SFOC: A NOVEL MULTI-DIRECTIONAL AND MULTI-SCALE STRUCTURAL DESCRIPTOR FOR MULTIMODAL REMOTE SENSING IMAGE MATCHING

Bai Zhu, Jiachen Zhang, Tengfeng Tang, Yuanxin Ye*

Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 610031, China

Commission II, WG II/1

KEY WORDS: Multimodal images, Image matching, SFOC, Fast-NCCsfo, integral images

ABSTRACT:

Accurate matching of multimodal remote sensing (RS) images (e.g., optical, infrared, LiDAR, SAR, and rasterized maps) is still an ongoing challenge because of nonlinear radiometric differences (NRD) between these images. Considering that structural properties are preserved between multimodal images, this paper proposes a robust matching method based on multi-directional and multi-scale structural features, which consist of two critical steps. Firstly, a novel structural descriptor named the Steerable Filters of first- and second-Order Channels (SFOC) is constructed to address severe NRD, which combines the first- and second-order gradient information by using the steerable filters to depict multi-directional and multi-scale structural features of images. Meanwhile, SFOC is further enhanced by performing the dilated Gaussian convolutions with different dilated rates on it, which can capture multi-level context structural features and improve the ability to resist noise. Then, a fast similarity measure, called Fast Normalized Cross-Correlation (Fast-NCCsfo), is established to detect correspondences by a template matching scheme, which employs the Fast Fourier Transform (FFT) technique and the integral image to improve the matching efficiency. The performance of the proposed SFOC has been evaluated with many different kinds of multimodal RS images, and experimental results show its superior matching performance compared with the state-of-the-art methods.

1. INTRODUCTION

Image matching is a prerequisite step for remote sensing (RS) image processing and analysis applications, such as image fusion, change detection, and environmental monitoring. The key of RS image matching is to find an evenly distributed and high-precision set of control points (CPs) as much as possible. Generally, RS images can be directly georeferenced by employing the rigorous sensor models or the generic sensor model. However, the georeferencing of RS image is usually biased that caused by the inaccurate measurement of the satellite ephemeris and instrument calibration, which results in the georeferencing having an offset typically ranging from several pixels to dozens of pixels in the image space (Jiang et al., 2015).

Figure 1 exemplarily shows two pairs of multimodal images with direct georeferencing. (a) Google (Left) and GaoFen-2 Panchromatic (Right) images, (b) Google (Nether) and Sentinel-1 SAR (Upper) images.

Figure 1. Example of multimodal images with direct georeferencing. (a) Google (Left) and GaoFen-2 Panchromatic (Right) images, (b) Google (Nether) and Sentinel-1 SAR (Upper) images.

* Corresponding author: Yuanxin Ye, yeyuanxin@swjtu.edu.cn

This contribution has been peer-reviewed.

number of stable features from multimodal images with significant NRD. To tackle these problems, a growing number of valid descriptors have been designed based on structural and shape features. Given the advantages of phase congruency in image perception, numerous phase-congruency-based methods have been proven to improve the performance of multimodal matching (Ye et al., 2017; Li et al., 2019). Although these phase-congruency-based methods have shown the superiority of the phase congruency in resisting NRD, they required the amplitude and orientation of phase congruency, leading to the complicated calculation and time-consuming processes.

As deep learning has shown superior performance in image matching in the field of computer vision (Dusmanu et al., 2019), LBM has also been introduced into the RS image matching field (Wang et al., 2018b; Zhou et al., 2021). Although current LBM's have achieved remarkable progress, their disadvantages are also quite significant. The main drawback is that LBM usually requires a large amount of training and labeled data, which will greatly affect the practical application of image matching. Due to the number of neural network parameters being huge, the training efficiency is greatly related to the basic configuration of computer infrastructure. LBM’s superiority only is brought into play in multimodal image matching based on high-performance computer infrastructures, which is another disadvantage to limit its widespread use.

