Information Fusion 102 (2024) 102027

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

POS-GIFT: A geometric and intensity-invariant feature transformation for multimodal images

Zhuolu Hou, Yuxuan Liu *, Li Zhang

Institute of Photogrammetry and Remote Sensing, Chinese Academy of Surveying and Mapping (CASM), Beijing 100036, China

ARTICLE INFO

Keywords:
Multimodal image matching
Nonlinear intensity distortion (NID)
Rotation invariance
Position-orientation-scale guided inlier recovery strategy (POS)

ABSTRACT

Multimodal image matching suffers from severe geometric and nonlinear intensity distortion (NID). Towards this problem, we propose a multimodal image matching algorithm based on multi-orientation filtering results, called position-orientation-scale guided geometric and intensity-invariant feature transformation (POS-GIFT). First, we design a multi-layer circular point sampling pattern to effectively capture the local image structure. Then, we propose a novel feature descriptor that can work robustly across rotational differences in [0°, 360°) in the presence of NID. Specifically, we (1) integrate the multi-orientation filtering response in the local neighborhood with a Gaussian weight to form the feature of each sampled point (GFP), (2) build feature vectors for each orientation by concatenating the features of points grouped by orientation, (3) estimate the primary orientation by finding the feature vector with the largest norm which is constructed in the previous step, (4) modify the order of elements of GFP, and (5) finally concatenate the features of all sampled points in a certain order to form the complete feature descriptor. At last, we propose a position-orientation-scale guided inlier recovery strategy (POS) by integrating the global position, orientation, and scale information to further improve the matching performance, especially the number and distribution of correct matches in texture-less and complex areas. Experimental results on various multimodal datasets from remote sensing, medical, and computer vision imaging domains show that POS-GIFT outperforms eight state-of-the-art multimodal image feature matching algorithms which are five handcrafted-based methods, OS-SIFT, PSO-SIFT, LGHD, RIFT, and LNIFT, and three learning-based methods RedFeat, MatchFormer, and SemLA by several times in terms of correct matches while improving the root-mean-square error to around 1 pixel. Our implementation is available at https://github.com/Zhuolu-Hou/POS-GIFT.

1. Introduction

Image matching, termed as finding correspondences between images, is a fundamental work for many photogrammetry and computer tasks [1–5], such as image registration [6–8], image stitching [9], change detection [10,11], SLAM [12–14], visual localization [15,16], navigation [17–19] and 3D reconstruction [20–22]. Even though this topic has been investigated for decades, accurate image matching for multimodal images is still far from being solved due to severe nonlinear intensity and geometric distortions caused by different imaging mechanisms, viewpoints, shooting times, etc.

Various image matching methods have been put forward to combat the modal difference by modifying the classical region-based and feature-based algorithms designed for the same modality of images [23,24]. Region-based matching methods typically use the correlation coefficient or mutual information between image blocks to determine their similarity [25,26]. However, the correlation coefficient is sensitive to nonlinear intensity changes, and the mutual information is prone to local optima [27]. Towards this, Ye et al. proposed using phase-consistent amplitude and orientation to generate orientation-phase consistency histograms (HOPC) [28] and computing orientated gradients to form channel features of orientated gradients (CFOG) [29] to overcome nonlinear intensity distortion(NID). Even though these area-based methods have relatively matching accuracy, they heavily rely on prior positional information between images and are sensitive to geometric distortions.

Different from area-based methods, feature-based methods extract scale and orientation invariant features, showing better robustness against geometric distortions. Yet, traditional image features, such as SIFT[30] and SURF [31], are based on intensity, gradient, and oriented

* Corresponding author.
E-mail address: yxliu@casm.ac.cn (Y. Liu).

https://doi.org/10.1016/j.inffus.2023.102027
Received 5 May 2023; Received in revised form 10 August 2023; Accepted 18 September 2023
Available online 18 September 2023

gradient information, which are sensitive to nonlinear intensity differences. The performance will decrease significantly when handling multimodal images. Recent research [29,32,33] shows that the features constructed based on multi-scale and multi-orientation filtering results have good robustness against NID, such as the Log-Gabor histogram descriptor (LGHD) [34] and the radiation invariant feature transform (RIFT) [35]. In terms of geometric distortions, these methods cannot handle large image rotations, and the reason is that these methods apply the traditional primary orientation calculation methods to eliminate the effect of rotation, which performs poorly on multimodal images. These methods exhibit limited robustness in handling significant image rotations due to their reliance on conventional dominant orientation estimation algorithms. As a result, their performance is unsatisfactory in image matching tasks involving severe geometric distortions, such as rotation and scale variations, notably in scenarios like street-to-aerial image geo-localization [36], UAV-based precision agriculture [37], traffic management [38], and marine organism observation by underwater robots equipped with multiple sensors [39]. In addition, compared with the same-modal images, more false matches are produced during the image matching process, significantly decreasing the creditability of the obtained matches and leading to matching failure.

With the rapid development of artificial intelligence technology, deep learning is also introduced to solve the task of multimodal image matching, and several methods have been developed, such as MU-Net [40], Loftr [41], and SemLA [42]. Even though these methods achieved promising results in their datasets, they have limitations and are currently challenging to apply to diverse engineering applications. Firstly, as far as we know, no multimodal image datasets can cover all types of image modalities and include sufficient images with labeled information for each modality. However, the actual situation can be very complex, and the performance could be decreased dramatically when it encounters cases not involved in the training process. Secondly, the mainstream learning-based image matching frameworks are oriented to small-size natural vision images and do not consider specific spectral characteristics involved in the remote sensing images, medical images, etc., and the geometric distortion such as large scale change and image rotation commonly existing in the large-size remote sensing images. Therefore, they may not handle severe NID and significant geometric distortions well. Lastly, the training process requires extensive computation resources, which is expensive and inefficient, further constraining its application.

In this paper, we propose a novel handcrafted multimodal image matching method, position-orientation-scale guided geometric and intensity invariant feature transformation (POS-GIFT), which innovatively put forward a robust feature descriptor robust to NID and an accurate and robust primary orientation method. We construct the descriptor based on the multi-scale and multi-orientation filtering results to exploit their advantage of good robustness to NID. Considering that the local neighborhood information reflects the local image structure, we design a multi-layer circular point sampling pattern that mimics the distribution of photoreceptor cells in the human visual system and integrates the feature of the sampled points with a Gaussian weighted method. Specifically, we robustly estimate the primary orientation by analyzing the norm information of the feature at different angles and modifying the order of different orientations of filtering results to achieve rotation invariance. At last, we develop an effective inlier recovery strategy that considers the consistency of position, orientation, and scale information simultaneously.

In detail, the main contributions of this work are as follows:

1) We propose a novel multimodal image matching method, POS-GIFT, invariant to geometric transformations (translation, rotation, scale change) and non-linear intensity distortions. POS-GIFT outperforms eight state-of-the-art multimodal matching algorithms, which are five handcrafted-based methods, OS-SIFT, PSO-SIFT, LGHD, RIFT, and LNIFT, and three learning-based methods RedFeat, MatchFormer, and SemLA by several times in terms of the number of correct matching points (*NCM*) while maintaining high accuracy on various multimodal image datasets.
2) We propose a robust primary orientation estimation method based on multi-scale Gaussian-weighted multi-orientation filter responses, achieving rotation invariant at a full range of rotation angles of [0°, 360°). Since the weighted norm information of multi-orientation filter responses can maintain rotational invariance under various NID, this method can be applied to various related engineering applications.
3) We introduce an effective inlier recovery strategy, which comprehensively utilizes the characteristic of position-orientation-scale consistency of correspondences and significantly improves the credibility of obtained matches with respect to *NCM* and accuracy.

2. Related work

Generally, multimodal image matching algorithms can be divided into three categories: region-based, feature-based, and learning-based image matching methods [8]. A brief review of the multimodal image matching methods is as follows.

Region-based Matching: The core idea is to calculate the similarity between the target image block and the reference image block based on a predefined similarity measure and select the reference image block with the highest similarity as the correct correspondence [43]. Correlation coefficients [44] and mutual information [45–47] are commonly used similarity metrics. However, the correlation coefficient is sensitive to nonlinear intensity difference, and MI is prone to fall into local optimum, making them unable to match the multimodal images effectively. Considering that the image structure remains unchanged across different image modalities, recent studies have developed several area descriptors based on the image structure information. For example, Ye et al. proposed the histogram of oriented phase consistency (HOPC) [28] algorithm, Ye et al. proposed an algorithm based on the channel feature of orientated gradient (CFOG) [29], and Fan et al. [48] proposed a pixel-level feature based on angle-weighted orientated gradient. Even though these methods demonstrate good robustness against NID, they are easily affected by geometric distortions. Therefore, the area-based methods [49] highly rely on prior information to coarsely eliminate the geometric distortions, including the scale change and image rotation. For example, CFOG requires using the rational function model (RFM) of remote sensing images to determine the approximate range of matching.

Feature-based Matching: Many classical feature-based methods have been proposed in recent decades, such as scale-invariant feature transform (SIFT) [30], speeded up robust features(SURF) [31], affine-SIFT (ASIFT) [50], and oriented FAST and rotated BRIEF (ORB) [51]. These methods are designed for linear intensity changes, making them unsuitable for multimodal images. To adapt to NID, Xiang et al. [52] proposed the OS-SIFT method, which uses multi-scale Sobel and ROEWA filters to extract gradients and match them with the SIFT algorithm. Aguilera et al. [34] proposed a multimodal image feature descriptor, LGHD, based on multi-scale and multi-orientation Log-Gabor convolutional values. Based on LGHD, Li et al. [35] proposed the radiation–variation insensitive feature transformation (RIFT) algorithm. Instead of using the filtering responses, RIFT extracts the maximum index map (MIM) by searching the index of the maximum value among all orientations of filtering responses. It is proven to be more robust against NID. To resist rotational differences, RIFT builds an end-to-end circular structure with multiple index mappings to simulate different rotational differences between images. However, its performance degrades dramatically at large rotations. Liu et al. [53] further improved RIFT by constructing a BRIEF descriptor based on MIM, improving efficiency. The use of MIM increases the robustness to nonlinear intensity change, and it only encodes the index information rather than the complete filtering sequence with richer information.

Learning-based methods: The use of deep learning technology to achieve high-precision image matching has recently attracted much attention. SuperGlue [54] exploits self-attention and cross-attention between descriptors in graph neural networks for learned features and descriptors. Based on SuperGlue, LightGlue [55] adaptively adjusts the image matching model according to the matching difficulty, making it more efficient. COTR [56] proposes to use the TransFormer decoder and recursive amplification strategy to handle the matching task. LoFTR [41] proposes a detector-free dense matching algorithm based on the visual TransFormer model, which improves the ability to handle textureless regions. However, it uses a light TransFormer module to reduce the amount of computation, which results in LoFTR being subject to incorrect diffusion of attention [57]. Towards this, AspanFormer [58] introduces a Transformer-based detector-free matcher that is built on the hierarchical attention structure and can adjust attention span in a self-adaptive manner. Matchformer [59] employs a lightweight decoder with multi-scale features to reduce computation and uses cross-attention to improve robustness, improving the matching performance. These methods use natural vision images, like the datasets, Hpatches [60], ScanNet [61], and MegaDepth [62], and are not specifically designed for multimodal images. To treat multimodal images explicitly, Hybrid [63] proposes a new convolutional neural network (CNN) architecture that utilizes Siamese CNN and dual non-weight-sharing CNNs to tightly couple the generated feature detectors with the feature descriptors and obtains good results in VIS–NIR cross-modal scenarios. RedFeat [64] recouples the independent constraints of feature detection and matching, and proposes a super-detector with a large receptive field and a learnable non-maximal suppression layer, realizing four cross-modal scene matching. SemLA [42] introduces semantic information to constrain and guide the matching process. However, Hybrid, RedFeat, and SemLA cannot handle large rotation between images. After all, only a few image matching methods are focused on the image matching of multimodal images currently, and the performance and generalizability of these methods are still limited.

In summary, even though much effort has been made into the effective matching of multimodal images, it is still a challenging problem caused by the NID and the geometric distortions [65]. Most current methods focus on tackling the NID, ignoring geometric distortions. A few studies [66] try to achieve rotation invariance by using image gradient or intensity information to estimate a primary orientation for each feature. However, both image gradient and intensity are unstable because of the NID. Besides, there may be many false matches in the obtained matches even if strict parameters are set for the traditional error elimination method. In light of these problems, we propose a novel primary orientation estimation method that is not insensitive to image modalities and introduces the image pyramid strategy to combat scale change. Besides, we found that the global information (scale, rotation, position) of initial matches can effectively reduce incorrect matches caused by the similarity of local structures in multimodal images. We also designed a new POS inlier recovery method correspondingly. After integrating these improvements, POS-GIFT can obtain sufficient matches with high credibility under complex nonlinear intensity, scale, and rotation change circumstances.

3. The POS-GIFT method

As shown in Fig. 1, our method consists of four main steps. First, we build image pyramids on the reference image and sensed images to release the effect of scale change. Then, we generate a phase congruency (PC) [67] map based on the multi-scale and multi-orientation filtered images obtained with a Log-Gabor filter, and detect distinctive feature points on the PC maps. After that, we construct a GIFT descriptor which includes primary orientation estimation to achieve orientation invariance. At last, we match the descriptors with a nearest neighbor matching method and refine the matches with the proposed POS strategy.

3.1. Multiscale feature points detection

Due to NID, it is challenging to detect repeated and highly distinctive features on multimodal image pairs. Previous studies [28] prove that the PC map can well keep the salient structures regardless of image modalities. In this paper, we also extract the PC maps of each layer image in the image pyramid first and then detect feature points on them with the FAST [68] algorithm. Moreover, we apply the pyramid strategy to tackle scale change. Specifically, two octaves of pyramid images are generated. The first octave is generated by conducting a sequence of down-sampling and Gaussian smoothing processes on the original image. The second octave is constructed the same as the first octave, and the only difference is the base image is obtained by down-sampling the original image with a scale factor of 1.5.

Fig. 2 presents the detected feature points on a pair of an optical image and a depth image. We can see that the PC map can well keep the image structure regardless of different image modalities, and a large amount of repeated feature points are evenly distributed on the two images.

