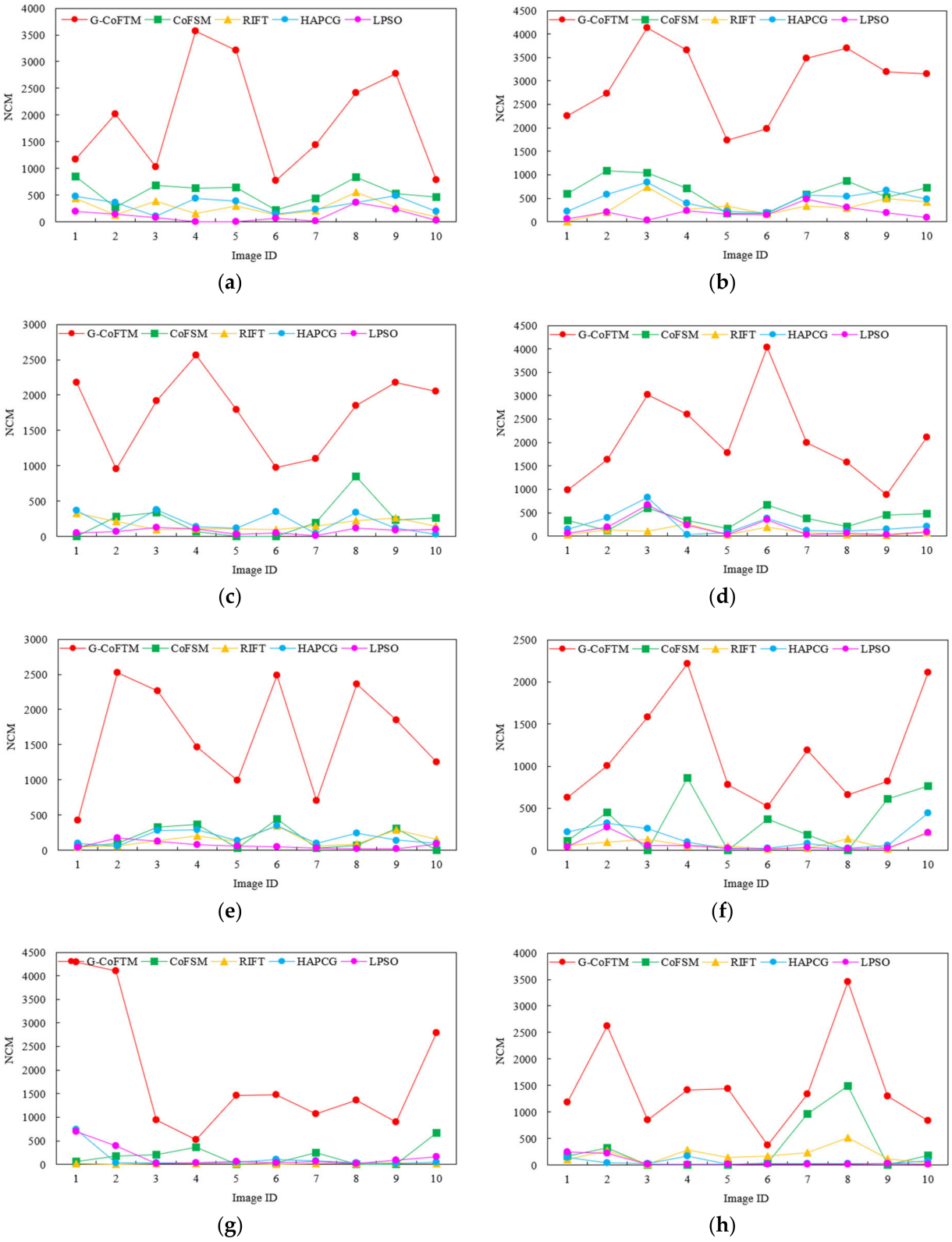

Figure 10. Comparisons on the NCM metrics: (**a**) Optical–Optical; (**b**) Optical–Infrared; (**c**) Optical–Depth; (**d**) Optical–Map; (**e**) Optical–SAR; (**f**) Day–Night; (**g**) Scale; (**h**) Rotation.

