Robust Multimodal Image Matching Based on Radiation Invariant Phase Correlation

Tao PENG[1], Liang ZHOU[1], Guangyang LEI[1], Peizhen YANG[1], Yuanxin YE[1, *]

[1] Faculty of Geosciences and Engineering, Southwest Jiaotong University, 611756 Chengdu, China
(ptao0824, zhouliangmale)@163.com, LGY0520@my.swjtu.edu.cn, ybacon999@gmail.com, yeyuanxin@home.swjtu.edu.cn

Keywords: Multimodal image matching, Phase correlation, Multimodal Local Self-correlation, Log-polar coordinates, RIPC.

Abstract

Due to the influence of nonlinear radiation distortion and geometric deformation, achieving multimodal image matching remains a challenging task. To address these issues, this paper proposes a method called radiation invariant phase correlation (RIPC) to simultaneously estimate the rotation, scale, and displacement changes of multimodal image pairs. Firstly, based on the local structure characteristics of the image itself, we harness the nonlinear invariance of kernel canonical correlation analysis to devise the multimodal local self-correlation (MLSC) descriptor. This descriptor is resilient to nonlinear radiative differences, as well as local rotation and scale variations. Subsequently, we incorporate the log-polar coordinate transformation to capture the overall rotation and scale changes in the image, enabling independent representation of these factors on the Cartesian coordinate system. Finally, drawing upon the continuity of displacement estimation, as well as rotation and scale estimation, we construct a five-dimensional descriptor tailored for phase correlation. Extensive experiments conducted on five open-source datasets demonstrate that our proposed method surpasses state-of-the-art (SOTA) techniques in matching performance. Furthermore, our RIPC method achieves matching accuracy within 2-pixel threshold, which underscores its effectiveness in multimodal remote sensing image matching.

1. Introduction

Over the past few decades, remote sensing technology has witnessed remarkable advancements, and evolving towards the joint application of multi-sensor, multi-resolution, and multi-temporal data. In contrast to relying solely on single modal data, jointly analyzing observation data collected by heterogeneous sensors can offer a richer and more comprehensive portrayal of scene information. However, due to the physical model of the sensor and the relative position of the imaging platform, there still exists offset between different images. These misalignments in remote sensing images can significantly hinder their subsequent utilization. Therefore, achieving the precise matching of multimodal remote sensing image pairs serves as a crucial prerequisite for downstream tasks, such as image fusion (Ye et al., 2024b), change detection (Wang et al., 2023), and three-dimension reconstruction (Qiu et al., 2018) coupled with multi-sensor observation data.

Nevertheless, when designing a multimodal image matching algorithm, two crucial issues must be taken into account.

(1) Radiation distortion. Usually, radiation distortion occurs due to the diverse imaging principles employed by different sensors. For instance, optical sensors capture color information while infrared sensors record radiation information. This diversity leads to distinct textures and intensities within the resulting multimodal images, even some image types may even contain irreparable system noise.

(2) Geometric deformation. Geometric deformation arises from the varying imaging poses in natural scenes. The imaging area, direction, amplitude, and resolution can vary significantly across different imaging modes. Consequently, differences in rotation and proportion between remote sensing images are inevitable. When performing multimodal image matching, it is imperative to consider the impact of geometric deformation. An illustrative example is presented in Figure 1.

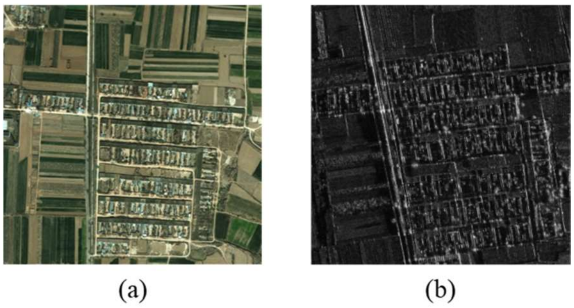


Figure 1. Multimodal image pair. (a) Optical image. (b) SAR image with rotation and scale change.

Therefore, the objective of this article is to devise a robust matching method that can effectively withstand nonlinear radiation differences, rotations, and scale variations that exist between multimodal remote sensing images.