3.2. GIFT descriptor

In this section, we present a novel robust feature descriptor for multimodal images, which can effectively work with any angle of image rotation regardless of image modalities. In addition, we label the feature point being processed as the target point. The construction of the

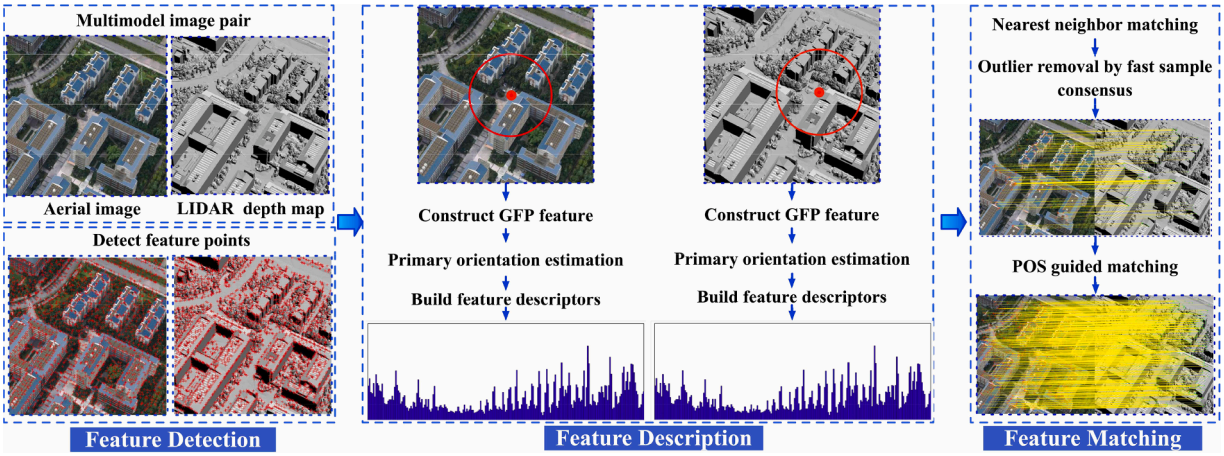

Fig. 1. The flowchart of the POS-GIFT.

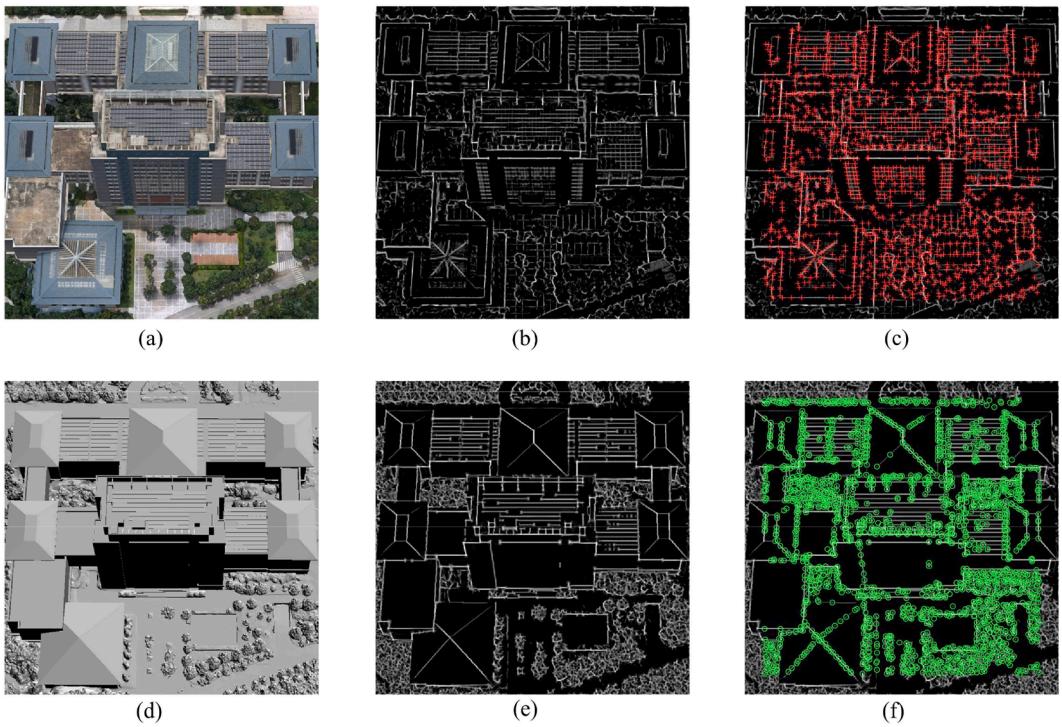

Fig. 2. The process of feature detection on a pair of multimodal images. (a) An optical image; (b) The generated PC map of (a); (c) The detected feature points on (b); (d) The corresponding depth map of (a); (e) The generated PC map of (d); (f) The detected feature points on (e).

descriptor involves four steps: (1) build an LG feature; (2) sample evenly distributed neighboring points; (3) construct a Gaussian weight feature (GFP)of each sampled point; (4) estimate the primary orientation, reorder of these features accordingly, and integrate them in a specific sequence to generate a rotation-invariant feature descriptor for the feature point.

3.2.1. LG feature

Note that multi-scale and multi-orientation filtered results have been obtained in the feature detection process. To comprehensively describe the feature in orientation, we sum up the filtered results across all scales for each orientation, which is helpful for the following primary orientation estimation but may lose the scale information of the multi-scale Log-Gabor filtered images [53]. As a result, a feature vector with a dimensional of a pre-defined number of orientations can be obtained for each point. To be simple, we call it as LG feature. The construction of LG features can be represented by the following equation.

$$LG(x,y) = \sum_{s} A_{so}(x,y) \tag{1}$$

where A_{so} is the filtering feature of the image by the Log-Gabor filter with a scale of s and direction o. A_{so} exhibits symmetry, meaning that A at a degrees is exactly identical to A at $(a+180)$ degrees. To avoid redundancy, we exclusively construct multi-directional Log-Gabor filters within the range of [0°, 180°). In this paper, we respectively set s to 4 and o to 6, then the LG feature has six layers, each of which integrates all the scaled Log-Gabor filter information in one orientation.

3.2.2. Points sampling pattern

Compared with a single point, the image structure can be better kept across image modalities. To capture image structure, we adopt a point sampling pattern analog to the human vision system. Some research [69, 70] has revealed that the retina is densely populated with photoreceptor cells that convert light signals into neural signals. As illustrated in Fig. 3 (a), the human retina can be partitioned into distinct regions, with photoreceptor cells in each region exhibiting varying sensitivities to light signals. To replicate this sensitivity to image intensity, we assign Gaussian kernels with varying sizes to sampling points in different locations. The detailed point sampling process is illustrated in Fig. 3(b), which can be described as follows. First, we generate $\boldsymbol{n}_1$ concentric circles with various radii around the target point. Next, each circle is divided into $\boldsymbol{n}_2$ equally spaced orientations, and a point is sampled for each orientation on each circle. In total, $\boldsymbol{n}_1$*$\boldsymbol{n}_2$ evenly distributed points are sampled, and these sampled points, together with the target point, form the proposed point sampling pattern. Large experiments (Section 4.2) reveal that good performance can be achieved when $\boldsymbol{n}_1$=3, which is consistent with the retinal region structure depicted in Fig. 3(a). The radius of concentric circles can be calculated according to the following equation.

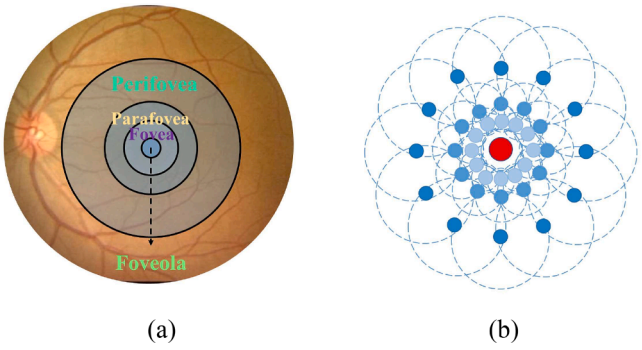

Fig. 3. Our point sampling pattern. (a) shows the distribution of photoreceptor cells in the human visual system; (b) the blue and red dots respectively represent the sampled and target points, and the blue dashed circles represent their sampling neighborhoods.

$$P_{i+1} = P_i + P_1 * i \tag{2}$$

where P_1, P_i, and P_{i+1} represent the radii of the first, i-th, and $(i+1)$-th concentric circles, respectively.

3.2.3. Gaussian weight features of the sampled points (GFPs)

In this section, we build a feature vector for each sampled point based on the multi-scale and multi-orientation filtering results. To fully apply the multi-orientation filtering results, we design a new encoding method called Gaussian weighting (GAUSS), as shown in Fig. 4. Essentially, GAUSS integrates the LG features in the neighborhood of the sampled point to form a feature vector to describe the sampled points. To be simple, we refer to the feature vector of a sampled point as GFP, which is constructed as follows.

1) We first determine a local circular area for each sampled point. To utilize more neighborhood information, the radius of the circular area is determined by the distance between the sampling point and the target point. The smaller the distance, the smaller the sampling radius, and vice versa. The sampling radius can be calculated with the following equation.

$$R_i = \frac{P_i}{2} \tag{3}$$

where R_i represents the sampling radius of the sampled point, $\lceil . \rceil$is a ceiling function, P_i represents the radius of the ith concentric circle. Specifically, the sampling radius of the target point is equal to R_1.

1) For each circular area, we extract the LG features of the points within the sampling area and build a Gaussian kernel whose radius is equal to the sampling radius with the following equation.

$$Gauss(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x^2 + y^2)}{2\sigma^2} \right) \tag{4}$$

where $Gauss(x, y, \sigma)$ represents the Gaussian kernel with a variance of σ.

1) We integrate the LG features in the local area to form the GFP feature for each sampled point (as shown in Fig. 5). Each layer of the LG features is separately added according to the weights calculated based on the constructed Gaussian kernel. In detail, the GFP can be calculated as follows.

$$GFP_{i,j} = \left[V_{i,j,1}; V_{i,j,2}; \ldots; V_{i,j,o} \right] \tag{5}$$

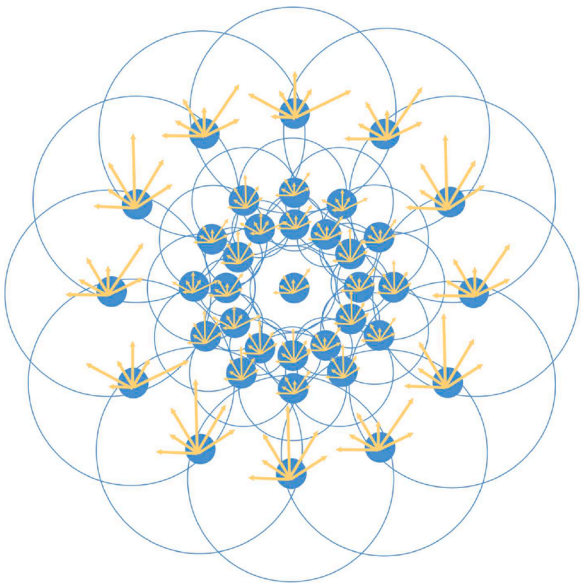

Fig. 5. Gaussian weight features of all sampled points (GFPs). The arrows in six directions represent the six dimensions of GFP, and the length of the arrows reflects the magnitude of the values in each dimension.

where (i, j) represents the index of concentric circle and orientation for the sampled point, and o is the number of filter orientations.

$$V_{i,j,o} = \sum_{\sqrt{(x-x_p)^2+(y-y_p)^2} < R_i} Gauss\left(x - x_p, y - y_p, \sigma_i\right) * LG(x, y, o) \tag{6}$$

$$\sigma_i = a * R_i + b \tag{7}$$

where (x_p, y_p) and R_i represents respectively the coordinates and sampling radius of the sampled point, a and b are constants, with an empirical value of 0.15 for a and 0.35 for b, σ_i is the Gaussian kernel variance of the sampling point located at the ith concentric circle. Specifically, the Gaussian kernel variance of target point is equal to σ_i.

3.2.4. Rotation-invariant feature descriptor

Image rotation is inevitable in most cases, and it is a bottleneck problem for current multimodal image matching methods due to the ineffective estimation of primary orientation caused by NID. We deeply exploit the characteristics of the LG features under image rotation and

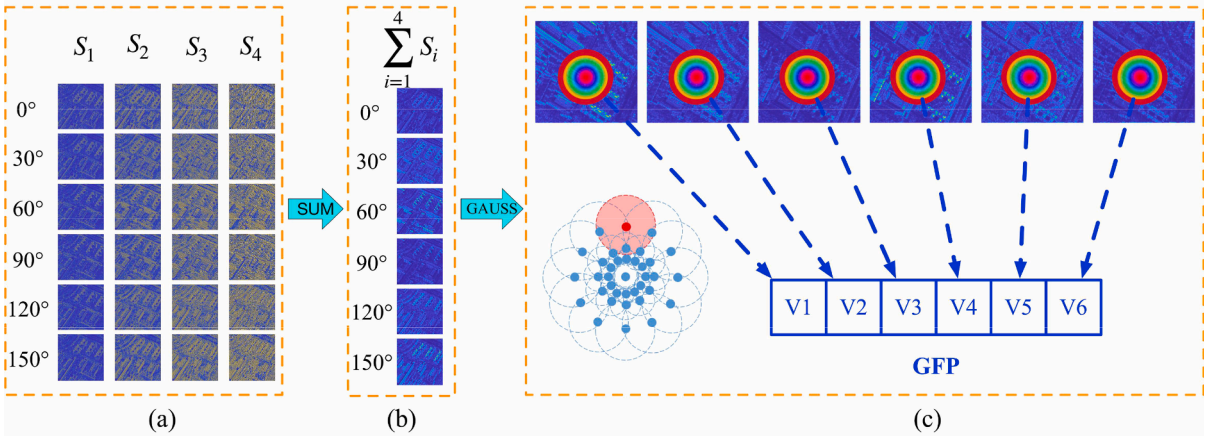

Fig. 4. The detailed GFP construction process for the red sampled point. (a) Multi-scale and multi-orientation Log-Gabor filtered results; (b) LG feature maps constructed from summing up the filtered results across all scales for each orientation; (c) The construction process of our proposed GFP feature, which integrates the multi-orientation filtering responses in a local neighborhood with a Gaussian weight.

found that the change of LG features meets a certain pattern. As shown in Fig. 6, the structure of image content represented by the norm values remains consistent when the image is rotated, and this gives us the possibility to find the primary orientation. However, we can also see that the LG feature map cannot be converted to the same as that of the LG feature map without rotation by simply rotating the LG feature map (Fig. 7). As displayed in Fig. 8, the orientation of the multi-orientation filter is fixed while the image is rotated, so we need to change the order of the filter sequences according to the rotation angle correspondingly as well. Fortunately, even though the order of the values of the sequence is changed, the norm values computed from the all-orientation filter response remain unchanged (Fig. 6). Based on the above analysis, we first use the norm map to estimate the primary orientation, then modify the order errors and rotation errors of LG features to achieve rotation invariance. Specifically, the norm value of the LG feature can be calculated as follows.

$$Norm_{LG}(x,y) = \sqrt{\sum_{i=1}^{O} LG(x,y,i)^2} \tag{8}$$

where O represents the orientation number of the Log-Gabor filter, $Norm_{LG}$ represents the norm information of LG features.

Fig. 9 gives the detailed process of primary orientation estimation and GIFT feature descriptor construction.