Recently, many descriptors based on multi-orientated gradient information to depict structural features have also proved to be robust to NRD, among which channel features of oriented gradients (CFOG) (Ye et al., 2019), angle-weighted oriented gradient (AWOG) (Fan et al., 2021), and multi-Scale and multi-Directional Features of odd Gabor (SDFG) (Zhu et al., 2021) are the most representative ones. Moreover, recent studies have shown that many local feature descriptors based on the first-order gradient information, such as SIFT and CFOG, are far from accurate in capturing visual features of human perception. Since the first- and second-order gradients are related to different geometric and structural features of images (Wallis and Georgeson, 2012), the second-order gradients have better performance in describing detailed information than the first-order gradients.

Although the CFOG, AWOG, and SDFG descriptors have been successfully used for multimodal image matching, the construction of gradient channels for CFOG is calculated by simple pixel differences, which are very sensitive to noises. While the horizontal and vertical gradients of AWOG are calculated by the Sobel operator that simply comprises the first-order Gradient information, such as SIFT and CFOG, are far from accurate in capturing visual features of human perception. Since the first- and second-order gradients are related to different geometric and structural features of images (Wallis and Georgeson, 2012), the second-order gradients have better performance in describing detailed information than the first-order gradients.

Although the CFOG, AWOG, and SDFG descriptors have been successfully used for multimodal image matching, the construction of gradient channels for CFOG is calculated by simple pixel differences, which are very sensitive to noises. While the horizontal and vertical gradients of AWOG are calculated by the Sobel operator that simply comprises the first-order Gradient information, such as SIFT and CFOG, are far from accurate in capturing visual features of human perception. Since the first- and second-order gradients are related to different geometric and structural features of images (Wallis and Georgeson, 2012), the second-order gradients have better performance in describing detailed information than the first-order gradients. The first-order steerable filters, which is utilized to address significant NRD of the gradient (AWOG) (Fan et al., 2021), and multi-scale and multi-directional features of odd Gabor (SDFG) (Zhu et al., 2021) are the most representative ones. Moreover, recent studies have shown that many local feature descriptors based on the first-order gradient information, such as SIFT and CFOG, are far from accurate in capturing visual features of human perception. Since the first- and second-order gradients are related to different geometric and structural features of images (Wallis and Georgeson, 2012), the second-order gradients have better performance in describing detailed information than the first-order gradients.

Although the CFOG, AWOG, and SDFG descriptors have been successfully used for multimodal image matching, the construction of gradient channels for CFOG is calculated by simple pixel differences, which are very sensitive to noises. While the horizontal and vertical gradients of AWOG are calculated by the Sobel operator that simply comprises the first-order Gradient information, such as SIFT and CFOG, are far from accurate in capturing visual features of human perception. Since the first- and second-order gradients are related to different geometric and structural features of images (Wallis and Georgeson, 2012), the second-order gradients have better performance in describing detailed information than the first-order gradients.

Although the CFOG, AWOG, and SDFG descriptors have been successfully used for multimodal image matching, the construction of gradient channels for CFOG is calculated by simple pixel differences, which are very sensitive to noises. While the horizontal and vertical gradients of AWOG are calculated by the Sobel operator that simply comprises the first-order Gradient information, such as SIFT and CFOG, are far from accurate in capturing visual features of human perception. Since the first- and second-order gradients are related to different geometric and structural features of images (Wallis and Georgeson, 2012), the second-order gradients have better performance in describing detailed information than the first-order gradients.

Although the CFOG, AWOG, and SDFG descriptors have been successfully used for multimodal image matching, the construction of gradient channels for CFOG is calculated by simple pixel differences, which are very sensitive to noises. While the horizontal and vertical gradients of AWOG are calculated by the Sobel operator that simply comprises the first-order Gradient information, such as SIFT and CFOG, are far from accurate in capturing visual features of human perception. Since the first- and second-order gradients are related to different geometric and structural features of images (Wallis and Georgeson, 2012), the second-order gradients have better performance in describing detailed information than the first-order gradients.