Multimodal image matching has always been a research hotspot. Numerous matching algorithms have emerged, primarily categorized into three groups: area-based, feature-based, and learning-based methods. Recently, with the growing popularity of deep learning, its applications in multimodal remote sensing image matching have increased significantly. Prominent methods include CNN-based methods (Zhou et al., 2021, Ye et al., 2024a), Transformer-based methods (Chen et al., 2023) and GAN-based methods (Du et al., 2020). Nevertheless, these methods typically rely on extensive training datasets and exhibit limited transferability. While unsupervised learning methods do not depend on training datasets, they often face challenges in parameter transformation and loss function configuration (Ye et al., 2022), as well as a reliance on high-performance computing resources. These limitations significantly restrict their application in the realm of multimodal image matching.

Feature-based methods involve extracting invariant features from images and aligning them by assessing the similarity between those features. This category encompasses various methods, such as feature-point-based methods (Li et al., 2019), feature-line and edge-based methods (Sun et al., 2015), correlation-region-based methods (Li et al, 2022), and local-feature-based methods (Xiong et al., 2019). However, a significant challenge with these

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-309-2024 | © Author(s) 2024. CC BY 4.0 License.

methods is the difficulty in achieving high repeatability in feature detection (Ye et al., 2017), which ultimately hinders their matching performance.

The area-based method is a completely different method, with the minimum unit of these methods being a region. This method aligns two images by calculating the similarity between the predefined template and the region to be matched (Jiang et al., 2021). This type of method can avoid feature detection steps with low repeatability between images, and can detect control points in small search areas. In addition, commercial software such as ENVI also use area-based methods in the automatic matching module, which indirectly indicates that this type of method is more in line with practical application requirements.

Phase correlation (Kuglin, 1975), as a classic area-based measurement method, has been widely applied in image matching. However, this similarity measurement method lacks adaptability to the inherent nonlinear radiation differences in multimodal image pairs. Phase correlation essentially relies on the time shift invariance of Fourier transform, so this method is only applicable to image pairs with completely consistent or linearly varying amplitude spectra. Nevertheless, the significant intensity and texture disparities in multimodal images result in nonlinear variations in their amplitude spectra after Fourier transformation, so the traditional phase correlation method is ineffective. Despite the profound differences between multimodal images, the topological relationship between their overall and local structures remains consistent. Leveraging this locally invariant topological relationship, we can estimate the correlation of the structures to devise a novel descriptor. By incorporating this descriptor into phase correlation methods, we can address the challenge of nonlinear radiation differences in multimodal image matching. Furthermore, the logarithmic polar coordinate transformation offers a means to quantify rotation and scale changes between image pairs on the Cartesian coordinate system. This transformation enables us to estimate these two changes through phase correlation. As a result, we propose a radiation invariant phase correlation (RIPC) method. This method constructs multimodal local self-correlation descriptors (MLSC) that capitalize on the similarity of local structures and describes rotation and scale changes using the logarithmic polar coordinates transformation. Finally, according to the continuity of displacement estimation along with rotation and scale estimation, we ultimately construct a five-dimensional descriptor for matching.

The main contributions of this article are as follows: (1) Constructing a descriptor MLSC that can resist nonlinear radiation differences between multimodal images; (2) A template matching method, RIPC, has been proposed that can simultaneously estimate rotation, scale and displacement changes in multimodal images.

2. Methodology

The essence of image matching is to find and restore the relative position of image patches in another image. In this section, we will introduce phase correlation and its limitations, and provide a detailed description of the RIPC method proposed in this paper.

Based on the local invariant structure of the image, this method constructs a multimodal local self-correlation descriptor, then, describes the scale and rotation changes of the image by using the log-polar coordinate transformation, and finally constructs a robust five-dimensional descriptor for matching.