(1) As shown in Fig. 9(a), all the sampled points, excluding the target point, are divided into n_2 groups according to their orientation, and each group contains n_1 points.
(2) For each group, we concatenated the GFPs for the n_1 sampled points and concatenated them in a from-inside-to-outside order to form a vector with a dimension of $n_1 \times 6$ (Fig. 9(b)). We calculate the norm of the created vectors and find the vector with the largest norm, and take the orientation corresponding to the largest norm as the primary orientation. This process can be expressed in the following equation.

$$Ori_{Feature} = \underset{j\in\{1,2..n_2\}}{\operatorname{argmax}} \sqrt{\sum_{i} \left|GFP_{i,j}\right|^2} \tag{9}$$

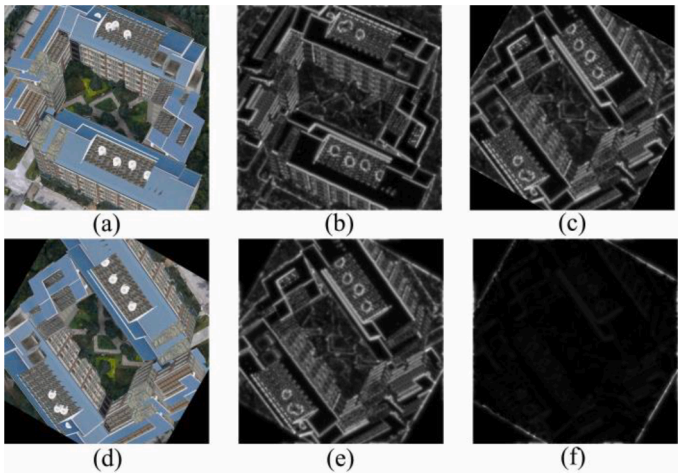

Fig. 6. The effect of rotation on the norm map of LG features. (a) The original image; (b) The norm map of (a); (c) is obtained by rotating (b) counterclockwise by 150°; (d) is obtained by rotating (b) counterclockwise by 150°; (e) The norm map (d); (f) The absolute of the difference map by subtracting (c) from (e).

where $Ori_{Feature}$ represents the estimated primary orientation, $|GFP_{i,j}|$ represents the norm of *GFP* for the sampled point located at the ith circle and jth orientation.

(1) If the value of the second largest norm is larger than 80 % of that of the largest norm, the corresponding orientation is taken as the second primary orientation, and building a feature descriptor for this direction is in the following steps. This strategy can increase the robustness against image rotation and improve the matching success.
(2) After obtaining the primary orientation, we further modify the order of the orientation of the Log-Gabor filter, which can be done by simply changing the orders of the initial filtering results instead. Specifically, we set the primary orientation as the start orientation of the Log-Gabor filters and change the order of the filter results correspondingly. As a consequence, the order of elements in GFP has changed accordingly. Moreover, A_{so} is consistent before and after rotation by 180°. Accordingly, the index of the first element in GFP can be calculated by the following equation.

$$IFE = \begin{cases} Ori_{Feature}, & Ori_{Feature} \le O \\ Ori_{Feature} - O, & Ori_{Feature} > O \end{cases} \tag{10}$$

where O represents the orientation number of the Log-Gabor filter, *IFE* is the index of the first element of GFP. Table 1 gives an example of the modification of the indices of the elements in GFP when $n_2 = 12$, $O = 6$.

(1) Note that the primary orientation has been estimated. We concatenate the GFPs of all sampled points to form the feature descriptor (Fig. 9(e)). First, the GFPs in each group are concatenated from insider to outsider, and the vector of $6*n_1$ is obtained. Then, we concatenated the obtained vectors of the n_2 groups from the primary orientation clockwise, appended the GFP of the target point to the end, and a $6*(n_1*n_2+1)$ vector is achieved, which is taken as the feature vector for a feature point. Large experiments showed that a good performance can be achieved when the values of n_1 and n_2 are set as 3 and 12, respectively. As a consequence, the proposed feature descriptor is a 222-dimensional vector.

3.3. Inlier recovery strategy based on position-orientation-scale guidance (POS)

To fulfill the matching task, we first calculate the distance between the constructed feature descriptors, and the point pair with the nearest distance is deemed as a match. Then, we use the fast sample consensus algorithm(FSC) [71] to further refine the matches and obtain the affine transformation matrix **M**. However, there are still incorrect matches in the obtained matches, especially in the areas with similar structures and texture-less areas, considering that our feature descriptor is designed based on the structural information to tackle the NID.

Towards this, we propose a position-orientation-scale guided inlier recovery strategy (as shown in Fig. 10), which uses global information from the initial matching to guide the matching process. By constraining the matching range and modifying the scale and primary orientation information of feature points, the matching performance, including the number of correct matches and matching accuracy, is significantly improved.

Then, we further decompose **M** with the following equation.

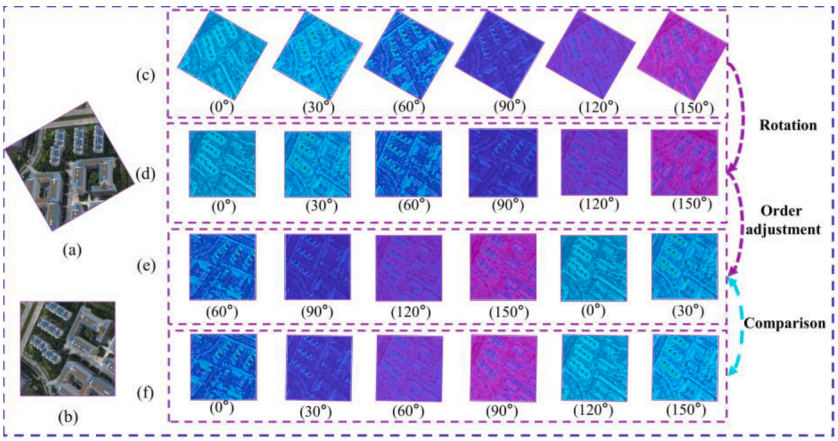

Fig. 7. The impact of rotation on LG features. (a) is an aerial image; (b) is obtained by rotating (a) counterclockwise by 60°; (c) is the generated LG feature maps of (a) in six orientations; (d) is obtained by rotating (c) counterclockwise by 60°; (e) is obtained by modifying the order of elements of the LG features based on the orientation angle; (f) is the generated LG feature maps of (b) in six orientations.

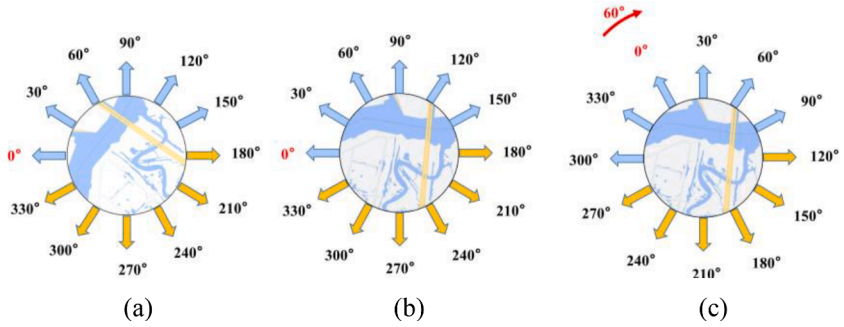

Fig. 8. The effect of rotation on multi-orientation Log-Gabor filters when o is 6. The arrows represent the multi-orientation filters. Given that the Log-Gabor filtering results of the orientation of a degrees are the same as that of $a + 180$ degrees, we only used the orientations of filters marked in blue, and the ones marked in yellow are calculated correspondingly. (a) The image and the orientations of Log-Gabor filters in the initial state; (b) the image is rotated clockwise by 60° while the orientations of Log-Gabor filters are kept unchanged based on the initial state; (c) the image and the orientations of the Log-Gabor filters are both rotated clockwise by 60° on the basis of the initial state.

$$\underset{\{\mathbf{M}\}}{\begin{bmatrix} a & c & e \\ b & d & f \\ 0 & 0 & 1 \end{bmatrix}} = \underset{\{\mathbf{T}\}}{\begin{bmatrix} 1 & 0 & e \\ 0 & 1 & f \\ 0 & 0 & 1 \end{bmatrix}} * \underset{\{\mathbf{R}\}}{\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}} * \underset{\{\mathbf{S}\}}{\begin{bmatrix} S_x & 0 & 0 \\ 0 & S_y & 0 \\ 0 & 0 & 1 \end{bmatrix}} * \underset{\{\mathbf{H}\}}{\begin{bmatrix} 1 & s & 0 \\ t & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \tag{11}$$

where **T**, **R**, **S**, and **H** respectively represent the translation, rotation, scaling, and shearing components, respectively. θ is the rotational difference, S_x represent the scale difference in the X direction, S_y represent the scale difference in the Y direction, s is the shear coefficient, and $t = 0$.

By deforming (11), we can obtain simultaneous equations as follows.

$$\begin{cases} a = S_x * \cos\theta - t * S_y * \sin\theta \\ b = S_x * \sin\theta + t * S_y * \cos\theta \\ c = s * S_x * \cos\theta - S_y * \sin\theta \\ d = s * S_x * \sin\theta + S_y * \cos\theta \end{cases} \tag{12}$$

Finally, the rotation θ and scale differences S_{image} of the image pair can be calculated with the following equations.

$$\theta = \arctan\left(\frac{b}{a}\right) \tag{13}$$

$$S_x = \sqrt{a^2 + b^2} \tag{14}$$

$$S_y = \frac{a * d - b * c}{\sqrt{a^2 + b^2}} \tag{15}$$

$$S_{image} = \sqrt{S_x^2 + S_y^2} \tag{16}$$

Based on the obtained **M**, θ, and S_{image}, the details of the position, orientation, and scale guidance are elaborated as follows.

Position guidance: Based on the initial matches, the affine transformation matrix **M** between the images is estimated first, then the feature points on the reference image are mapped to the sensed image. Each feature point on the referenced image is only matched to the predicted point, and the K points nearest to the predicted point on the sensed image, the default value of K is 20. By restricting the search range of the matching area, the possibility of false matching can be largely reduced.

Orientation guidance: Based on the initial matching results, we also get the rotation difference θ between the images. Therefore, we set the primary orientation of the reference image to 0 and the primary orientation of the target image to θ, as shown in Fig.10(d). By adjusting the primary orientations of all feature points in two images, the influence of principal orientation estimation on matching performance is eliminated.

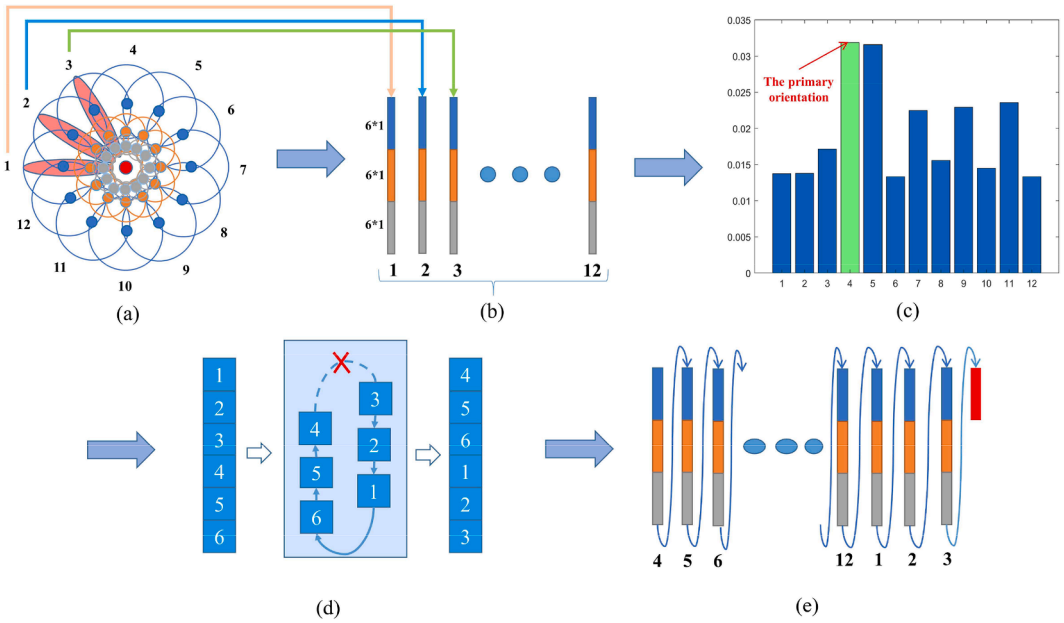

Fig. 9. The flowchart for constructing the GIFT descriptor. (a) shows the distribution of sampling points, which are grouped by orientation; (b) shows the 18-dimensional feature vector built by concatenating the features of the three points in the same direction; (c) calculates the norm value of the 12 constructed orientation vectors and takes the orientation (marked green) with the largest norm as the primary orientation; (d) shows the process of adjusting the element order of the GFP according to the primary orientation of 4; (e) concatenate the feature vectors of all the orientations from the primary orientation clockwise and append the GFP of the target point at the end to form the complete feature descriptor.

Table 1
The correspondence between the element order of GFP and the primary orientation. The red number represents *IFE* in Eq. (10).

The primary orientation	The element order of GFP
1,7	[1,2,3,4,5,6]
2,8	[2,3,4,5,6,1]
3,9	[3,4,5,6,1,2]
4,10	[4,5,6,1,2,3]
5,11	[5,6,1,2,3,4]
6,12	[6,1,2,3,4,5]

To reduce the computational complexity, we align the primary orientation of the image to the nearest 30-degree multiple, which enables us to rotate the feature points by only changing the order of the elements in the feature vector without regenerating the descriptors.

Scale guidance: Based on the initial matching results, we calculate the scale difference S_{image} between the two images. Then, we adjust the P_1 in Eq. (2) and recreate the feature descriptor with scale invariance.