Considering the accuracy of the matching, Figure 11 shows the RMSE of all methods on different types of data, where RMSE = 10 indicates a failed match. The results show that our matching accuracy is the highest among all methods due to our two-step matching

strategy using optimized 3DPC descriptors for accurate matching. It can be seen from Table 7 that the RMSE of our matching is about 1 pixel, and that the $RMSE_{ave}$ of eight times is 1.004, 0.890, 1.129, 0.990, 1.093, 1.169, 1.017, and 1.074, respectively. CoFSM has a high matching accuracy on the optical–optical, optical–infrared and optical–map types of images. The $RMSE_{ave}$ of the three types of images are 1.900, 1.853, and 1.801, respectively. CoFSM has poor matching accuracy on four types of images: optical–depth, day–night, scale, and rotation. Figure 10 shows that CoFSM often fails to match the above four types of images. CoFSM is quite unstable in the face of these four types of data. RIFT has stable matching performance on seven types of images except for scale, with an $RMSE_{ave}$ of around 2 pixels. The $RMSE_{ave}$ of HAPCG is lower than that of LPSO on four types of images: optical–optical, optical–infrared, optical–SAR, and day–night, which are improved by 0.802, 0.974, 1.79, and 0.727, respectively. However, the $RMSE_{ave}$ of HAPCG is higher than that of LPSO in optical–map, optical–depth, scale, and rotation types of images. Combined with Figure 11 and Table 7, the matching effect of HAPCG is generally better than that of LPSO. However, both methods cannot deal with various problems in multimodal image matching. Comparing the performances of all methods on eight data types, we can find that the scale and rotation types of images are more challenging for the matching algorithms. This is because the scale and rotation types of images are randomly selected and generated from the first six images, making it more challenging. Day–night images are also a challenge in multimodal image matching. We analyzed 10 day–night image pairs in the dataset, and found that the day–night images contain less information than the other five types of data (excluding scale and rotation), and they have a larger gap with the reference image.