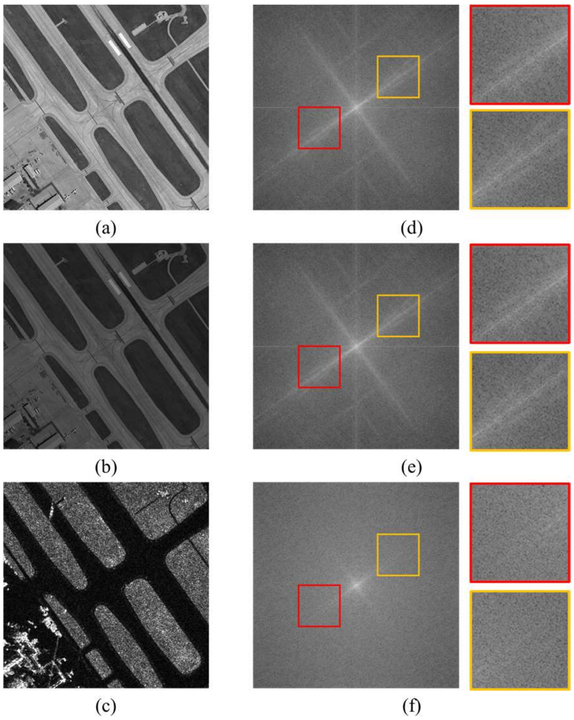


Figure 2. The variation and invariance of amplitude. (a) Optical image. (b) Optical images with illuminance changes compared to (a). (c) SAR image. (d)-(f) represents the visualization results and detailed display of the amplitude spectrum of (a)-(c).

2.1 Prior Knowledge—Phase Correlation

If image I_2 is obtained by translating image I_1 by (x_0, y_0) pixels, then the cross-power spectrum C of I_1 and I_2 in the frequency domain is:

$$C = \frac{F_1(\mu,\nu)F_2^*(\mu,\nu)}{\left|F_1(\mu,\nu)F_2(\mu,\nu)\right|} = \exp\left[-j2\pi\left(ux_0 + vy_0\right)\right] \qquad (1)$$

where F_1 and F_2 are the results of the Fourier transform of I_1 and I_2, respectively, and F^* represents conjugate complex numbers. It is not difficult to find that if the content of I_1 and I_2 are the same, the amplitude term in equation (1) will be reduced, and the value of C is only related to the phase difference. If this phase difference is subjected to an inverse Fourier transform, the result is an impulse function that approximates a 2D Dirac function, with the only non-zero coordinate being the translation (x_0, y_0). In an ideal situation, this method is only affected by phase information, so it is called phase correlation method. Therefore, if the amplitude spectra of the two images used for matching are different, the amplitude related terms in equation (1) cannot be cancelled out, and the inverse Fourier transform result of C will no longer approximate a 2D Dirac function.

Coincidentally, multimodal image pairs often have different amplitude spectra, as shown in Figure 2. (a) and (b) are optical images with only illuminance differences, and there is basically

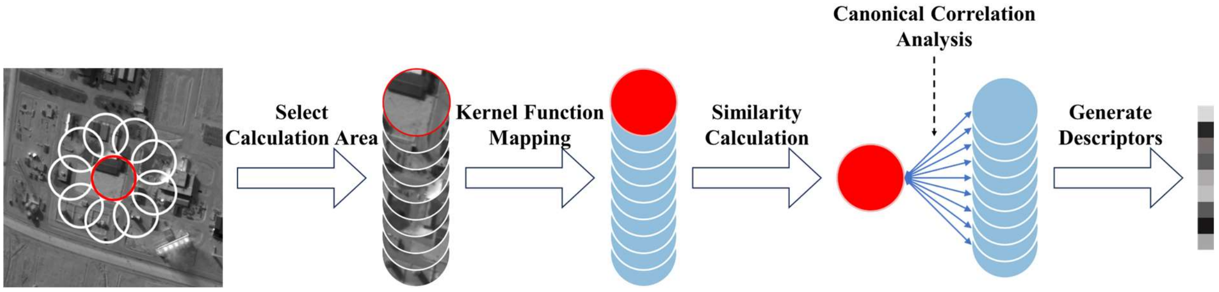


Figure 3. The construction process of the multimodal local self-correlation (MLSC) descriptor.