4. Experimental results

To demonstrate the effectiveness of the proposed methods, we

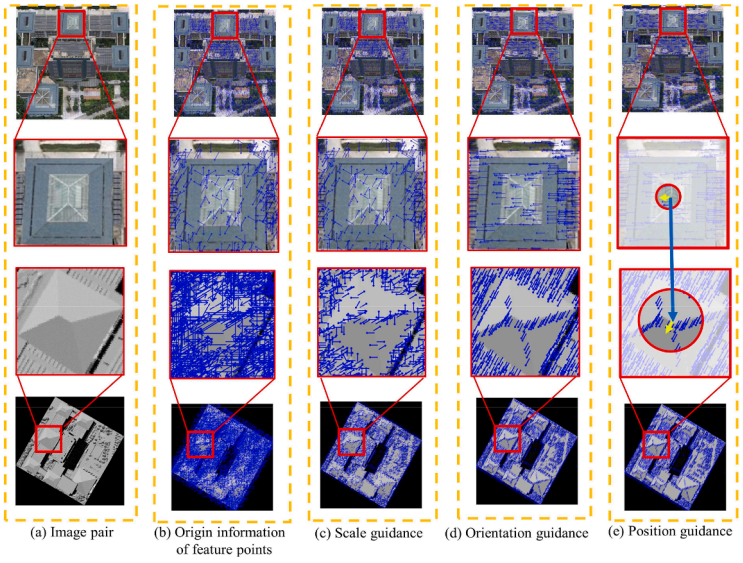

Fig. 10. The proposed position-orientation-scale (POS) inlier recovery strategy. The arrows in the figure represent feature point information: the starting point, direction, and length of the arrow represent the position, orientation, and scale information of the feature point, respectively. (a) is the multimodal image pair to be matched; (b) displays initial extracted feature points; (c) presents the feature points after scale modification on the basis of (b); (d) shows the feature points after primary orientation modification from (c); and (e) illustrates the position constraint where the searching area is marked with a mask.

designed four groups of experiments. Firstly, we conduct a parameter study to test the sensitivity to different parameters and find the proper parameters. Secondly, we prove the superiority of our method by comparing the qualitative and quantitative results of our method with that of eight latest state-of-the-art multimodal image matching methods, including five handcrafted feature-based matching methods, PSO-SIFT [72], OS-SIFT [52], LGHD [34], RIFT [35], LNIFT [7], and three deep learning-based matching methods, RedFeat [64], MatchFormer [59], SemLA [42], on various multimodality of images. In addition, we also add the method of POS-GIFT without the POS strategy as a comparative method, which is named GIFT. For handcrafted feature-based methods, the author-provided source codes and parameters (Table 2) are applied; for the deep-learning methods, the author-provided pre-trained models and default parameters are applied (Table 2). Specifically, we lowered the thresholds of feature detection for PSO-SIFT and OS-SFIT to obtain more feature points considering that insufficient feature points can be found with the pre-defined parameters. Thirdly, we show the strong rotation invariance of our method on a large number of images with different image rotations. Lastly, we evaluate the running efficiency of POS-GIFT, OS-SIFT, LGHD, PSO-SIFT, RIFT, and LNIFT in detail.

For quantitative evaluation, three measure metrics are employed, which are the number of correct matches (*NCM*), success rate (*SR*), and root mean square error (*RMSE*). *NCM* reflects the strength of the matching ability; *RMSE* reflects the matching accuracy, with lower *RMSE* indicating higher accuracy; *SR* reflects the robustness of the matching algorithm, with higher *SR* indicating greater robustness. Moreover, we consider the matching point with a distance greater than five pixels from the corresponding ground truth point an incorrect match and the image pair with *NCM* smaller than four as a matching failure. The *RMSE*, and *SR* are calculated as follows:

$$RMSE = \sqrt{\frac{1}{C}\sum_{i=1}^{C}\left(x_i^{'} - x_i^{''}\right)^2 + \left(y_i^{'} - y_i^{''}\right)^2} \tag{17}$$

$$SR = \frac{N_{success}}{N_{total}} * 100 \tag{18}$$

where C represents the number of correct matched points, $(x_i^{'}, y_i^{'})$ and $(x_i^{''}, y_i^{''})$ are respectively the coordinates of the matched point and the corresponding ground truth point; $N_{success}$ and N_{total} represent the number of successfully matched image pairs and the number of image pairs used for matching, respectively. For each image pair, 10–20 evenly distributed correspondences are employed as the ground truth.

Table 2
Detailed settings of nine methods.

Method	Parameters
POS-GIFT	Concentric circles number (n_1): 3; The radius of the first concentric circles (P_1): 6; The number of directions (n_2) per circle: 12; The number of nearest neighbors (K) in the POS strategy: 20; The scale number of Log-Gabor filter: 4; The orientation number of Log-Gabor filter: 6.
PSO-SIFT	Initial variance: 1.6; Patch size: 24*scale; Descriptor size: 136; Edge threshold 31;Contrast threshold: 0.001.
OS-SIFT	Initial variance: 1.6; Patch size: 24*scale; Descriptor size: 136; Harris function threshold 0.001; Scale ratio: $\sqrt[3]{2}$.
LGHD	Patch size: 100; Descriptor size: 384; FAST mincontrast: 0.1; Scale ratio: $\sqrt[3]{2}$;Log-Gabor scale: 4; Log-Gabor orientation: 6.
RIFT	Patch size: 96; Descriptor size: 216;Fast mincontrast: 0.001; Log-Gabor scale: 4;Log-Gabor oritation: 6.
LNIFT	Patch size: 96; Descriptor size: 256; ORB edge threshold: 5; Filter windows size: $s = 3$.
RedFeat	Descriptor size: 128; Min/Max size: 100/1000; Scale ratio: 2×0.25; Keypoint number: 4096; Reliability threshold: 0.5; Repeatability threshold: 0.4.
MatchFormer	Backbone: “largela”; Image size: {640,480}; Scens: “outdoor ”; Resolution: (8,2).
SemLA	Match mode: "scene"; Image size: {640,480}; Semantic threshold γ: 0.

4.1. Datasets

We use three groups of multimodal image datasets from the areas of remote sensing, medicine, and computer vision. Dataset 1 consists of 6 types of remote sensing image pairs, which are optical-optical images with different time differences, optical-SAR images, optical-map images, optical-infrared images, optical-depth images, and day-night images. Each type contains 10 image pair, taken from [8][1] and [73][2] Dataset 2 consists of 6 types of medical image pairs, which are Magnetic-Resonance-Image (MRI)–Single-Photon-Emission-Computed-Tomography (SPECT) images, Proton-Density-Weighted (PD)–T1-weighted (T1) images, PD–T2-weighted (T2) images, retina images acquired by different angiography techniques, SPECT–CT, and T1–T2 images. The type of retina-retina contains 23 image pairs, and the other types include ten image pairs. Specifically, the T1, T2, and PD images are given by [74][3]; the MRI, SPECT, and CT images are obtained from Harvard University[4]; the retina images are from [75]. Dataset 3 consists of 235 RGB–NIR image pairs with 9 scenes: country, field, forest, indoor, mountains, old buildings, street, urban, and water, which are taken from [76][5] We manually labeled 10–20 ground point pairs for each image pair and constructed the corresponding affine transformation matrix. These images differ in imaging mechanism, waveband, shooting time, usage, etc. Their detailed information is shown in Table 3.

Datasets 2 and 3 contain different strengths and types of nonlinear intensity NID and image translation (NT), which can be used to test the robustness against NID, while Dataset 1 contains NID, image translation, scale change, and image rotation (NTSR), simultaneously, which can be used to evaluate the performance under NID and complex geometric distortions. To enrich the experimental data, we rectify the image pairs of Dataset 1 with the transformation estimated based on manually selected matches to eliminate scale change and image rotation. The rectified image pairs only involve NT. At the same time, we randomly add different extents of scale change ranging from 1/2 to 2 and image rotation spanning 0–360° to Datasets 2 and 3. To be simple, we name the datasets with NT as DataSet1$_{NT}$, DataSet2$_{NT}$, DataSet3$_{NT}$, the datasets with NTSR as DataSet1$_{NTSR}$, DataSet2$_{NTSR}$, DataSet3$_{NTSR}$. Representative image pairs and their matching results can be found in Figs.11–13.

4.2. Parameter study

In this section, we test the performance of POS-GIFT under different parameter settings. Most of the parameters come from the construction of feature descriptors, which are the number of concentric circles (n_1), radius (P_1), the number of directions (n_2) per circle, and the number of nearest neighbors (K) in the POS strategy when constructing sampling points. For a fair comparison, we limit the maximum number of feature points to 5000 and only adjust the test parameter while keeping other parameters constant. Table 4 displays the detailed parameter settings and experimental results on 60 multimodal image pairs from DataSet1$_{NT}$.

As shown in Table 4, the different parameters have little effect on *RMSE*, which is maintained at around 1 pixel but has a significant impact on *NCM*. We varied n_1 from 2 to 5, the optimal matching performance was achieved at $n_1 = 3$. When the number of concentric circles exceeded 3, the *NCM* decreased significantly. For P_1, The maximum value of *NCM* was obtained when the radius of concentric circles P_1 was set to 6. Additionally, *NCM* gradually increased with the increase in the number

[1] https://skyearth.org/publication/project/CoFSM/
[2] https://skyearth.org/publication/project/HOWP/
[3] https://brainweb.bic.mni.mcgill.ca/brainweb/
[4] http://www.med.harvard.edu/aanlib/home.html
[5] https://ivrlwww.epfl.ch/supplementary_material/cvpr11/index.html

Table 3
The detailed information of the experimental datasets.

Dataset1	Image type	Optical	Optical infrared	Optical depth	Optical map	Optical SAR	Day night			
	Modal ID	1	2	3	4	5	6			
	Number	10	10	10	10	10	10			
	Size	500 × 271~992 × 602								
Dataset2	Image type	MRI SPECT	PD T1	PD T2	retina	SPECT CT	T1 T2			
	Modal ID	7	8	9	10	11	12			
	Number	10	10	10	23	10	10			
	Size	256 × 256~640 × 640								
Dataset3	Imge type	RGB-NIR								
	scene	Country	Filed	Forest	Indoor	Mountain	Old building	Street	Urban	Water
	Modal ID	13								
	Number	34	39	11	14	47	28	11	20	31
	Size	509 × 768~1024 × 754								

Fig. 11. The qualitative comparison results of (a) OS-SIFT, (b) LGHD, (c) PSO-SIFT, (d) RIFT, (e) LNIFT, (f) RedFeat, (g) MatchFormer, (h) SemLA, (i) GIFT, (j) POS-GIFT on six typical image pairs of DataSet1$_{NT}$ and six typical image pairs of DataSet1$_{NTSR}$. For each modality pair, the top row and the bottom row show the results of the images with NT and NTSR, respectively. The yellow and red line represents the correct and incorrect matches, respectively.

of directions n_2, possibly because a higher n_2 value resulted in a denser distribution of sampling points, enabling the feature descriptor to reflect more details. However, when n_2 exceeded 12, the increase in *NCM* was limited. For K, increasing the value of K can enlarge the searching range of the potential matched point, but may introduce incorrect matches. Conversely, decreasing the value of K will narrow the search range,

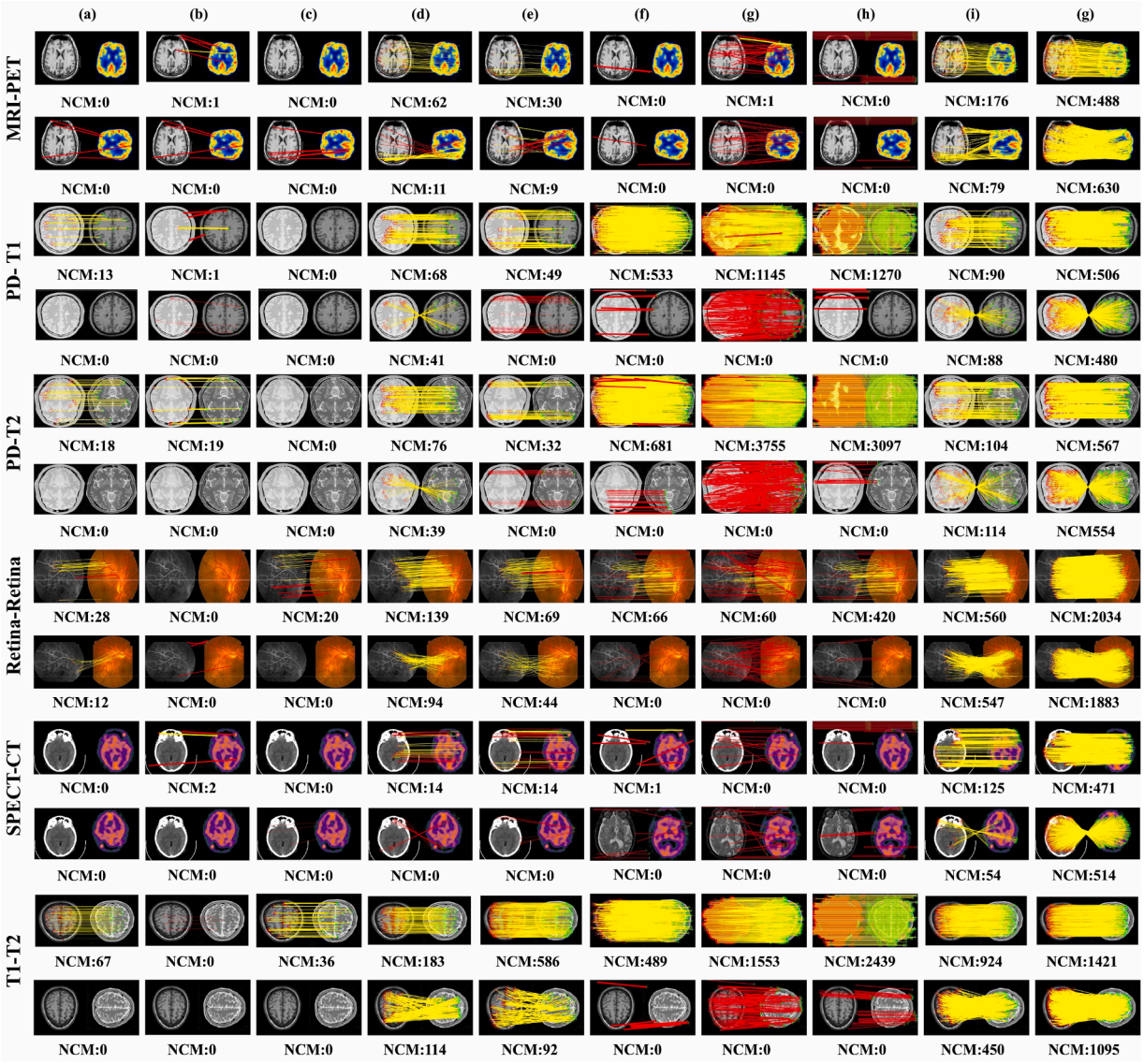

Fig. 12. The qualitative comparison results of (a) OS-SIFT, (b) LGHD, (c) PSO-SIFT, (d) RIFT, (e) LNIFT, (f) RedFeat, (g) MatchFormer, (h) SemLA, (i) GIFT, (j) POS-GIFT on six typical image pairs of DataSet2$_{NT}$ and six typical image pairs of DataSet2$_{NTSR}$. For each modality pair, the top row and the bottom row show the results of the images with NT and NTSR, respectively. The yellow and red line represents the correct and incorrect matches, respectively.

which can ensure high accuracy. However, a relatively precise initial matching position is required for the POS strategy. Experiments show that the largest NCM with high precision is achieved when K is set to 20. Therefore, based on the above analysis, we set the default parameters of POS-GIFT as $n_1 = 3$, $P_1 = 6$, $n_2 = 12$, and $K = 20$, and employed them in the following experiments.