Although the CFOG, AWOG, and SDFG descriptors have been successfully used for multimodal image matching, the construction of gradient channels for CFOG is calculated by simple pixel differences, which are very sensitive to noises. While the horizontal and vertical gradients of AWOG are calculated by the Sobel operator that simply comprises the first-order Gradient information, such as SIFT and CFOG, are far from accurate in capturing visual features of human perception. Since the first- and second-order gradients are related to different geometric and structural features of images (Wallis and Georgeson, 2012), the second-order gradients have better performance in describing detailed information than the first-order gradients.
In addition to the first-order steerable $G_1$ filter, the second-order steerable $G_2$ filter is also used in subsequent descriptor construction. Similar to the steerable $G_1$ filter, the second-order Gaussian steerable filter $G_2$ is defined as follows:

\[
\begin{align*}
G_{2}^\sigma &= k_1(\theta)G_{1}^\sigma + k_2(\theta)G_{2}^\sigma + k_3(\theta)G_{2}^{2\sigma} \\
G_{1}^\sigma &= G_\mu = (-\frac{1}{2\pi\sigma^2})(1-\frac{x^2}{\sigma^2})e^{-\frac{x^2}{2\sigma^2}} \\
G_{2}^\sigma &= G_\mu = (-\frac{1}{4\pi^2\sigma^4})(1-\frac{x^2}{\sigma^2})e^{-\frac{x^2}{\sigma^2}} + (-\frac{x^2}{4\pi^2\sigma^6})(1-\frac{x^2}{\sigma^2})e^{-\frac{x^2}{\sigma^2}} \\
k_1(\theta) &= \frac{1}{3}[1 + 2\cos(2(\theta - \theta^o))] \quad \theta^o = 0^o, 60^o, 120^o
\end{align*}
\]

Where $\sigma$ represents the Gaussian standard deviation, and $*$ denotes convolution operation.

Furthermore, in order to enhance the detailed information of images, the second-order gradients based on the three basic filters (i.e., $G_{2,2}^{0}$, $G_{2,2}^{0}$, and $G_{2,2}^{120}$) are applied in the construction process of the second-order channels. Similarly, the second-order steerable channels of the image $I(x, y)$ at an arbitrary orientation $\theta$ can be computed by convoluting the image with $G_{2,2}^{0}$, $G_{2,2}^{0}$, and $G_{2,2}^{120}$, which is expressed as Eq. (7).

\[
\begin{align*}
S_{2,2}^{0} &= G_{2,2}^{0} * I(x, y) \\
S_{2,2}^{0} &= G_{2,2}^{120} * I(x, y) \\
S_{2,2}^{120} &= \cos(\theta)S_{2,2}^{0} + \sin(\theta)S_{2,2}^{120}
\end{align*}
\]

Once the synthetical first- and second-order steerable channels are constructed, the specified direction features at different scales are summed to obtain as much useful information as possible in each direction. Subsequently, these synthetical steerable channels in specified directions are convoluted by three parallel dilated Gaussian kernels, and the three parallel convolutional results are combined through one summation operation, which is designed to integrate a wealth of local inter-pixel information of images. The dilated Gaussian convolution with different dilated rates by inserting “holes” in the convolution kernels to expand its receptive field, which is inspired by the recent deep convolutional neural networks (Chen et al., 2017). In addition, the dilation rates $r$ are set to $[1, 2, 3]$ for avoiding the inherent “gridding” problem that exists in the current dilated

![Figure 2. Construction flowchart of the proposed SFOC descriptor.](image-url)
convolution framework (Wang et al., 2018a). By this means, the multi-level context structural features of the synthetical first- and second-order steerable channels can be captured by utilizing dilated Gaussian weighting without increasing the computational complexity, and play a role in smoothing noise as well.

Figure 3 clearly illustrates the advantages of utilizing the dilated Gaussian convolution for the construction of SFOC. Four different types of heatmaps concerning different features are acquired by performing template matching. It is obvious that the heatmap of the original image pairs is the messiest, and the heatmap of the SFOC features without Gaussian convolution has several peaks but the peak is not distinct, because it’s greatly affected by significant noise. In contrast, Gaussian convolution can effectively resist the interference of noise and make the peak more discriminative (see Figure 3 (e) and (f)). Furthermore, the dilated Gaussian convolution can not only smooth the noise, but also integrate the multi-level context structural features by the dilated Gaussian weighting. This is the reason why the heatmap of the SFOC features with parallel Dilated Gaussian convolution presents a smoother and more discriminative peak than the general Gaussian convolution, which indicates the matching robustness of SFOC with parallel Dilated Gaussian convolution may be superior.