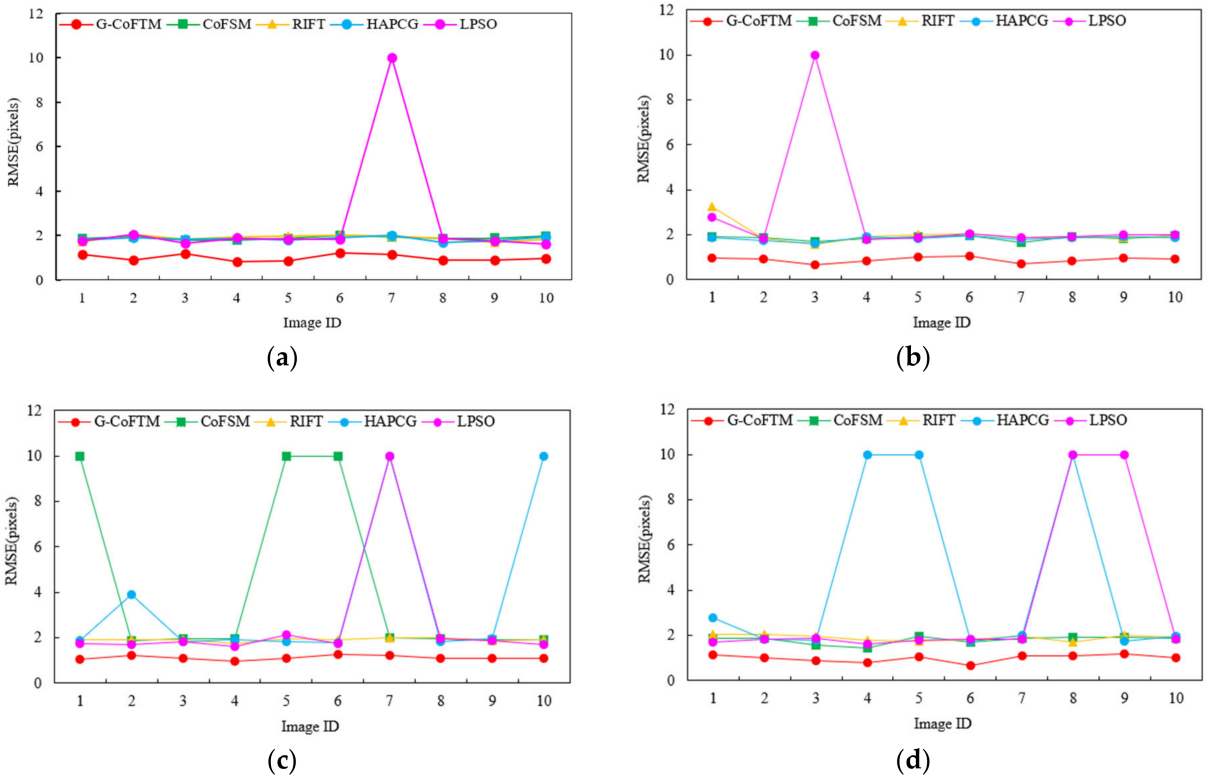

Figure 11. *Cont.*

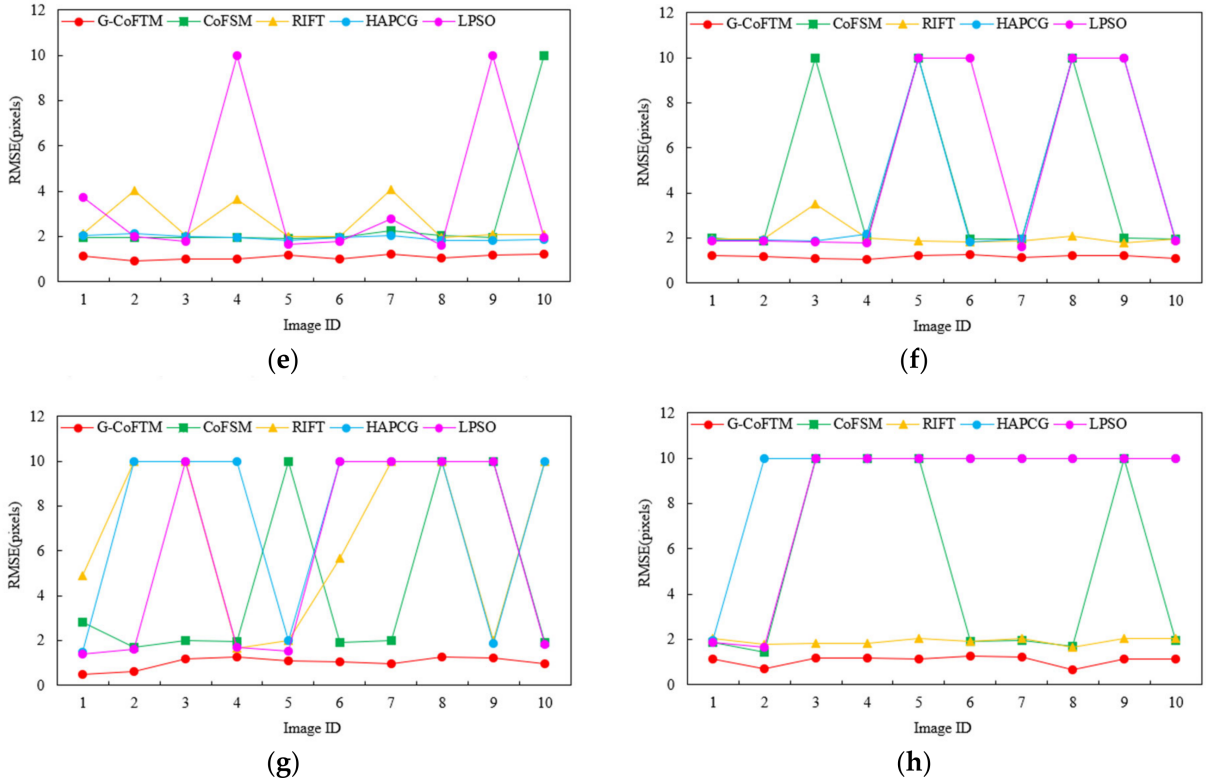

Figure 11. Comparisons on the RMSE metrics: (**a**) Optical–Optical; (**b**) Optical–Infrared; (**c**) Optical–Depth; (**d**) Optical–Map; (**e**) Optical–SAR; (**f**) Day–Night; (**g**) Scale; (**h**) Rotation.

Table 7. Comparison on $RMSE_{ave}$ metric.

Method	$RMSE_{ave}$							
	Optical–Optical	Optical–Infrared	Optical–Depth	Optical–Map	Optical–SAR	Day–Night	Scale	Rotation
G-CoFTM	1.004	0.890	1.129	0.990	1.093	1.169	1.017	1.074
CoFSM	1.900	1.853	4.358	1.801	2.800	4.367	4.426	5.087
RIFT	1.891	2.027	1.921	1.903	2.599	2.084	6.621	1.926
HAPCG	1.853	1.836	3.696	4.380	1.945	4.357	7.532	9.193
LPSO	2.632	2.810	2.630	3.431	3.735	5.084	5.803	8.353

5. Discussion

5.1. Performance Analysis

From the qualitative and quantitative results, we can see that the proposed G-CoFSM achieves good results on all types of data. This is because we decompose the matching problem into two sub-problems: we obtain relatively accurate transformation relations in the coarse matching stage, and we obtain uniformly distributed and large numbers of correct matches in the fine matching stage. To obtain reliable transformation relations, we construct co-occurrence scale-spaces combined with the second-order Gaussian steerable filter. The construction of scale space has two main purposes: to solve the scale problem in image matching through different scale images, and to strengthen the structural features and smooth texture information of the image. The image rotation problem is mainly solved by the main direction angle of the image. Compared with the texture information in the multimodal image, the main direction angle calculated by the structure information is more reliable. We can better describe the feature points for different types of multimodal images by designing a reasonable local feature descriptor. The above solution strategy ensures that we can obtain relatively accurate transformation relations on the eight data

types. After obtaining a reliable transformation relationship, we use template matching to extract the correct match further. Therefore, our proposed G-CoFSM can handle multiple simultaneous problems in multimodal image matching.