4.3. Qualitative and quantitative comparison experiments

We first demonstrate the comparative visualization results (Figs. 11–13) accompanied by specific NCM of the ten methods on 30 typical multimodal image pairs, with 15 different modalities of image pairs with NT selected from DataSet1$_{NT}$, DataSet2$_{NT}$, and DataSet3$_{NT}$, and the corresponding 15 image pairs with NTSR selected from DataSet1$_{NTSR}$, DataSet2$_{NTSR}$, and DataSet3$_{NTSR}$. We can see that POS-GIFT and GIFT successfully matched all image pairs and obtained considerable NCM on all the image pairs, no matter whether the image pair involved NTSR or not. In addition, POS-GIFT outperforms GIFT. LGHD, PSO-SIFT, and OS-SIFT matched 9, 10, and 12 out of the 15 image pairs with NT but just matched 0, 4, and 6 out of 15 image pairs with NTSR, and the NCMs are small for the matched image pairs. This reveals that these three methods have some robustness against NID but are very sensitive to NTSR. RIFT and LNIFT present good robustness against NID, successfully matched all 15 image pairs with NT, and obtained larger NCM compared with LGHD, PSO-SIFT, and OS-SIFT. However, they are also subjected to NTSR, matching 5 and 10 pairs of the 15 image pairs involving large geometric changes, respectively. The three deep learning methods, RedFeat, Matchformer, and SemLA, successfully matched 13 out of the 15 image pairs with NT but failed to match the MRI-SPECT and SPECT-CT pairs, indicating that they can only tackle specific modalities differences. Moreover, their performance decreased dramatically on the image pairs with NTSR, only got a few matches on the image pairs with slight scale changes of SAR-optical and map-optical, and failed to match all the other image pairs. After all, POS-GIFT performs the best, successfully matching all 30 image pairs with substantially correct matches achieved.

Tables 5 and 6 present detailed matching results for DataSet1$_{NT}$, DataSet2$_{NT}$, DataSet3$_{NT}$, DataSet1$_{NTSR}$, DataSet2$_{NTSR}$, and DataSet3$_{NTSR}$. Generally, consistent results to the visualization results of selected image pairs are obtained.

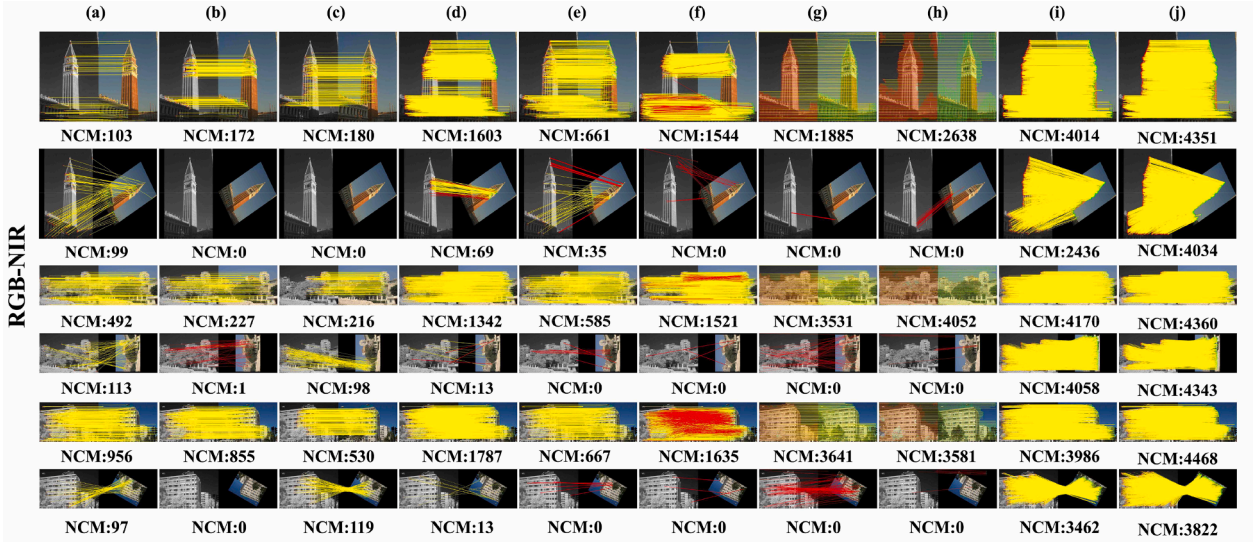

Fig. 13. The qualitative comparison results of (a) OS-SIFT, (b) LGHD, (c) PSO-SIFT, (d) RIFT, (e) LNIFT, (f) RedFeat, (g) MatchFormer, (h) SemLA, (i) GIFT, (j) POS-GIFT on three RGB-NIR image pairs of DataSet3$_{NT}$ and three RGB-NIR image pairs of DataSet3$_{NTSR}$. For each modality pair, the top row and the bottom row show the results of the images with NT and NTSR, respectively. The yellow and red line represents the correct and incorrect matches, respectively.

Table 4
The performance of POS-GIFT under different parameter settings.

	Parameter setting				*NCM*↑	*RMSE*↓
	n_1	P_1	n_2	K		
n_1	2	6	12	20	1638	1.15
	3	6	12	20	**1720**	**0.94**
	4	6	12	20	1586	0.94
	5	6	12	20	1295	0.93
P_1	3	4	12	20	1552	1.18
	3	5	12	20	1648	0.97
	3	**6**	12	20	**1720**	**0.94**
	3	7	12	20	1706	0.94
	3	8	12	20	1703	0.93
n_2	3	6	8	20	1666	1.13
	3	6	10	20	1714	1.02
	3	6	**12**	20	**1720**	**0.94**
	3	6	14	20	1725	0.93
	3	6	16	20	1727	0.93
K	3	6	12	16	1763	0.98
	3	6	12	18	1747	0.96
	3	6	12	**20**	**1720**	**0.94**
	3	6	12	22	1700	0.96
	3	6	12	24	1685	0.95

With respect to the results on the image pairs with NT, we can see that OS-SIFT and PSO-SIFT struggle to match the optical–SAR, day-night, MRI–SPECT, and SPECT–CT image modalities and obtains a small NCM on the other image modalities, showing limited resistance against complex NID. LGHD achieves a high SR of up to 90 % in the Optical–Optical, Optical–Infrared, Optical–Depth, Optical–SAR, and RGB–NIR image modalities, but the corresponding NCM was low, and the RMSE exceeds three pixels. Plus, it fails to match the Optical-Map, MRI–SPECT, PD–T1, Retina–Retina, SPECT–CT, and T1–T2 image modalities with complex NID. Through normalizing the multimodal images to reduce NID, LNIFT achieves a large SR above 90 % in the eight image modalities of Optical–Optical, Optical–Infrared, Optical–Depth, Optical-Map, PD–T1, PD–T2, Retina–Retina, T1–T2, and the metrics of NCM and RMSE of LNIFT are better than that of LGHD. RIFT outperforms LNIFT in SR in almost all modalities except for Optical-Map, thanks to the utilization of the radiometrically invariant maximum index map (MIM). The RMSE achieves around three to four pixels, even within two pixels on the RGB–NIR image pairs, demonstrating good radiometric invariance. However, the average NCMs of the Day–Night, MRI–SPECT, and SPECT–CT image modalities are 66, 43, and 11, which are relatively fewer. Compared to the handcrafted methods, RedFeat, MatchFormer, and SemLA successfully match most image modalities and achieve higher NCM and RMSE than RIFT but fail to produce any correct matches in the most challenging MRI–SPECT and SPECT–CT modalities. Additionally, there are a large number of incorrect matches in the matching results, with incorrect matching rates of 29 %, 35 %, and 27 % for RedFeat, MatchFormer, and SemLA. This suggests that deep learning methods have notable limitations, particularly when dealing with certain types of NID.

When NTSR is also involved in the multimodal image pairs, all comparative algorithms show varying degrees of performance degradation. Although OS-SIFT and PSO-SIFT consider scale change and rotation, the average SR of OS-SIFT in all modality pairs decreases from 58.1 to 23.4 while that of PSO-SIFT decreases from 65.5 to 35.1, and their average NCM decreases to 23 and 46, respectively. LNIFT adopts the orientation estimation approach of ORB, but its performance is still susceptible to rotation, with the average SR reduces from 85 % to 36.9 %, and the average NCM reduces from 195 to 60. Among them, it fails entirely in optical–SAR and SPECT–CT image modalities. Due to the lack of scale and orientation estimation modules, LGHD can only resist small-scale change and image rotation below 20°. It matches the four modality pairs of Optical–Infrared, Optical–Depth, Day–Night, and RGB–NIR, where the SRs are all less than 10%. When the rotation angle is greater than 30°, the performance of RedFeat, Matchformer, and SemLA decreases sharply. Specifically, the average NCM of the four methods reduces by more than 50 %, and their SRs decrease from more than 80 % to less than 30 % at all image modalities except for the PD-T2 and T1–T2, which drops to 40 %.

On the contrary, POS-GIFT is significantly superior to the other nine compared methods in all metrics on all the image datasets with or without scale and rotation distortion. For the multimodal image pairs with NT, POS-GIFT obtains large NCMs by several to tens of times than the other five handcrafted feature methods in all modalities. Moreover, POS-GIFT demonstrates remarkable superiority over deep learning matching methods, achieving absolute advantages in the challenging MRI–SPECT and SPECT–CT modalities and significantly improving NCM by 60–400 % in the six image modality pairs from the remote sensing area while increasing NCM by approximately 30 % in the RGB–NIR image modalities. Even GIFT excels with over 100 correct matches, outperforming the five comparative handcrafted feature matching

Table 5
The detailed matching results of ten algorithms on DataSet1$_{NT}$, DataSet2$_{NT}$, and DataSet3$_{NT}$.

Modality pair	Metric	OS-SIFT	LGHD	PSO-SIFT	LNIFT	RIFT	RedFeat	MatchFormer	SemLA	GIFT	POS-GIFT
Optical	NCM	33	256	102	111	226	312	1456	833	557	**2471**
Optical	RMSE	3.8	3.7	2.9	3.0	2.6	2.4	2.1	2.7	2.0	**0.9**
	SR(%)	70.0	90.0	60.0	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Optical	NCM	47	75	80	146	250	756	653	1232	830	**2374**
Infrared	RMSE	3.6	3.3	1.7	3.2	2.4	2.6	2.5	2.4	1.7	**0.8**
	SR(%)	90	90	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Optical	NCM	18	102	26	106	114	264	636	631	284	**2063**
Depth	RMSE	3.6	4.4	3.2	3.4	3.3	2.8	2.8	2.9	1.8	**0.9**
	SR(%)	90	**100**	60	90	**100**	**100**	**100**	**100**	**100**	**100**
Optical	NCM	30	13	76	130	119	373	483	568	1149	**2555**
Map	RMSE	3.9	4.9	1.7	2.8	2.6	2.6	2.7	3.0	2.1	**0.9**
	SR(%)	50	10	80	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Optical	NCM	15	82	13	69	87	110	292	382	257	**1886**
SAR	RMSE	4.2	3.5	4.7	4.9	4.6	2.9	3.0	2.9	2.3	**0.9**
	SR(%)	30	90	30	50	90	**100**	**100**	**100**	**100**	**100**
Day	NCM	19	71	26	31	66	56	682	202	210	**1303**
Night	RMSE	4.9	3.6	3.6	4.6	4.0	2.9	2.9	2.8	2.2	**1.0**
	SR(%)	10	80	40	50	60	80	80	80	**100**	**100**
MRI	NCM	0	11	0	27	43	0	0	0	149	**567**
SPECT	RMSE	5.0	4.3	5.0	4.9	4.4	5.0	5.0	5.0	3.2	**1.9**
	SR(%)	0	60	0	70	70	0	0	0	**100**	**100**
PD	NCM	17	0	105	274	421	1248	1000	1028	464	**1442**
T1	RMSE	4.1	5.0	1.4	2.9	2.6	2.7	2.1	2.4	2.6	**1.1**
	SR(%)	70	0	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
PD	NCM	47	18	265	617	749	1944	3459	3257	2033	**3894**
T2	RMSE	3.0	**1.0**	1.2	1.9	2.2	2.5	1.7	1.8	1.7	**1.0**
	SR(%)	**100**	70	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Retina	NCM	21	20	89	140	411	333	851	703	1461	**3059**
Retina	RMSE	4.6	2.9	2.1	3.3	2.5	2.7	2.8	2.9	1.8	**0.9**
	SR(%)	52	13	83	96	**100**	**100**	**100**	**100**	**100**	**100**
SPECT	NCM	0	0	0	14	11	0	0	0	140	**470**
CT	RMSE	5.0	5.0	5.0	4.6	3.9	5.0	5.0	5.0	2.5	**2.0**
	SR(%)	0	0	0	50	50	0	0	0	**100**	**100**
T1	NCM	136	0	209	447	688	2385	1604	2549	1234	**2672**
T2	RMSE	2.6	5.0	**0.9**	2.3	2.1	2.4	1.8	1.9	1.7	**0.9**
	SR(%)	**100**	0	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
RGB	NCM	177	202	268	430	1097	1084	2688	2685	3397	**3525**
NIR	RMSE	2.8	1.5	**0.8**	1.9	1.5	2.6	1.4	1.2	1.5	**0.8**
	SR(%)	93	93	98	**100**	**100**	**100**	**100**	**100**	**100**	**100**

methods and surpassing the deep learning methods in the Optical-Map, MRI–SPECT, PD–T2, and RGB–NIR image modalities. Besides, GIFT and POS-GIFT exhibit good rotation and scale invariance, effectively matching all image pairs even under complex geometric distortions. The average RMSE of GIFT is 2 pixels, while that of POS-GIFT reaches a very high accuracy of 1.2 pixels. Additionally, the NCM of POS-GIFT remains unchanged. The outstanding performance of POS-GIFT is attributed to multi-scale feature detection, robust and accurate primary orientation estimation, and the POS strategy. These large amounts of experiments indicate that POS-GIFT can be applied to various sources of multimodal images with different modalities.

4.4. Performance with respect to rotation

In this section, we demonstrate the robustness of our method for image rotation. In terms of image rotation, we select an optical-map image pair and a retina-retina image pair from the experimental datasets and manually modify the rotation difference between the images from 0 to 360°. Considering that the primary orientation method and the proposed POS strategy both affect the performance against rotation, we testify these two modules separately to fully evaluate the proposed method. Moreover, we introduce three other primary orientation estimation methods, GIFT-Gradient, GIFT-Centroid, and GIFT-Phase, as a comparison to prove the effectiveness of the proposed methods. To be fair, all the other steps are the same except for the primary orientation estimation method during experiments. The definitions of the comparative methods are as follows.