In particular, compared with the synthetical first-order steerable channels, a larger is used for the dilated Gaussian smoothing for the synthetical second-order steerable channels. This is because the second-order gradient describes more image detail but is accompanied by an increase in noises. Subsequently, the first- and second-order steerable channels are normalized respectively, then the final feature representation of SFOC is obtained by stacking them.

2.3 Establishment of fast similarity measure

The traditional normalized cross-correlation (NCC) is widely applied to determine corresponding CPs between the given image pairs with overlapping regions by evaluating the intensity similarity. However, it is often used only for CP detection of single-modal images and is often unable to keep the same performance for multimodal image matching. As mentioned above, the SFOC descriptor can capture the structural features of images, which effectively resists NRD between multimodal images. Accordingly, it makes sense to establish a novel similarity measure that combines NCC with the SFOC descriptor.

SFOC is a 3D descriptor with a large amount of data, as well as the NCC also has the disadvantage of large calculation amount and high computational complexity. Hence, in order to maintain the matching accuracy and improve the computational efficiency, a fast-matching similarity measure is designed based on NCC and SFOC, it’s expressed as Fast-NCCSFOC. The proposed Fast-NCCSFOC can be reformulated with more detail as follows in this subsection.

First of all, the SFOC descriptor is used to calculate the structural features in the template image and the search image, which are denoted by T and S, respectively. Their normalized correlation value NCCSFOC (S, T) represents the similarity of the template window T(i, j, z) and the search window S(x, y, z) at the location (x, y), which is defined as:

\[
NCC_{SFOC}(S,T) = \frac{\sum_{i} \sum_{j} \sum_{z} (S(x+i,y+j,z) - \bar{S})(T(i,j,z) - \bar{T})}{\sqrt{\sum_{i} \sum_{j} \sum_{z} (S(x+i,y+j,z) - \bar{S})^2 \sum_{i} \sum_{j} \sum_{z} (T(i,j,z) - \bar{T})^2}}
\]  

(8)

Where \( z \) presents the dimension of the SFOC descriptor. \( S(i, j, z) \) and \( T(i, j, z) \) are the feature value of the search window and the template window at the position \((i, j, z)\), respectively. The sizes of the template and search window are \( m \times n \times z \) pixels and \( M \times N \times z \) pixels, respectively. \( \bar{S} \) represents the average feature value of the template image, and \( \bar{T} \) represents the average feature value of the search image \( S \) under the current template image \( T \).

The reason for the high computational complexity of traditional correlation matching is that NCC is completely recalculated for any search position \((x, y)\), while the internal relation of the NCC of adjacent search points is ignored. In order to reduce the computational complexity, an equivalent transformation is performed on Eq. (8), as follows:

\[
NCC_{SFOC}(S,T) = \sqrt{\frac{R_{\text{xy}}(x,y,z) - R_{\text{xy}}(x,y,z)/\text{mnz}}{R_{\text{xy}}(x,y,z) - R_{\text{xy}}(x,y,z)/\text{mnz} + R_{\text{zy}}(i,j,z) - R_{\text{zy}}(i,j,z)/\text{mnz}}}
\]  

(9)

There are only three items related to \((x, y, z)\) included in the above formula, which is respectively denoted as:

\[
R_{\text{xy}}(x,y,z) = \sum_{i=1}^{M} \sum_{j=1}^{N} S(x+i,y+j,z)T(i,j,z)
\]  

(10)

\[
R_{\text{xy}}(x,y,z) = \sum_{i=1}^{M} \sum_{j=1}^{N} S(x+i,y+j,z)
\]  

(11)

\[
R_{\text{zy}}(i,j,z) = \sum_{x=1}^{M} \sum_{y=1}^{N} T(i,j,z)
\]  

(12)

It should be noticed that the first term \( R_{\text{xy}}(x,y,z) \) in the numerator is convolution operation, and the convolution in the spatial domain is equivalent to the dot production operation in the frequency domain. Therefore, it can be converted to the frequency domain, and the FFT technique is used to improve computational efficiency. Accordingly, the new expression of the term is equivalent to the following form:

\[
R_{\text{xy}}(x,y,z) = \left| \mathcal{F}(S) \right|^2 \mathcal{F}(T)
\]  

(13)
Where $\mathcal{F}$ is the signal of the Fourier transform, $\mathcal{F}^*$ represents the conjugate complex operation of the transformed result, and $\mathcal{F}^{-1}$ denotes the inverse FFT (i.e., IFFT).