Compared with G-CoFSM, CoFSM performs well on optical–optical, optical–infrared, optical–map, and optical–SAR image pairs. However, the performance drops significantly on day–night, scale, and rotation, especially as the rotation shows the worst matching effect. This is because CoF's enhancement of image edge structure and smoothing of texture information is much worse than that of $G^{\theta}_{2,\sigma} - CoF$. Therefore, the orientation angle calculated using CoFSM is not as accurate as that of G-CoFSM. At the same time, we found that CoFSM often has a non-response in matching. Through further analysis of CoFSM, we found that CoFSM adopts a strategy similar to that of this paper in the step of eliminating error points: first, we use the FSC algorithm to obtain the coarse matching result, and then we add the obtained transformation relationship to the nearest neighbor matching to obtain more correct matches. However, compared with our two-step matching method, this strategy has higher requirements on the coarse matching results of the CoFSM. A wrong transformation relationship or a small number of successful matches will not only lead to the failure of the matching, but also affect the incorrect convergence when using the FSC algorithm for the second time. Although the improved nearest neighbor constraint adopted using CoFSM can greatly improve the registration accuracy and the number of correct matches, it will also lead to an unacceptable result: the program does not respond.

The performances of the HAPCG and LPSO algorithms are quite unstable in the test. Although HAPCG and LPSO can obtain more correct matches than RIFT in some cases, the stability of matching is our first consideration. Compared with RIFT, both HAPCG and LPSO use the phase consistency direction as the main direction, and the GLOH descriptor to describe the feature points. The direction of phase consistency can reflect the change of image direction, and it has a good resistance to the NRDs of multimodal images. However, considering that the first-order direction information is sensitive to NRD and noise, the calculation of the second-order gradient based on the phase-consistent direction may be able to reflect the main direction of the image better.

RIFT has the best matching success rate among the five algorithms compared, but there are more failures on the scale images. Unlike the other methods, RIFT does not calculate the main direction directly. It uses the MIM feature map calculated based on the phase consistency information to construct the feature descriptor, and realizes the direction invariance by analyzing the influence of different orientation angles on the MIM feature, such that this method makes RIFT more robust. It is a pity that the RIFT algorithm does not consider how to deal with scale images. Therefore, RIFT cannot be applied to the matching of scale images. Of course, RIFT can solve the problem of not being able to match scale images by introducing scale space. However, considering the computational complexity of RIFT, this approach may have better improvements. The matching accuracy of RIFT is lower than that of the CoFSM algorithm among the five methods compared. Still, if the matching results of RIFT are used for fine matching, the matching accuracy will be greatly improved.

5.2. Influence Analysis of Rough Matching on the Final Result

As mentioned above, our final matching results are based on the results of preliminary matching. This section will discuss the effect of our preliminary matching algorithm. Generally speaking, if we want to obtain accurate matching results in the fine matching step, the transformation matrix we obtained in the coarse matching step should be as precise as possible. Therefore, this method's matching accuracy of rough matching is crucial. Figure 12 shows the proposed method's specific matching results on eight data types, and the matching evaluation indexes are NCM and RMSE. From Figure 12a, we can find that our proposed method has the best effect on extracting corresponding feature points on optical–infrared image data. The effect of extraction on three types of data: day–night, scale, and rotation, is relatively poor. The NCM_{ave} of the preliminary matching

method in this paper is 120.9, 79.3, and 265.5 on three data types of day–night, scale, and rotation, respectively, as shown in Table 8. The extraction effect on the scale image is the worst; NCM_{ave} is only 79.3. This shows that the performance of our preliminary matching method for day–night, scale, and rotation data remain to be improved. Figure 12b and Table 8 show the accuracy error of our method on all types of data. The $RMSE_{ave}$ of eight types of data are 1.764, 1.822, 1.852, 1.843, 1.909, 1.868, 1.841, and 1.843. Therefore, our coarse matching method is robust against eight types of data matching, which is also the key to obtaining accurate matching results in the fine matching step.