- GIFT-Gradient uses a gradient orientation histogram [30] to estimate the primary orientation. Similar to SIFT, we first calculate the gradient directions and magnitudes of pixels located at the neighborhood of the feature point. Then, a gradient histogram is constructed, and the direction bin with the largest value is taken as the primary orientation.
- GIFT-Centroid adopts the method of ORB [51] to estimate the primary orientation. Specifically, GIFT-Centroid computes a centroid point in a local area of the feature point on the PC map, and the direction of the line connecting the centroid point and the feature point center is taken as the primary orientation.
- GIFT-Phase uses the method of LHOPC [77] to estimate the primary orientation. The process can be represented by the following equations:

$$V = \sum_{s} O_{so}\cos(o) \tag{19}$$

$$H = \sum_{s} O_{so}\sin(o) \tag{20}$$

$$Ori_{Phase} = \tan^{-1}\frac{H}{V} \tag{21}$$

where O_{so} is the response of the Log-Gabor odd-symmetric filter with scale s and orientation o, and Ori_{Phase} represents the estimated phase direction.

Figs. 14–16 give the comparative quantitative results with different

Table 6
The detailed matching results of ten algorithms on DataSet1$_{NTSR}$, DataSet2$_{NTSR}$, and DataSet3$_{NTSR}$.

Modality pair	Metric	OS-SIFT	LGHD	PSO-SIFT	LNIFT	RIFT	RedFeat	MatchFormer	SemLA	GIFT	POS-GIFT
Optical	NCM	12	0	40	39	69	0	391	6	349	**2253**
	RMSE	4.2	5.0	1.7	4.2	4.9	5.0	3.0	2.9	1.8	**1.0**
	SR(%)	20.0	0.0	40.0	40.0	60.0	0.0	10.0	10.0	**100.0**	**100.0**
Optical Infrared	NCM	39	106	47	32	74	202	75	256	655	**2402**
	RMSE	2.7	4.5	1.8	4.9	4.2	2.8	3.4	2.8	1.8	**1.1**
	SR(%)	40.0	10.0	50.0	50.0	60.0	10.0	30.0	30.0	**100.0**	**100.0**
Optical Depth	NCM	15	33	14	30	78	209	450	192	264	**1897**
	RMSE	3.4	3.3	2.4	4.8	4.4	2.5	3.2	3.4	1.8	**1.0**
	SR(%)	20.0	10.0	30.0	40.0	30.0	10.0	30.0	20.0	**100.0**	**100.0**
Optical Map	NCM	12	0	29	22	36	200	175	260	1084	**2688**
	RMSE	4.2	5.0	1.6	4.2	4.5	1.6	2.3	2.2	1.7	**0.9**
	SR(%)	40.0	0.0	50.0	50.0	50.0	10.0	30.0	30.0	**100.0**	**100.0**
Optical SAR	NCM	8	0	7	0	15	26	88	46	235	**1676**
	RMSE	4.9	5.0	3.7	5.0	4.1	2.2	3.2	3.3	1.9	**1.0**
	SR(%)	10.0	0.0	10.0	0.0	70.0	10.0	30.0	20.0	**100.0**	**100.0**
Day Night	NCM	0	359	19	16	82	172	964	739	239	**1366**
	RMSE	5.0	2.3	2.3	3.3	3.9	2.3	2.3	2.4	1.9	**1.2**
	SR(%)	0.0	10.0	30.0	30.0	30.0	10.0	10.0	10.0	100.0	**100.0**
MRI SPECT	NCM	0	0	0	23	17	0	0	0	141	**608**
	RMSE	5.0	5.0	5.0	4.3	4.3	5.0	5.0	5.0	3.0	**1.9**
	SR(%)	0.0	0.0	0.0	10.0	60.0	0.0	0.0	0.0	**100.0**	**100.0**
PD T1	NCM	0	0	39	125	116	757	226	167	156	**1468**
	RMSE	5.0	5.0	2.3	3.4	3.9	2.6	2.8	2.1	1.9	**1.2**
	SR(%)	0.0	0.0	50.0	50.0	70.0	20.0	30.0	20.0	100.0	**100.0**
PD T2	NCM	19	0	121	137	176	466	398	237	1276	**3550**
	RMSE	2.9	5.0	1.0	3.3	4.2	2.1	2.9	1.8	1.7	**1.0**
	SR(%)	50.0	0.0	50.0	50.0	80.0	30.0	40.0	30.0	**100.0**	**100.0**
Retina	NCM	7	0	38	22	95	174	146	354	976	**3116**
	RMSE	4.2	5.0	2.3	4.2	4.3	2.8	3.0	2.7	1.8	**1.0**
	SR(%)	13.0	0.0	47.8	65.2	91.3	8.7	30.4	21.7	**100.0**	**100.0**
SPECT CT	NCM	0	0	0	0	0	0	0	0	141	**420**
	RMSE	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	2.8	**1.9**
	SR(%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**	**100.0**
T1 T2	NCM	40	0	94	127	226	792	251	360	852	**2693**
	RMSE	2.4	5.0	1.2	3.3	4.8	2.8	3.0	1.7	1.7	**0.9**
	SR(%)	50.0	0.0	60.0	60.0	60.0	30.0	40.0	20.0	**100.0**	**100.0**
RGB NIR	NCM	61	73	58	86	213	422	1218	914	1899	**3201**
	RMSE	2.3	3.0	1.2	2.7	2.4	2.4	2.3	2.8	1.7	**1.0**
	SR(%)	61.7	4.7	38.7	34.0	56.2	14.9	29.8	25.1	**100.0**	**100.0**

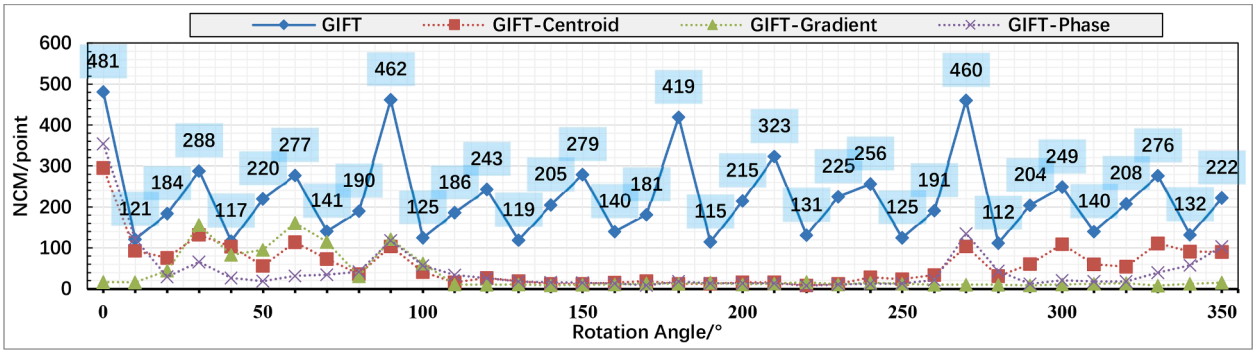

Fig. 14. The *NCM* changing curves of GIFT, GIFT-Centroid, GIFT-Gradient, GIFT-Phase on an optical-map image pair with different angles of image rotation.

primary orientation estimation methods, and Figs. 20–22 give the corresponding qualitative results of Figs. 14–16. From the results, we can see that similar patterns are obtained for the two image pairs. The performance of GIFT-Gradient, GIFT-Centroid, and GIFT-Phase decreases continuously with the increase of rotation angle, and it drops significantly when the angle difference is over 90°, resulting in matching failure. Among them, GIFT-Gradient performs the worst due to the sensitivity of nonlinear intensity difference to distortion. GIFT-Phase and GIFT-Centroid perform relatively better but struggle at large image rotation.

In comparison, GIFT not only obtains decent *NCM* but also maintains its performance without degradation as the rotation distortion increases, achieving rotation invariance at any angle. Additionally, we observed two periodic phenomena in the changing curve of the *NCM* of GIFT. Firstly, there is a significant periodicity of 90°, and this is because the manual rotation process of an image will change the number of effective pixels of the resampled sense image, resulting in a periodicity of 90°. Secondly, the small peak is at the integer multiples of 30°. The reason is that the number of orientations of the Log-Gabor filter is 6, which means that the angular difference between adjacent filters is 30°. When the rotation angle is a multiple of 30°, the multi-directional filtering features of the two images are most similar, therefore obtaining the largest *NCM*.

Figs. 17–22 give the comparative quantitative and qualitative results of POS-GIFT and GIFT. The above experiments have proven the

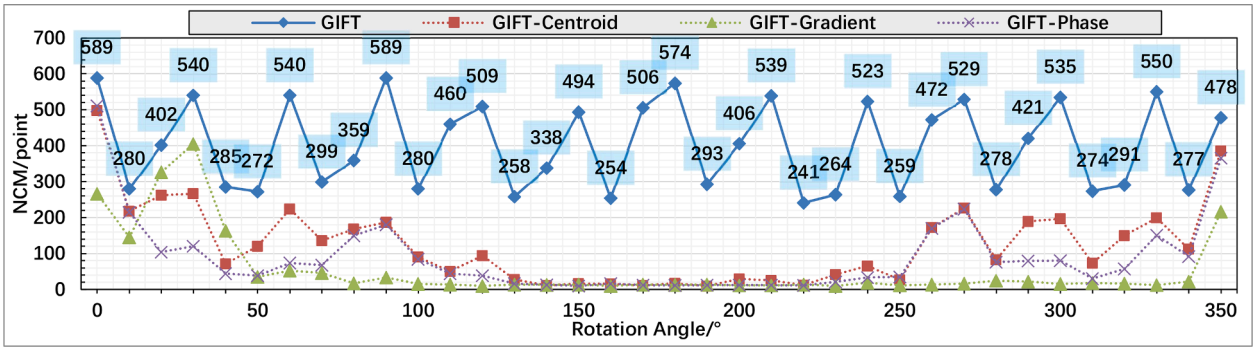

Fig. 15. The *NCM* changing curves of GIFT, GIFT-Centroid, GIFT-Gradient, and GIFT-Phase on a retina–retina image pair with different angles of image rotation.

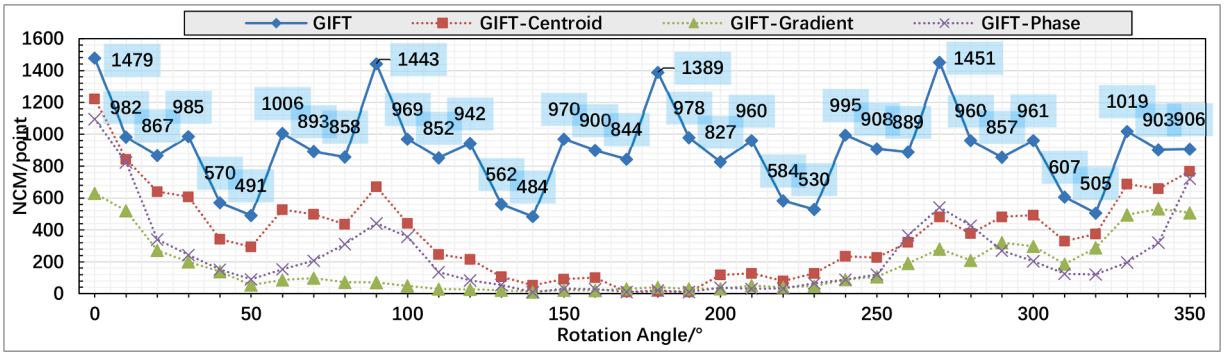

Fig. 16. The *NCM* changing curves of GIFT, GIFT-Centroid, GIFT-Gradient, and GIFT-Phase on an RGB–NIR image pair with different angles of image rotation.

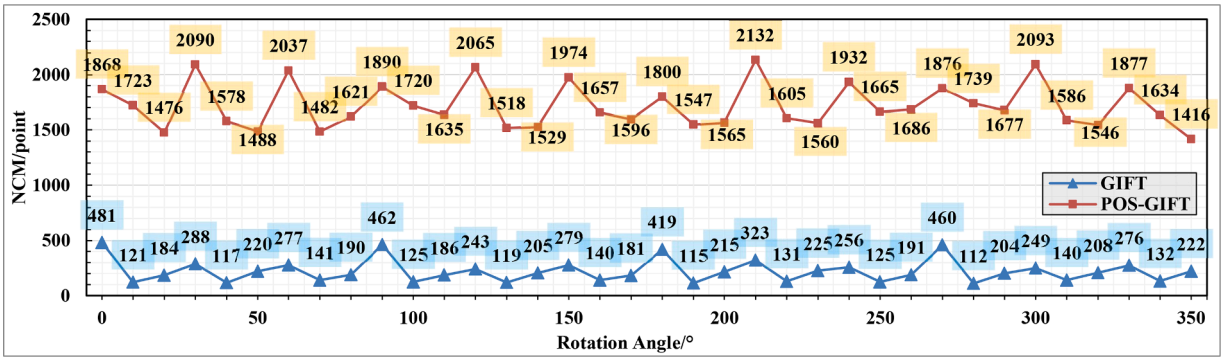

Fig. 17. The *NCM* changing curves of GIFT and POS-GIFT on an optical-map image pair with different angles of image rotation.

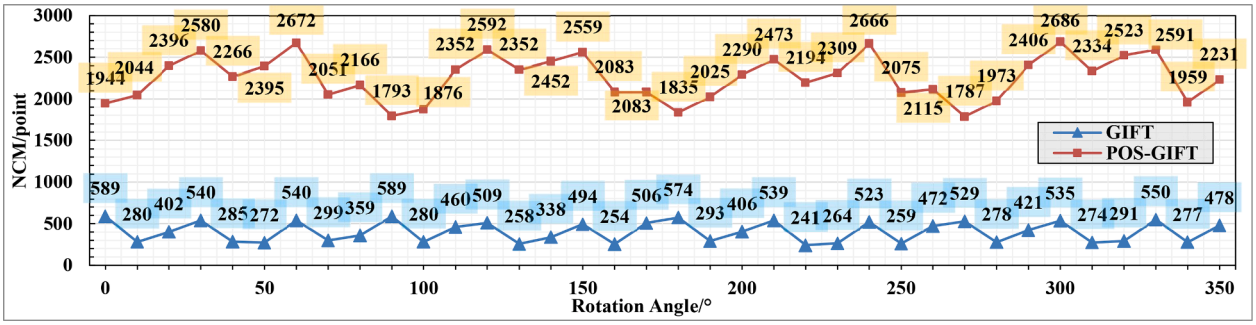

Fig. 18. The *NCM* changing curves of GIFT and POS-GIFT on a retina–retina image pair with different angles of image rotation.