Additionally, the terms in the denominator and the other terms in the numerator of Eq. (9) require a lot of multiplications and additions. When the template is sliding, the sum of the squares and correlation values are recalculated, which results in computation time increased enormously. It can be seen that these terms, $R_s$ and $R_{ss}$, fit the definition of the integral image (Viola and Jones, 2001). While the other two terms, $R_t$ and $R_{ts}$, are only related to the template image, which results in their values being fixed. Therefore, the integral image is used to replace the original summation process with three simple addition and subtraction operations, which can effectively reduce the computational complexity of the original algorithm to calculate NCC, and improve the running time.

As a result, these terms, $R_s$ and $R_{ss}$, can be efficiently calculated utilizing the integral image. Since the integral process only involves a limited number of additional operations, the complexity of the algorithm is mainly determined by FFT and IFFT in Eq.(13). Typical FFT and IFFT require about $2MN\log_2(MN)$ times of multiplication, and the Eq. (13) requires to calculate FFT and IFFT once in total. Accordingly, the total number of multiplications required by the proposed Fast-NCC_SFOC is as follows.

$$T_i \approx 4MNz\log_2(MN) \quad (14)$$

With regard to the template matching strategy, the Eq. (8) is directly used to calculate NCC at each sliding position, and the calculation amount mainly depends on the dominant times of multiplication operation. For any search position, the Eq. (8) is used to calculate NCC for about three times of multiplication, and a total of $(M+m+1) \times (N+n+1)$ slidable positions need to be calculated for traversal search in the search window space. Thus, the number of multiplication operations required for NCC matching is:

$$T_c = 3mnz(M - m + 1)(N - n + 1) \quad (15)$$

From the Eqs. (14) and (15), we can see that the computational complexity of the proposed Fast-NCC_SFOC is independent of the template size, whereas the computational complexity of NCC is approximately proportional to the product of the template size and the search size, especially when $m$ and $n$ are small relative to $M$ and $N$. The ratio of the computational complexity of the two similarity measures is:

$$T = \frac{T_i}{T_c} \approx \frac{4MNz\log_2(MN)}{3mnz(M - m + 1)(N - n + 1)} \quad (16)$$

To facilitate the illustration of the computational advantage of Fast-NCC_SFOC, we assume that $M=N$, $m=n$, and $M=2m$. The curve of $T$ changing with $m$ is shown in Figure 4. As the template and search sizes increase, the ratio of the computational complexity between Fast-NCC_SFOC and NCC decreases rapidly, that is, the larger the template and search sizes are, the greater the computational advantage of Fast-NCC_SFOC is. Taking a template window $m=100$ as an example accompanied by a search window $M=200$, Fast-NCC_SFOC takes about 0.799% of the time required by NCC, which greatly improves the computational efficiency.

In this section, the performance of the proposed SFOC was experimentally evaluated with different types of multimodal RS datasets (e.g., optical, infrared, LiDAR, SAR, and rasterized maps). Firstly, the experimental settings were presented, which include the detailed information of test datasets, the evaluation criteria, the implementation details, and the parameters predefined. Then, SFOC was compared with the five state-of-the-art methods for verifying its effectiveness, including MI, matching by tone mapping (MTM) (Hel-Or et al., 2013), phase congruency structural descriptor (PCSD) (Fan et al., 2018), CFOG, and SDFG. Finally, we analysed the robustness of SFOC against Gaussian white noise and speckle noise.

### 3. EXPERIMENTS

Eight cases of multimodal image pairs with significant NRD were employed to evaluate the performance of SFOC. The detailed information of these cases is given in Table 1, and these image pairs of each case are displayed in Figure 6. In addition, the two images of each case have been pre-registered with the same ground sample distance (GSD) to remove obvious rotation and scale differences.