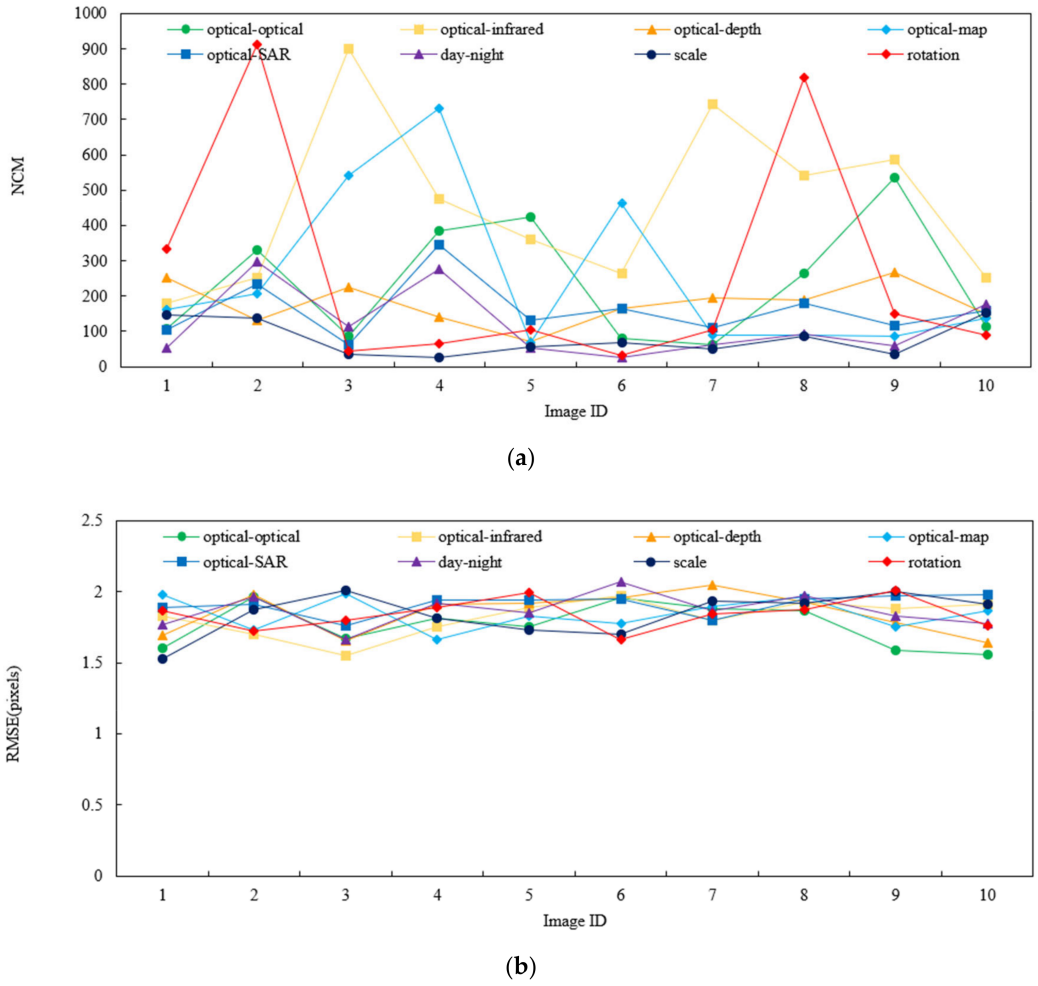

Figure 12. Coarse matching results of G-CoFTM on the NCM (**a**) and RMSE metrics (**b**).

Table 8. The average results of NCM and RMSE in Figure 12 for each type of image pair.

Criteria	Optical–Optical	Optical–Infrared	Optical–Depth	Optical–Map	Optical–SAR	Day–Night	Scale	Rotation
NCM_{ave}	238.6	455.6	178.4	257.1	160.7	120.9	79.3	265.5
$RMSE_{ave}$	1.764	1.822	1.852	1.843	1.909	1.868	1.841	1.843

5.3. Fusion and Registration Performance Analysis

In this section, we visualize the final registration results, obtain the optimal transformation parameters according to the least squares method, and display them using fusion graphs and checkerboard grid graphs. It can be seen in Figures 13 and 14 that although there are apparent NRD, rotation, and scale differences in multimodal image matching, our proposed **method** can still complete the image matching work well. The fusion graph

shown in Figure 13 shows no obvious ghosting using the proposed algorithm. At the same time, in the chessboard shown in Figure 14, each chessboard edge can match well without obvious misalignment, verifying that the method has good universality.