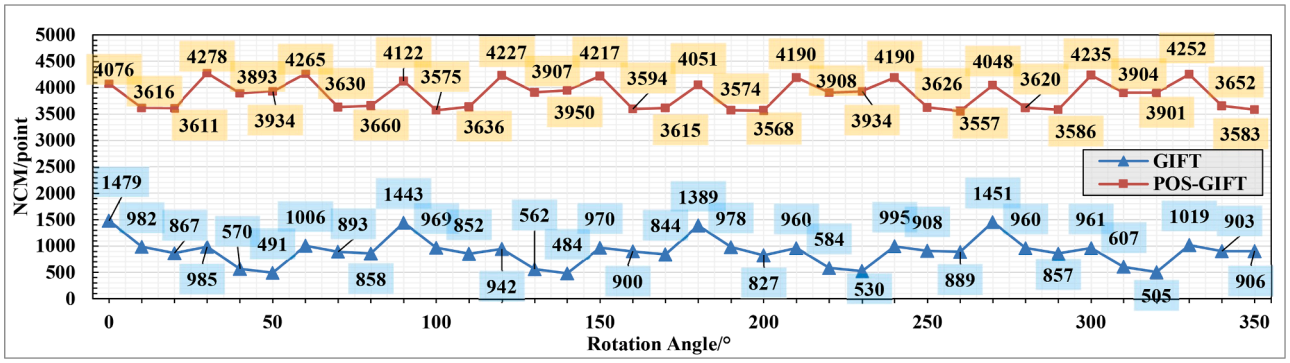

Fig. 19. The *NCM* changing curves of GIFT and POS-GIFT on an RGB–NIR image pair with different angles of image rotation.

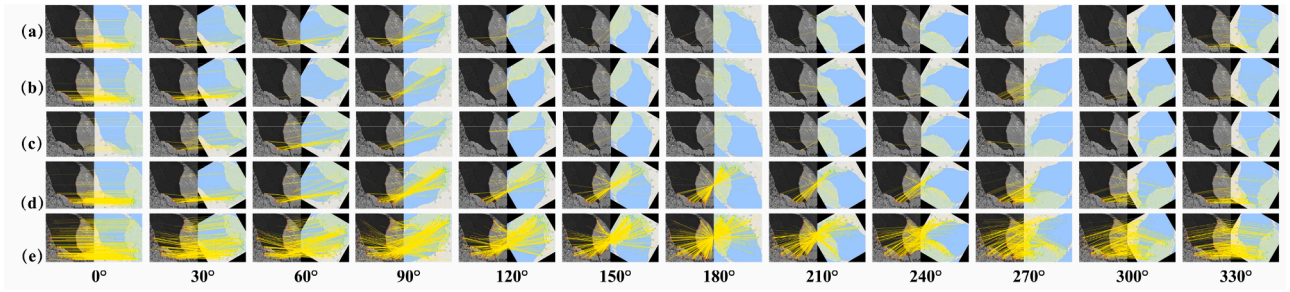

Fig. 20. The visualization results of Fig. 16 and Fig. 19. (a) GIFT-Centroid; (b) GIFT-Gradient; (c) GIFT-Phase; (d) GIFT;(e) POS-GIFT.

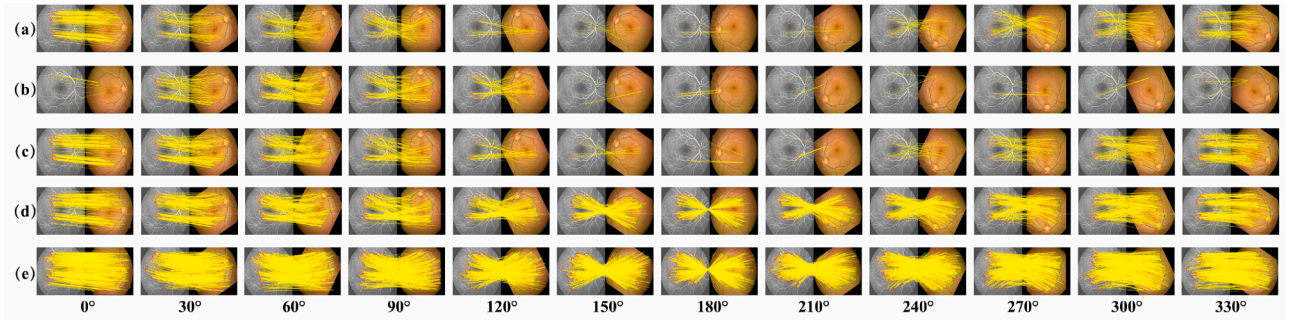

Fig. 21. The visualization results of Fig. 17 and Fig. 20. (a) GIFT-Centroid; (b) GIFT-Gradient; (c) GIFT-Phase; (d) GIFT;(e) POS-GIFT.

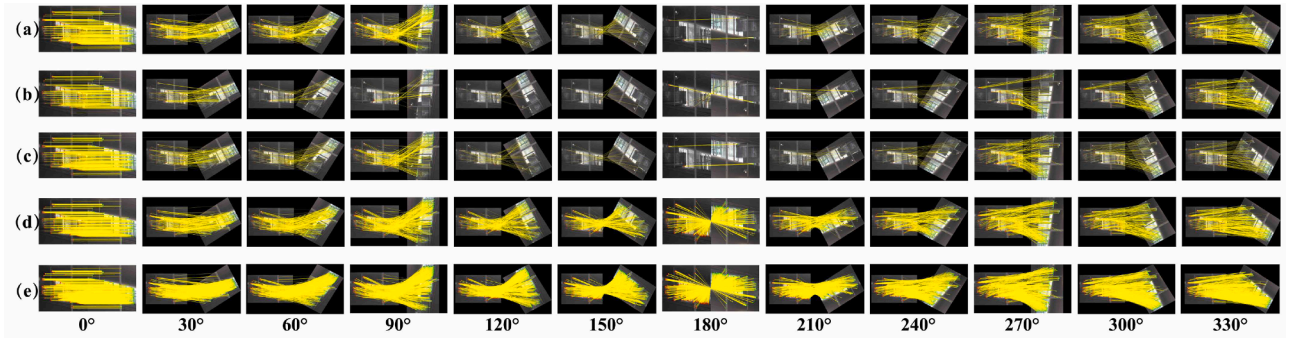

Fig. 22. The visualization results of Fig.18 and Fig. 21. (a) GIFT-Centroid; (b) GIFT-Gradient; (c) GIFT-Phase; (d) GIFT;(e) POS-GIFT.

effectiveness of the proposed primary orientation method. Here, we can see that both GIFT and POS-GIFT can handle any angle of image rotation, revealing that the POS strategy can work well on multimodal images with rotation. Similarly, we noticed the large periodicity of 90° and the "small peak" phenomenon of 30° in the ***NCM*** curve of POS-GIFT.

4.5. Running time

As shown in Table 7, we present the average running time of our proposed methods, POS-GIFT and GIFT, and the other five handcrafted methods, OS-SIFT, PSO-SIFT, LGHD, LNIFT, and RIFT, on DataSet1$_{NT}$

Table 7
The average running times of the seven algorithms on DataSet1$_{NT}$.

Method	OS-SIFT	PSO-SIFT	LGHD	LNIFT	RIFT	GIFT	POS-GIFT
Time/s	9.9	27.0	18.2	15.8	23.2	10.3	11.9

with 60 pairs of multimodal images. All the experiments were conducted with the same configuration: Windows 10 64 professional system, Intel (R) Xeon(R) CPU E5-1650 v4@ 3.60 GHz, 32 GB RAM, Matlab 2022a development environment.

We can see that OS-SIFT, GIFT, and POS-GIFT are at the same efficiency level, using the least time. PSO-SIFT cost the longest time. The time consumed by LGHD, LNIFT, and RIFT is moderate, while RIFT is significantly slower than the other two methods. The reason why OS-SIFT is fast is that it detects a relatively small number of feature points and is easy to fail, which will stop the time count. LNIFT consumes most of the time on rotating to direction gradient features to achieve rotation invariance. LGHD, RIFT, and GIFT all construct feature descriptors based on multi-scale and multi-orientation filtering results, but the different descriptor construction pattern leads to different efficiency. LGHD detects a MIM map on each scale of filtered image and builds the descriptor with histogram statistics, which are time-consuming. RIFT first added all scales of images and then constructed the feature descriptor on one MIM map. This strategy saves lots of time compared with LGHD. However, it employs an end-to-end annular feature matching approach, taking lots of time and making it slower than LGHD overall. On the contrary, GIFT utilizes an efficient Gaussian-weighted approach to build feature descriptors based on the LG feature, given that the filtering results have been generated in the feature detection process. Besides, the rotation invariance of GIFT is achieved by simply adjusting the elements order of the feature vector without complex rotation operations. These two schemes help GIFT be much faster than LGHD and RIFT. Additionally, POS-GIFT is only slightly slower than GIFT benefiting from the high efficiency of our POS strategy.

5. Discussion

In this section, we deeply analyze the performance of different methods from the perspective of algorithm design and principles. For the handcrafted methods, OS-SIFT was initially developed to match optical and SAR satellite images, introducing a new operator, ROEWA, improving the gradients calculation accuracy of SAR images. However, this design limits its application to specific multimodal images. PSO-SIFT proposes to use the second image derivative to reduce the modalities difference, But the second derivative is still not very good against NID. Besides, the PSO matching enhanced scheme it employed only has limited accuracy improvement. Unlike PSO-SIFT, LNIFIT first improves the similarity of the multimodal images through normalization operation and then conducts feature detection and description on the normalized images. The normalization process can help detect repeated feature points and increase the consistency of feature descriptors, but it can only decrease the effect of NID rather than eliminate it. Besides, it adopts the orientation estimation approach of ORB, which is sensitive to rotation. LGHD and RIFT construct feature descriptors based on the multi-scale and multi-orientation filtering results, increasing the feature description and discrimination ability towards NID. However, LGHD uses the FAST operator to detect feature points on the multimodal images, which are sensitive to NID and struggle to find common feature points, resulting in poor matching performance. Even worse, LGHD does not consider scale change and image rotation. Instead, RIFT detects feature points on the phase congruency, increasing the feature repeatability. Additionally, even if RIFT employs an end-to-end annular matching approach to tackle image rotation, it performs poorly under large rotation angles and still does not consider scale change.

For the learning-based methods, RedFeat introduces a mutual weighting strategy to appropriately couple the task of feature detection and description, improving the matching ability. Moreover, it employs a multi-scale feature extraction strategy to attain scale invariance; however, it exhibits heightened sensitivity to image rotation and perspective transforms. MatchFormer adopts a detector-free dense matching framework to enhance matching credibility in textureless regions, but its effectiveness is limited in handling complex NID and rotations. To mitigate the impact of modality differences, SemLA introduces the utilization of semantic information to constrain and guide the matching process. However, due to limitations in the training data, SemLA lacks semantic awareness for small objects, which prevents stable application in remote sensing images or images containing only small objects or objects devoid of semantic information. Remarkably, RedFeat, MatchFormer, and SemLA encounter total failure in matching the MRI–SPECT and SPECT–CT modalities, likely because of the ambiguous structure, significant NID, and the absence of suitable annotated datasets. These factors could hinder the effective application of deep learning in multimodal matching as well.

Our approach, POS-GIFT, absorbs the merits of LGHD and RIFT to handle NID by conducting feature detection and description on the multi-scale and multi-orientation log-Gabor filtering results. We also introduce a novel GFP feature, which concatenates the Gaussian weight features of evenly distributed sample points to capture the image structure, further increasing the resistance against modality difference. Then, we develop a robust primary orientation method by exploiting the characteristics of the GFP feature to make it strong to any rotation angle. However, similar structures and other unpredictable factors will still cause ambiguity and lead to incorrect matches even if the outlier removal algorithm, like RANSAC and FSC, is applied. Therefore, we propose the POS strategy, which refines the initial matching results and improves the matching accuracy using the positions-orientation-scale information. These innovations ensure the excellent performance of POS-GIFT on various multimodal images with severe NID and geometric distortions accounting for all evaluation metrics.

6. Conclusion

This paper proposes a novel multimodal matching method, POS-GIFT, which can robustly resist NID and geometric distortion (rotation, scale) in multimodal images. POS-GIFT is highly automatic, only containing a few parameters, and is not sensitive to them. Experiments prove the excellent resistance to image rotation. Furthermore, the Extensive qualitative and quantitative comparative results on various multimodal images reveal that POS-GIFT is superior to the state-of-the-art methods, PSO-SIFT, OS-SIFT, LGHD, RIFT, LNIFT, RedFeat, MatchFormer, and SemLA. The main contribution of this study lies in a novel feature descriptor invariant to image rotation and a novel inlier recovery.

Based on the characteristics of multimodal images, we designed a novel feature descriptor that captures the image structure, is invariant to image rotation at any angle, and can roughly work under severe NID. (1) We develop a point sampling pattern simulating the human vision system to capture image structure effectively; (2) We define the LG feature based on the multi-scale and multi-orientation Log-Gabor filter sequences, generate a feature for each sampled point by integrating the local LG values with a Gaussian weight, and further create a feature descriptor for the target point by concatenating the features of all sampled points in a predefined order. This way of constructing feature descriptor sufficiently apply surrounding information and capture image structure. Moreover, the descriptor can be easily adapted to image rotation by changing the order of the constructed features of sampled points with the estimated primary orientation; (3) Discarding the traditional primary orientation estimation methods, we propose a novel primary orientation estimation method under the proposed framework, which can work well with large NID. We figure out the changing pattern of the LG map and identify that the change comes from both the image

rotation and fixed filtering orientation. We further discover that the norm values computed from the all-orientation filter response remain unchanged. Accordingly, we first use the norm map to estimate the primary orientation to eliminate one aspect of change, then modify the order errors and rotation errors of LG features to achieve image rotation invariance.

Next, considering that our feature descriptor is based on the image structure, it is prone to produce false matches or even matching failures. We introduce a position-orientation-scale constraint inlier recovery strategy to eliminate the incorrect match points and build more correct correspondences. (1) Based on initial matches obtained from nearest neighbor matching, a coarse affine transformation between the images is estimated. The position of a certain point in the reference on the sensed image can be calculated. Even though the predicted position is not very accurate, the deviation will be significant. By constraining the searching area into the predicted position, the false matches produced by similar structures can be largely avoided. (2) By deforming the transformation matrix, we can get the rotation angle and scale change between the two images. Leveraging the calculated rotation angle, we reset the primary orientation of the feature points and rematch the feature points with the given primary orientation. By using the globally estimated primary orientation to substitute the locally estimated orientation, the estimation accuracy is improved, and the robustness against image rotation is enhanced. (3) Similarly, we unify the scale differences between images, improving the robustness against scale change.

In the future, we will try to improve the efficiency, accuracy, and robustness of the proposed method from the following aspects: (1) create a comprehensive multimodal image benchmark and further evaluate the potential of the proposed method in practical applications; (2) improve our method to make it adapt to non-rigid distortions in multimodal images so that it can be applied to broader applications; (3) exploit the potential of deep learning technique in multimodal image matching, for example, to apply it into the steps of feature detection or feature description, or even the complete process, to leverage the benefits brought by the rapid development of artificial intelligence.

CRediT authorship contribution statement

Zhuolu Hou: Writing – original draft, Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Yuxuan Liu:** Writing – original draft, Writing – review & editing, Resources, Methodology, Project administration, Funding acquisition. **Li Zhang:** Methodology, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to the code of our manuscript.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (42201494).