#### 3.1 Experimental settings

In the experiments, the block-based FAST operator was first employed to extract 200 uniformly distributed IPs from the reference image. Then, the CP detection was performed using different methods with the same template size of $80 \times 80$ pixels. Furthermore, four criteria were used to quantitatively evaluate the matching performance in terms of the number of correct matches (NCM), the correct matching ratio (CMR), the root-mean-square errors (RMSE), and the matching time (MT). The
Among the six methods used for comparison, the worse matching was determined by manually selecting 50 evenly distributed CPs to estimate the projective model for the image pairs of each case. The projective model was used to calculate the location errors of the matches obtained by different methods, and the match within positioning errors of 1.5 pixels was defined as the correct CP. CMR was defined as CMR = NCM / total matches, where total matches refer to all matched CPs, including the outliers with large errors.

To make a fair comparison, MI was calculated using a histogram with 32 bins, as this is usually accompanied by an optimal matching performance (Ye et al., 2019). And the parameters of the other comparative methods (i.e., MTM, PCSD, CFOG, and SDFG) used the best parameters recommended in their related papers. All experiments were performed using a personal computer (PC) with the configuration of Intel (R) Core (TM) CPU i7-10750H 2.6GHz and 16GB RAM.

3.2 Comparison and analysis of matching performance

In this section, the performance of the proposed method was quantitatively and qualitatively evaluated. Moreover, to evaluate the effectiveness of the second-order gradient in the generation process of SFOC, the SFOC descriptor was degraded by only using the first-order steerable channels without the second-order steerable channels. The degraded SFOC descriptor was represented by F-SFOC, and it was also used for matching performance comparison with other methods.

The seven different methods, i.e., MI, MTM, PCSD, CFOG, SDFG, F-SFOC, and SFOC, were applied to eight multimodal image cases (Table 1) for the comparison of matching performance. Figure 5 depicts the comparison results of all the evaluation criteria (i.e., NCM, CMR, RMSE, and MT) for the different methods on each multimodal image pair. It is obvious that SFOC outperformed the other methods for the above four criteria in all test cases, which effectively demonstrates the superiority and robustness of the proposed SFOC.

![Figure 5. Performance comparison of different methods on the eight multimodal image cases with the template size of 80 × 80 pixels. (a) NCM. (b) CMR. (c) RMSE. (d) MT.](image)

Among the six methods used for comparison, the worst matching performance was found in the MI and MTM. MI and MTM had comparable matching performance, but MTM performed slightly better than MI on two Optical-to-Infrared cases, while MI performed better than MTM on cases 3-8. This may be related to the fact that MTM only utilizes a piecewise linear function to fit the intensity changes widely existing in the multimodal images. However, the intensity relationship between optical and SAR (or LiDAR) images is too complex to be fitted by MTM, which results in its performance degradation. Although the performance of MI was slightly better than that of MTM, it was the most time-consuming among all the methods because it requires calculating a large number of joint probability histograms.

From the comparison results in Figure 5, we can also observe that PCSD, CFOG, and SDFG performed significantly better than MI and MTM, while SDFG had slightly better performance compared with CFOG and PCSD. The main reason is that PCSD is constructed by using the multi-scale phase congruency structural features, and CFOG is built making use of the dense channel features of orientated gradients, which is more robust to NRD than MI and MTM. When comparing PCSD with CFOG, its performance was slightly worse than that of CFOG. The reason for that is the PCSD may lose some detailed structural information because it employed the strategy of the phase congruency order-based region division for descriptor construction. As for SDFG, since it further increasingly adopted the multi-scale strategy on the basis of multi-direction using odd Gabor functions, its matching performance was more robust than CFOG, but the matching process was more time-consuming. In addition, the construction of PCSD relies on multi-scale phase congruency features, which results in it being time-consuming. Therefore, PCSD and MTM were the most time-consuming apart from MI in all the compared methods.