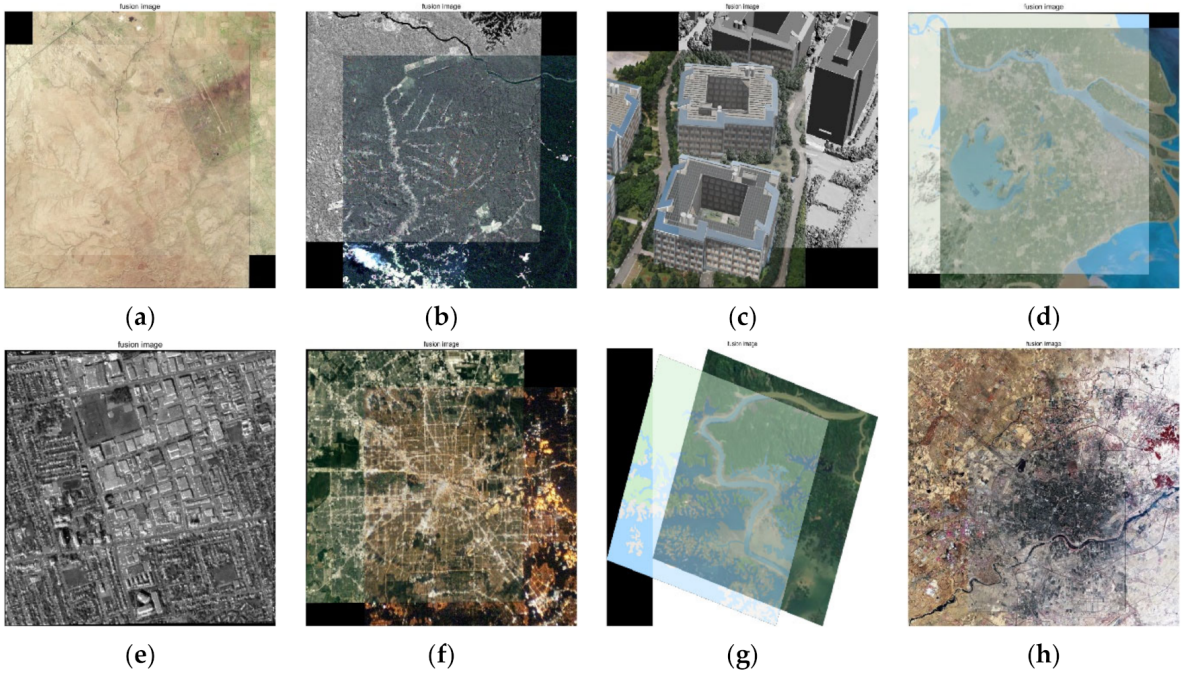

Figure 13. Fusion effect of image pairs: (**a**) optical–optical; (**b**) optical–infrared; (**c**) optical–depth; (**d**) optical–map; (**e**) optical–SAR; (**f**) day–night; (**g**) rotation; (**h**) scale.

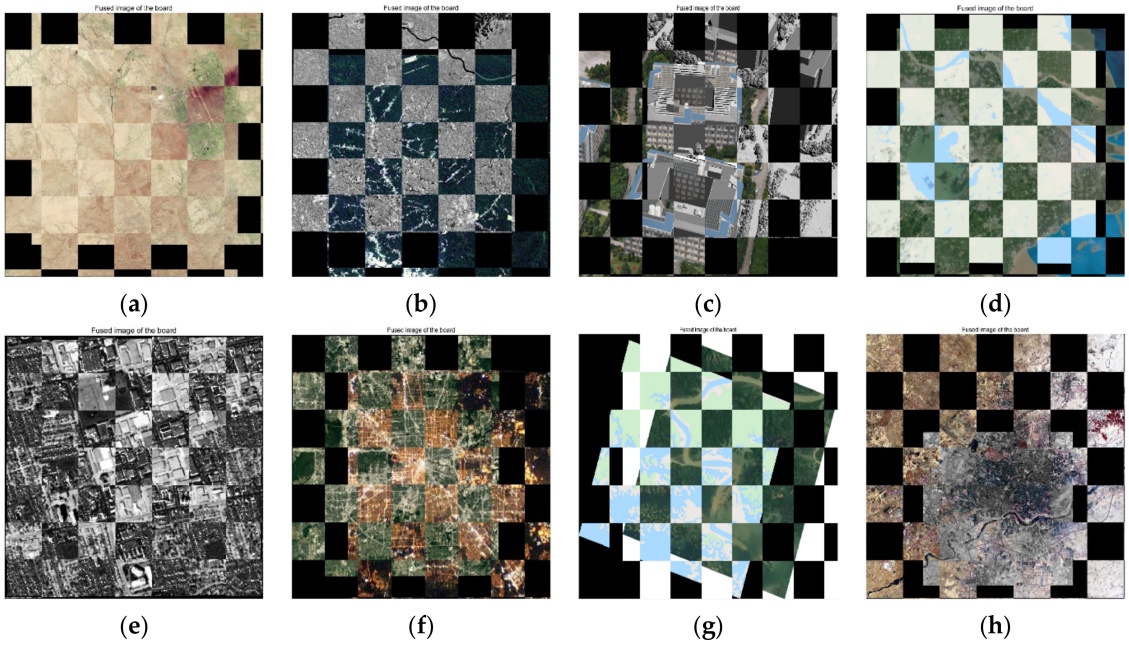

Figure 14. Registration effect of image pairs: (**a**) optical–optical; (**b**) optical–infrared; (**c**) optical–depth; (**d**) optical–map; (**e**) optical–SAR; (**f**) day–night; (**g**) rotation; (**h**) scale.