References

[1] Q. Zhu, Z. Wang, H. Hu, L. Xie, X. Ge, Y. Zhang, Leveraging photogrammetric mesh models for aerial-ground feature point matching toward integrated 3D reconstruction, ISPRS J. Photogramm. Remote Sens. 166 (2020) 26–40.
[2] C. Sun, X. Wu, J. Sun, N. Qiao, C. Sun, Multi-stage refinement feature matching using adaptive ORB features for robotic vision navigation, IEEE Sens. J. 22 (2022) 2603–2617.
[3] J. Wang, Q. Zhu, S. Liu, W. Wang, Robust line feature matching based on pair-wise geometric constraints and matching redundancy, ISPRS J. Photogramm. Remote Sens. 172 (2021) 41–58.
[4] Y. Jin, D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K.M. Yi, E. Trulls, Image matching across wide baselines: from paper to practice, Int. J. Comput. Vis. 129 (2021) 517–547.
[5] J. Chen, X. Chen, S. Chen, Y. Liu, Y. Rao, Y. Yang, H. Wang, D. Wu, Shape-former: bridging CNN and transformer via ShapeConv for multimodal image matching, Inf. Fusion 91 (2023) 445–457.
[6] S. Chen, S. Zhong, B. Xue, X. Li, L. Zhao, C.I. Chang, Iterative scale-invariant feature transform for remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 59 (2021) 3244–3265.
[7] J. Li, W. Xu, P. Shi, Y. Zhang, Q. Hu, LNIFT: locally normalized image for rotation invariant multimodal feature matching, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–14.
[8] Y. Zhang, Y. Yao, Y. Wan, W. Liu, W. Yang, Z. Zheng, R. Xiao, Histogram of the orientation of the weighted phase descriptor for multi-modal remote sensing image matching, ISPRS J. Photogramm. Remote Sens. 196 (2023) 1–15.
[9] T. Liao, N. Li, Single-perspective warps in natural image stitching, IEEE Trans. Image Proces. 29 (2020) 724–735.
[10] A. Asokan, J. Anitha, Change detection techniques for remote sensing applications: a survey, Earth Sci. Inform. 12 (2019) 143–160.
[11] D. Cao, B. Zhang, X. Zhang, L. Yin, X. Man, Optimization methods on dynamic monitoring of mineral reserves for open pit mine based on UAV oblique photogrammetry, Measurement 207 (2023), 112364.
[12] S. Ji, Z. Qin, J. Shan, M. Lu, Panoramic SLAM from a multiple fisheye camera rig, ISPRS J. Photogramm. Remote Sens. 159 (2020) 169–183.
[13] C. Campos, R. Elvira, J.J.G. Rodríguez, J.M.M. Montiel, J.D. Tardós, ORB-SLAM3: an accurate open-source library for visual, visual–inertial, and multimap SLAM, IEEE Trans. Robot. 37 (2021) 1874–1890.
[14] H. Yang, J. Yuan, Y. Gao, X. Sun, X. Zhang, UPLP-SLAM: unified point-line-plane feature fusion for RGB-D visual SLAM, Inf. Fusion 96 (2023) 51–65.
[15] E. Stenborg, C. Toft, L. Hammarstrand, Long-term visual localization using semantically segmented images, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 6484–6490.
[16] A. Moreau, T. Gilles, N. Piasco, D. Tsishkou, B. Stanciulescu, A. de La Fortelle, ImPosing: implicit pose encoding for efficient visual localization, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2892–2902.
[17] J. Markiewicz, K. Abratkiewicz, A. Gromek, W. Ostrowski, P. Samczyński, D. Gromek, Geometrical matching of SAR and optical images utilizing ASIFT features for SAR-based navigation aided systems, Sensors 19 (2019) 5500.
[18] X. Zhang, W. Sultani, S. Wshah, Cross-view image sequence geo-localization, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2914–2923.
[19] Q. Li, R. Cao, J. Zhu, H. Fu, B. Zhou, X. Fang, S. Jia, S. Zhang, K. Liu, Q. Li, Learn then match: a fast coarse-to-fine depth image-based indoor localization framework for dark environments via deep learning and keypoint-based geometry alignment, ISPRS J. Photogramm. Remote Sens. 195 (2023) 169–177.
[20] Z. Hu, Y. Hou, P. Tao, J. Shan, IMGTR: image-triangle based multi-view 3D reconstruction for urban scenes, ISPRS J. Photogramm. Remote Sens. 181 (2021) 191–204.
[21] L. Zhang, Y. Liu, Y. Sun, C. Lan, H. Ai, Z. Fan, A review of developments in the theory and technology of three-dimensional reconstruction in digital aerial photogrammetry, Cehui Xuebao/Acta Geod. Cartogr. Sin. 51 (2022) 1437–1457.
[22] P. Maken, A. Gupta, 2D-to-3D: a review for computational 3D image reconstruction from x-ray images, Arch.Comput. Methods Eng. 30 (2023) 85–114.
[23] J.Y. Ma, X.Y. Jiang, A.X. Fan, J.J. Jiang, J.C. Yan, Image matching from handcrafted to deep features: a survey, Int. J. Comput. Vis. 129 (2021) 23–79.
[24] Y. Ye, B. Zhu, T. Tang, C. Yang, Q. Xu, G. Zhang, A robust multimodal remote sensing image registration method and system using steerable filters with first- and second-order gradients, ISPRS J. Photogramm. Remote Sens. 188 (2022) 331–350.
[25] W. Zhaoxia, L. Yongxin, Z. Jie, F. Chenqing, Z. Hui, Interference image registration combined by enhanced scale-invariant feature transform characteristics and correlation coefficient, J. Appl. Remote Sens. 16 (2022), 026508.
[26] M. Pan, F. Zhang, Medical image registration based on Renyi's quadratic mutual information, IETE J. Res. 68 (2022) 4100–4108.
[27] X. Liu, S. Chen, L. Zhuo, J. Li, K. Huang, Multi-sensor image registration by combining local self-similarity matching and mutual information, Front. Earth Sci. 12 (2018) 779–790.
[28] Y. Ye, J. Shan, L. Bruzzone, L. Shen, Robust registration of multimodal remote sensing images based on structural similarity, IEEE Trans. Geosci. Remote Sens. 55 (2017) 2941–2958.
[29] Y.X. Ye, L. Bruzzone, J. Shan, F. Bovolo, Q. Zhu, Fast and robust matching for multimodal remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 57 (2019) 9059–9070.
[30] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[31] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vis. Image Underst. 110 (2008) 346–359.
[32] C. Gao, W. Li, R. Tao, Q. Du, MS-HLMO: multiscale histogram of local main orientation for remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–14.

[33] B. Zhu, C. Yang, J. Dai, J. Fan, Y. Qin, Y. Ye, R2FD2: fast and robust matching of multimodal remote sensing images via repeatable feature detector and rotation-invariant feature descriptor, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–15.

[34] C.A. Aguilera, A.D. Sappa, R. Toledo, LGHD: a feature descriptor for matching across non-linear intensity variations, in: 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 178–181.

[35] J. Li, Q. Hu, M. Ai, RIFT: multi-modal image matching based on radiation-variation insensitive feature transform, IEEE Trans. Image Process. 29 (2020) 3296–3310.

[36] S. Zhu, T. Yang, C. Chen, Revisiting street-to-aerial view image geo-localization and orientation estimation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 756–765.

[37] E. Honkavaara, H. Saari, J. Kaivosoja, I. Pölönen, T. Hakala, P. Litkey, J. Mäkynen, L. Pesonen, Processing and assessment of spectrometric, stereoscopic imagery collected using a lightweight UAV spectral camera for precision agriculture, Remote Sens. 5 (2013) 5006–5039.

[38] F. Rovira-Más, Q. Zhang, J.F. Reid, Stereo vision three-dimensional terrain maps for precision agriculture, Comput. Electron. Agric. 60 (2008) 133–143.

[39] H. Huang, T. Guo, T. Jiang, F. Cai, B. Niu, X. Han, Q. Zhang, Tightly coupled binocular vision-DVL fusion positioning feedback for real-time autonomous sea organism capture, IEEE Trans. Instrum. Meas. 72 (2023) 1–14.

[40] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li, S. Hao, A multiscale framework with unsupervised learning for remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.

[41] J. Sun, Z. Shen, Y. Wang, H. Bao, X. Zhou, LoFTR: detector-free local feature matching with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8922–8931.

[42] H. Xie, Y. Zhang, J. Qiu, X. Zhai, X. Liu, Y. Yang, S. Zhao, Y. Luo, J. Zhong, Semantics lead all: towards unified image registration and fusion from a semantic perspective, Inf. Fusion 98 (2023), 101835.

[43] L. Meng, J. Zhou, S. Liu, Z. Wang, X. Zhang, L. Ding, L. Shen, S. Wang, A robust registration method for UAV thermal infrared and visible images taken by dual-cameras, ISPRS J. Photogramm. Remote Sens. 192 (2022) 189–214.

[44] A.A. Cole-Rhodes, K.L. Johnson, J. LeMoigne, I. Zavorin, Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient, IEEE Trans. Image Process. 12 (2003) 1495–1511.

[45] A. Tashlinskii, G. Safina, R. Kovalenko, R. Ibragimov, Usage of mutual information as similarity measures for stochastic binding images, in: 2021 International Conference on Information Technology and Nanotechnology (ITNT), IEEE, 2021, pp. 1–6.

[46] V. Aggarwal, A. Gupta, Integrating morphological edge detection and mutual information for nonrigid registration of medical images, Curr. Med. Imaging 15 (2019) 292–300.

[47] H.m. Chen, M.K. Arora, P.K. Varshney, Mutual information-based image registration for remote sensing data, Int. J. Remote Sens. 24 (2003) 3701–3706.

[48] Z. Fan, L. Zhang, Y. Liu, Q. Wang, S. Zlatanova, Exploiting high geopositioning accuracy of SAR data to obtain accurate geometric orientation of optical satellite images, Remote Sens 13 (2021) 3535.

[49] L. Zhou, Y. Ye, T. Tang, K. Nan, Y. Qin, Robust matching for SAR and optical images using multiscale convolutional gradient features, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5.

[50] J.M. Morel, G.S. Yu, ASIFT: a new framework for fully affine invariant image comparison, SIAM J. Imaging Sci. 2 (2009) 438–469.

[51] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: an efficient alternative to SIFT or SURF, in: 2011 International Conference on Computer Vision, 2011, pp. 2564–2571.

[52] Y. Xiang, F. Wang, H. You, OS-SIFT: a robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas, IEEE Trans. Geosci. Remote Sens. 56 (2018) 3078–3090.

[53] X. Liu, J. Xue, X. Xu, Z. Lu, R. Liu, B. Zhao, Y. Li, Q. Miao, Robust multimodal remote sensing image registration based on local statistical frequency information, Remote Sens. 14 (2022) 1051.

[54] P.-E. Sarlin, D. DeTone, T. Malisiewicz, A. Rabinovich, Superglue: learning feature matching with graph neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4938–4947.

[55] P. Lindenberger, P.-E. Sarlin, M. Pollefeys, LightGlue: local feature matching at light speed, arXiv preprint arXiv:2306.13643, (2023).

[56] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, K.M. Yi, Cotr: correspondence transformer for matching across images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6207–6217.

[57] S. Tang, J. Zhang, S. Zhu, P. Tan, Quadtree attention for vision transformers, arXiv preprint arXiv:2201.02767, (2022).

[58] H. Chen, Z. Luo, L. Zhou, Y. Tian, M. Zhen, T. Fang, D. Mckinnon, Y. Tsin, L. Quan, Aspanformer: detector-free image matching with adaptive span transformer, in: European Conference on Computer Vision, Springer, 2022, pp. 20–36.

[59] Q. Wang, J. Zhang, K. Yang, K. Peng, R. Stiefelhagen, Matchformer: interleaving attention in transformers for feature matching, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 2746–2762.

[60] V. Balntas, K. Lenc, A. Vedaldi, K. Mikolajczyk, HPatches: a benchmark and evaluation of handcrafted and learned local descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5173–5182.

[61] A. Dai, A.X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Nießner, Scannet: richly-annotated 3d reconstructions of indoor scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5828–5839.

[62] Z. Li, N. Snavely, Megadepth: learning single-view depth prediction from internet photos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2041–2050.

[63] E.B. Baruch, Y. Keller, Joint detection and matching of feature points in multimodal images, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2021) 6585–6593.

[64] Y. Deng, J. Ma, ReDFeat: recoupling detection and description for multimodal feature learning, IEEE Trans. Image Process. 32 (2023) 591–602.

[65] Q. Yu, D. Ni, Y. Jiang, Y. Yan, J. An, T. Sun, Universal SAR and optical image registration via a novel SIFT framework based on nonlinear diffusion and a polar spatial-frequency descriptor, 2021.

[66] A. Sedaghat, H. Ebadi, Remote sensing image matching based on adaptive binning SIFT descriptor, IEEE Trans. Geosci. Remote Sens. 53 (2015) 5283–5293.

[67] P. Kovesi, Image features from phase congruency, Videre: J. Comput. Vis. Res. 1 (1999) 1–26.

[68] E. Rosten, R. Porter, T. Drummond, Faster and better: a machine learning approach to corner detection, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 105–119.

[69] L. Sawides, A. de Castro, S.A. Burns, The organization of the cone photoreceptor mosaic measured in the living human retina, Vis. Res. 132 (2017) 34–44.

[70] J.B. Jonas, U. Schneider, G.O.H. Naumann, Count and density of human retinal photoreceptors, Graefe's Arch. Clin. Exp. Ophthalmol. 230 (1992) 505–510.

[71] Y. Wu, W. Ma, M. Gong, L. Su, L. Jiao, A novel point-matching algorithm based on fast sample consensus for image registration, IEEE Geosci. Remote Sens. Lett. 12 (2015) 43–47.

[72] W. Ma, Z. Wen, Y. Wu, L. Jiao, M. Gong, Y. Zheng, L. Liu, Remote sensing image registration with modified SIFT and enhanced feature matching, IEEE Geosci. Remote Sens. Lett. 14 (2017) 3–7.

[73] Y. Yao, Y. Zhang, Y. Wan, X. Liu, X. Yan, J. Li, Multi-modal remote sensing image matching considering co-occurrence filter, IEEE Trans. Image Process. 31 (2022) 2584–2597.

[74] C.A. Cocosco, V. Kollokian, R.K.-S. Kwan, A.C. Evans, BrainWeb: online interface to a 3D MRI simulated brain database, Neuroimage 5 (1997) 425.

[75] J. Ma, J. Zhao, J. Jiang, H. Zhou, X. Guo, Locality preserving matching, Int. J. Comput. Vis. 127 (2019) 512–531.

[76] M. Brown, S. Süsstrunk, Multi-spectral SIFT for scene category recognition, in: CVPR 2011, 2011, pp. 177–184.

[77] Y. Ye, L. Shen, M. Chen, J. Wang, An Automatic Matching Method Based on Local Phase Feature Descriptor for Multi-source Remote Sensing Images, 42, Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of Wuhan University, 2017, pp. 1278–1284.