For our degraded descriptor (i.e., F-SFOC), its matching performance was comparable to SDFG, and it yielded better results than CFOG on the criterion of RMSE, especially in the LiDAR-to-Optical and Optical-to-SAR cases. This phenomenon illustrates that the first-order Gaussian steerable filters and the dilated Gaussian convolution are effective to construct the descriptor. While the matching performance of F-SFOC was obviously lower than SFOC, which verified the feasibility and effectiveness of adding the second-order gradient in the generation of SFOC. In this way, the robustness and discriminability of SFOC can be effectively increased. As far as the MT, F-SFOC was slightly faster than CFOG, because it only took advantage of the first-order steerable channels without the second-order steerable channels resulting in a smaller dimensionality of its features than that of CFOG. Whereas SFOC required slightly more time-consuming than CFOG and SDFG, this is related to the multi-scale strategy with different Gaussian STD and the dilated Gaussian convolution with different dilated rates were embedded in the generation process of SFOC. Hence, considering the improvement of the matching performance for SFOC, it is acceptable to sacrifice a little running time.

Moreover, qualitative evaluation was performed by displaying the correct matched CPs for the visual inspection. As shown in Figure 6, these CPs were established by SFOC on the image pairs of each case with a template size of 80 × 80 pixels. It is obvious that these obtained CPs on the image pairs of each case were evenly distributed, and the location accuracy of these CPs was reliable despite significant NRD and noise between these multimodal image pairs.

3.3 Comparison and analysis of noise sensitivity

In this section, the anti-noise performance of the above-mentioned methods was evaluated and analyzed by adding...
different levels of Gaussian white noise and speckle noise to the images, respectively. Because the NRD between multimodal images is difficult to be precisely fitted only by a simple mathematical model. Meanwhile, LiDAR and SAR images typically contain more noise than infrared images, which is not conducive to the assessment of noise sensitivity. Consequently, all the methods were performed with the template size of 80 × 80 pixels for the selected four pairs of Optical-to-Infrared cases, and their average value of CMR was used for the subsequent analysis. Specifically, two types of series noisy images were generated by adding the different levels of Gaussian white noise with mean 0 and variance $\nu$ in the range $[0, 0.01]$ with an interval of 0.001, and the different levels of speckle-noise with variances $\nu$ in the range $[0, 0.1]$ with an interval of 0.01, respectively. Figure 7 presents the average CMRs of different methods versus various speckle noise. (a) Average CMRs of different methods versus various Gaussian white noise. (b) Average CMRs of different methods versus various speckle noise.

Figure 7. Average CMRs of different methods versus various noise. (a) Average CMRs of different methods versus various Gaussian white noise. (b) Average CMRs of different methods versus various speckle noise.

The above results and coherence analysis demonstrate that SFOC has apparent effectiveness and advantages for resisting significant NRD and noise between multimodal images, as well as high computational efficiency. The good adaptive performance was mainly due to the following reasons. On the one hand, it not only employed the first-order steerable filters with the multi-scale strategy to depict the multi-directional and multi-scale structural features between multimodal images, but it also utilized the second-order steerable filters and three parallel dilated Gaussian kernels to emphasize more detailed information and multi-level context structural features, respectively, which further improves the discriminative and anti-noise capability of the proposed method. On the other hand, the improved Fast-NCCSFOC based on the FFT and integral image technique ensured its fast computational efficiency.
4. CONCLUSIONS

This paper presented a robust matching method of multimodal RS images, involving both a novel SFoC descriptor and a fast similarity measure (i.e., Fast-NCCSFOC). SFoC is first proposed by making use of the first- and second-order Gaussian steerable filters, which aims to capture distinctive multi-directional and multi-scale structural features for resisting significant NRD between multimodal images. Then Fast-NCCSFOC is established by combining NCC and SFoC, and it speeds up the image matching process by using the FFT technique and the integral image. The experimental results on eight various multimodal images have demonstrated the robustness and effectiveness of SFoC. In contrast to other state-of-the-art methods (i.e., MI, MTM, PCSD, CFOG, SDFG), the proposed SFoC achieved the best matching performance in the quantitative evaluation.

Although SFoC presented robust performance for multimodal image matching, it is sensitive to global geometric distortions between images, that is, it cannot be adapted to multimodal image matching with large scale or rotation differences. Future research aims to design an enhanced descriptor that is adaptable to geometric distortions without the assist of geo-referenced information, and explore the matching technique with scale and rotation invariance.

ACKNOWLEDGEMENTS

This paper is supported by the National Natural Science Foundation of China (No.41971281).

REFERENCES