6. Conclusions

In order to better solve the NRD problem in multimodal images matching, we propose a two-step matching method based on co-occurrence scale space combined with the second-order Gaussian steerable filter called G-CoFTM. We first combine the second-order Gaussian

steerable filter and co-occurrence filter to construct the scale space of the image. This is to better preserve the edge and detail features of the images. Then, the preliminary matching of the image is completed by using the second-order gradient of the image and the improved GLOH descriptor to describe the features. Finally, based on preliminary matching, a more accurate matching is performed by using the optimized 3DPC descriptor combined with the preliminary matching results. Compared with other state-of-the-art multimodal image matching methods, the proposed method has more evident advantages in universality, NCM and RMSE. Compared with CoFSM, RIFT, HAPCG, and LPSO, the mean NCMs are 5.322, 11.503, 8.607, and 16.429 times those of the above four methods, and the mean RMSEs are increased by 2.278, 1.576, 3.393, and 3.264, respectively. At the moment, we are trying to apply the method proposed in this paper to the registration of panoramic images and mobile LiDAR data. If you are interested in our work, please read the article [51].

However, in this paper, we find that although G-CoFTM can effectively deal with the scale and rotation problems in multimodal image matching, the number of corresponding feature points in the extracted scale and rotation images is less than 100 in most cases. That is to say, if there are large scaling and rotation differences between the two images, our method may not be applicable. At the same time, for the day–night problem, how to use the night light information more effectively is also a problem that we will face. Therefore, in a future study, we will explore this more deeply: (1) design a better edge-preserving filter, smooth the texture information of the image, and strengthen the contour features; (2) build a more reliable scale space to solve the scaling problem of image matching; and (3) think about how to use the local features of the image to enhance the uniqueness of the feature descriptor.

Author Contributions: G.W. carried out the empirical studies and the literature review, and drafted the manuscript; R.Z., C.W., Y.X., Z.Y. and K.Y. helped to draft and review the manuscript, and communicated with the editor of the journal. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Technologies Research and Development, grant number: 2022YFB3904101, and the National Natural Science Foundation of China, grant number: U22A20568 and 42071444.

Data Availability Statement: Not applicable.

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

References

1. Li, J.; Xu, W.; Shi, P.; Zhang, Y.; Hu, Q. LNIFT: Locally Normalized Image for Rotation Invariant Multimodal Feature Matching. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
2. Alam, F.; Rahman, S.U. Challenges and Solutions in Multimodal Medical Image Subregion Detection and Registration. *J. Med. Imaging Radiat. Sci.* **2019**, *50*, 24–30. [CrossRef] [PubMed]
3. Hill, D.L.G.; Batchelor, P.G.; Holden, M.; Hawkes, D.J. Medical image registration. *Phys. Med. Biol.* **2001**, *46*, R1–R45. [CrossRef] [PubMed]
4. Yogheedha, K.; Nasir, A.; Jaafar, H.; Mamduh, S. Automatic vehicle license plate recognition system based on image processing and template matching approach. In Proceedings of the 2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA), Kuching, Malaysia, 15–17 August 2018; pp. 1–8.
5. Markiewicz, J.; Abratkiewicz, K.; Gromek, A.; Samczynski, W.; Gromek, D. Geometrical Matching of SAR and Optical Images Utilizing ASIFT Features for SAR-based Navigation Aided Systems. *Sensors* **2019**, *19*, 5500. [CrossRef]
6. Hou, B.; Wang, J.; Zhou, H. Navigation landmark recognition and matching algorithm based on the improved SURF. In Proceedings of the 2019 Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; pp. 1356–1361.
7. 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]
8. 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]
9. Niu, X.; Gong, M.; Zhan, T.; Yang, Y. A Conditional Adversarial Network for Change Detection in Heterogeneous Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 45–49. [CrossRef]

10. 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]
11. Maes, F.; Collignon, A.; Vandermeulen, D.; Marchal, G.; Suetens, P. Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imaging* **1997**, *16*, 187–198. [CrossRef]
12. Bay, H.; Tuytelaars, T.; van Gool, L. SURF: Speeded Up Robust Features. In *Computer Vision–ECCV 2006*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 404–417.
13. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
14. Viola, P.; Wells, W.M., III. Alignment by maximization of mutual information. *Int. J. Comput. Vision* **1997**, *24*, 137–154. [CrossRef]
15. 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 ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, Prague, Czech Republic, 12–19 July 2016; pp. 9–16.
16. 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]
17. Dellinger, F.; Delon, J.; Gousseau, Y.; Michel, J. SAR-SIFT: A SIFT-Like Algorithm for SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 453–466. [CrossRef]
18. Li, J.; Hu, Q.; Ai, M. RIFT: Multimodal image matching based on radiation-variation insensitive feature transform. *IEEE Trans. Image Process.* **2020**, *29*, 3296–3310. [CrossRef] [PubMed]
19. Yi, K.M.; Trulls, E.; Lepetit, V.; Fua, P. LIFT: Learned invariant feature transform. In *Computer Vision—ECCV 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 467–483.
20. Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; Sattler, T. D2-net: A trainable cnn for joint description and detection of local features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 8092–8101.
21. Efe, U.; Ince, K.G.; Alatan, A. Dfm: A performance baseline for deep feature matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–25 June 2019.
22. Yu, K.; Zheng, X.; Duan, Y. NCFT: Automatic Matching of Multimodal Image Based on Nonlinear Consistent Feature Transform. *IEEE Trans. Geosci. Remote Sens.* **2022**, *19*, 1–5. [CrossRef]
23. Jevnisek, R.; Shai, A. Co-occurrence Filter. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
24. Yao, Y.; Zhang, Y.; Wan, Y.; Liu, X.; Yan, X.; Li, J. Multi-Modal Remote Sensing Image Matching Considering Co-Occurrence Filter. *IEEE Trans. Image Process.* **2022**, *31*, 2584–2597. [CrossRef]
25. Jiang, Y. Optical/SAR image registration based on cross-correlation with multi-scale and multi-direction Gabor characteristic matrixes. In Proceedings of the 2013 IET International Radar Conference (IRC), Xi'an, China, 14–16 April 2013; pp. 1–4.
26. Zitová, B.; Flusser, J. Image registration methods: A survey. *Image Vis. Comput.* **2003**, *21*, 977–1000. [CrossRef]
27. Uss, M.; Vozel, B.; Lukin, V.; Chehdi, K. Multimodal remote sensing images registration with accuracy estimation at local and global scales. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6587–6605. [CrossRef]
28. Suri, S.; Reinartz, P. Mutual-Information-Based Registration of TerraSAR-X and Ikonos Imagery in Urban Areas. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 939–949. [CrossRef]
29. 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]
30. 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]
31. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
32. Ke, Y.; Sukthankar, R. PCA-SIFT: A more distinctive representation for local image descriptors. *CVPR* **2004**, *4*, 506–513.
33. Sedaghat, A.; Ebadi, H. Remote sensing image matching based on adaptive binning SIFT descriptor. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 5283–5293. [CrossRef]
34. Xiong, X.; Jin, G.; Xu, Q.; Zhang, H. Self-similarity features for multimodal remote sensing image matching. *IEEE J.-STARS.* **2021**, *14*, 12440–12454. [CrossRef]
35. 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.
36. Yang, W.; Xu, C.; Mei, L.; Yao, Y.; Liu, C. LPSO: Multi-source image matching considering the description of local phase sharpness orientation. *IEEE Photonics J.* **2022**, *14*, 7811109. [CrossRef]
37. Hughes, L.; Schmitt, M.; Zhu, X. Mining hard negative samples for SAR-optical image matching using generative adversarial networks. *Remote Sens.* **2018**, *10*, 1552. [CrossRef]
38. Wang, S.; Quan, D.; Liang, X.; Ning, M.; Guo, Y.; Jiao, L. A deep learning framework for remote sensing image registration. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 148–164. [CrossRef]
39. Han, Z.; Weiping, N.; Weidong, Y.; Deliang, X. Registration of multimodal remote sensing image based on deep fully convolutional neural network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3028–3042.

40. Ye, F.; Su, Y.; Xiao, H.; Zhao, X.; Min, W. Remote sensing image registration using convolutional neural network features. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 232–236. [CrossRef]
41. Ma, W.P.; Wen, Z.L.; Wu, Y.; Jiao, L.C.; Gong, M.G.; Zheng, Y.F.; Liu, L. Remote sensing image registration with modified SIFT and enhanced feature matching. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 3–7. [CrossRef]
42. 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]
43. Lindeberg, T. Scale-space theory: A basic tool for analyzing structures at different scales. *J. Appl. Stat.* **1994**, *21*, 225–270. [CrossRef]
44. Ye, Y.; Bai, Z.; Tang, T. A robust multimodal remote sensing image registration method and system using steerable filters with first- and second-order gradients. *ISPRS J. Photogramm. Remote Sens.* **2022**, *188*, 331–350. [CrossRef]
45. Studholme, C.; Hill, D.L.G.; Hawkes, D.J. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognit.* **1999**, *32*, 71–86. [CrossRef]
46. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Proc.* **2004**, *13*, 600–612. [CrossRef]
47. 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]
48. Yu, B.; Gabriel, D.; Noble, L.; An, K.-N. Estimate of the optimum cutoff frequency for the Butterworth low-pass digital filter. *J. Appl. Biomech.* **1999**, *15*, 318–329. [CrossRef]
49. Kovesi, P. Image features from phase congruency. *Videre J. Comput. Vis. Res.* **1999**, *1*, 1–26.
50. Kuglin, C.D. The phase correlation image alignment method. In Proceedings of the International Conference on Cybernetics and Society/IEEE Systems, Man, and Cybernetics Society, New York, NY, USA, 26–28 October 1975.
51. Wan, G.; Wang, Y.; Wang, T.; Zhu, N.; Zhang, R.; Zhong, R. Automatic Registration for Panoramic Images and Mobile LiDAR Data Based on Phase Hybrid Geometry Index Features. *Remote Sens.* **2022**, *14*, 4783. [CrossRef]