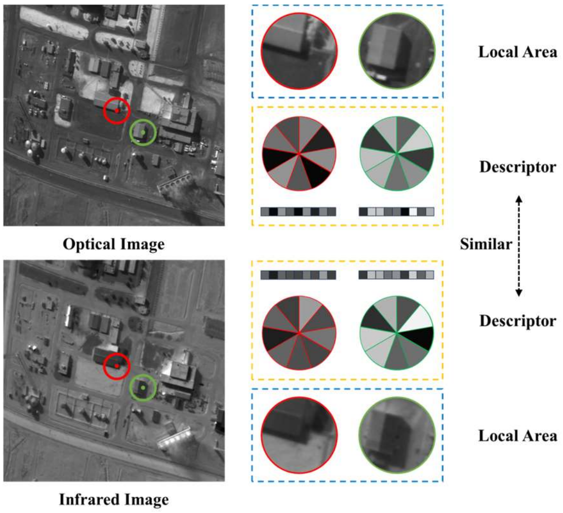


Figure 4. The MLSC descriptors of optical and infrared images.

no difference when stretching the amplitude spectrum to display [0, 255]. (c) represents the SAR image of the same region, and its amplitude spectrum differs significantly from (a) and (b). It should be explained here that although the amplitude spectrum may not appear to differ significantly, the actual amplitude spectrum has a very large order of magnitude. In order to facilitate display, the images in Figure 2 were logarithmically stretched, so all small differences can correspond to larger differences in the actual amplitude spectrum. Therefore, phase correlation is difficult to use for multimodal image matching. It is necessary to construct a new descriptor that can homogenize multimodal images.

2.2 Multimodal Local Self-correlation Descriptor (MLSC)

Although multimodal images exhibit significant differences in intensity and texture, the overall position and topological relationships of elements in the image remain unchanged. Specifically, regardless of the modality of the image, there is basically no significant change in the similarity between a pixel neighborhood and the neighboring pixel neighborhood. Therefore, we can fully utilize this local similarity relationship to construct descriptors that accurately describe image features. Such descriptors can better capture the structural information of images, providing strong support for subsequent image processing and analysis.

The multimodal local self-correlation (MLSC) descriptor proposed in this article is an improvement on the classic local self-similarity (LSS) descriptor. Although the LSS operator is widely used in basic image research, through extensive experiments, we have found that this method does not possess linear invariance and nonlinear invariance, and is susceptible to radiation distortion and noise interference. Meanwhile, the LSS descriptor is relatively sparse, making it difficult to accurately represent the local structure of multimodal images. The MLSC descriptor proposed in this article improves on the above shortcomings. This descriptor combines kernel functions with canonical correlation analysis to construct a dense descriptor with nonlinear radiative invariance, which is suitable for describing local structural features in multimodal images. The process of constructing feature descriptors for each pixel is shown in Figure 3. The specific process of this step is as follows:

(1) The first step is to define a circular neighborhood of a certain size centred on a pixel, denoted as α_0. And select n circular neighborhoods of the same size with equal angular spacing in an area m pixels away from α_0, denoted as$\alpha_1, \alpha_2, \dots, \alpha_\mathrm{n}$. Existing research has shown that using circular templates to represent local features can maintain the angle invariance of local features (Li et al., 2023).

(2) The second step is to define a kernel function κ that maps$\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_\mathrm{n}$ to a high-dimensional space, denoted as $\kappa(\alpha_0), \kappa(\alpha_1), \kappa(\alpha_2), \dots, \kappa(\alpha_\mathrm{n})$. By transforming low dimensional linearly indivisible spaces into high-dimensional linearly separable spaces, it is possible to find linear relationships in high-dimensional spaces that are difficult to determine with extremely low computational costs.

(3) For $\kappa(\alpha_0)$ and $\kappa(\alpha_\mathrm{n})$, there can always be a set of orthogonal bases ρ_0 and ρ_n that maximize the correlation coefficient between $\rho_0\kappa(\alpha_0)$ and $\rho_\mathrm{n}\kappa(\alpha_\mathrm{n})$. It is easy to prove that any orthogonal basis matrix is linearly invariant, and the k-th eigenvalue is the square of the k-th canonical correlation coefficient. So, it is possible to calculate invariant features through the traces of orthogonal basis matrices. The third step is to use the kernel canonical correlation analysis method to calculate the orthogonal basis, and use this to construct multi-directional features of the central and surrounding neighborhoods. If $\lfloor\cdot\rfloor$ represents the calculation process of a set of orthogonal bases, then this feature can be expressed as:

$$\tau_n = \mathrm{tr}\left(\left\lfloor \kappa(\alpha_0), \kappa(\alpha_n) \right\rfloor\right) = U^{-1}VW^{-1}V^T \tag{2}$$

where

$$\begin{aligned}
K_X &= \kappa(\alpha_0)^T \kappa(\alpha_0) \\
K_Y &= \kappa(\alpha_n)^T \kappa(\alpha_n) \\
U &= K_X^T J K_X / N + \eta K_X \\
V &= K_X^T J K_Y / N \\
W &= K_Y^T J K_Y / N + \eta K_Y \\
J &= I - aa^T \\
a &= \left(1, \cdots, 1\right)^T
\end{aligned} \tag{3}$$

where N is the length of vector $\kappa(\alpha_0)$ or $\kappa(\alpha_n)$, and η is a minimum term that prevents the generation of singular matrices.

(4) Finally, collect the multi-directional feature descriptors into the matching feature vectors. After calculating all pixels, normalize according to the direction to achieve better illumination invariance.

As mentioned above, MLSC descriptors obtain locally similar structural information of images, so as long as two images with different modalities have the same structural and geometric information, they can be described using this descriptor. Figure 4 shows the similarity level of this descriptor in multimodal image pairs.

2.3 Logarithmic Polar Coordinate Transformation

If the centers of I_a and I_b are aligned, and I_b is I_a with an angle difference of $\theta°$ and a scale difference of a. Assuming the image center is denoted as (0, 0), then the point (x, y) on I_a and the corresponding points (x', y') on I_b satisfies:

$$I_b\left(x',y'\right)=\frac{1}{|a|}I_a\left(x\cos\theta+y\sin\theta,-x\sin\theta+y\cos\theta\right) \quad (4)$$

Obviously, both the horizontal and vertical coordinates contain four unknown variables: x, y, a, and θ, so it is not practical to directly evaluate these values. We perform a logarithmic polar coordinate transformation on equation (4), let ρ be the minimum scale sampling interval, then:

$$I_b\left(\log\rho',\theta'\right)=I_a\left(\log\rho-\log|a|,\theta-\theta_0\right) \quad (5)$$

Equation (5) transforms the coupled rotation and scale changes into independent estimation problems regarding the translation of scale ρ and angle θ. Even if there are rotation and scaling changes between image pairs, the rotation and scaling changes of the image itself can still be estimated through the simplest coordinate translation. Similarly, rotation and scale changes can also be estimated using phase correlation methods. In other word, if the MLSC descriptors in section 2.2 are subjected to logarithmic polar transformation, it is possible to estimate the rotation and scale of multimodal images. If an image with a size of $m \times m$ in the Cartesian coordinate system has a size of $n \times n$ on the logarithmic polar coordinate plane, and the matching result in “scale - rotation” plane is (u, v), then the scale and angle can be expressed as:

$$\begin{aligned}\text{Angle} &= 2\pi v / n \\ \text{Scale} &= \rho^u\end{aligned} \quad (6)$$

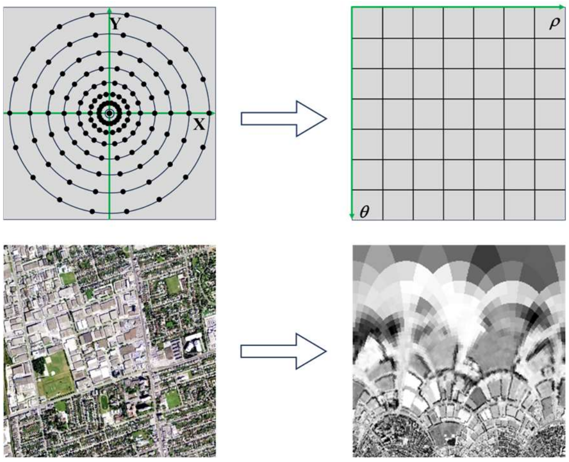


Figure 5. The schematic diagram of logarithmic polar coordinate transformation.

2.4 The Construction of Five-dimensional Descriptors

In fact, the estimation of rotation and scale changes in section 2.3 is flawed, as the features do not exhibit translation invariance after logarithmic polar coordinate transformation. Therefore, this method can only be applied to cases of center alignment.

However, any linear transformation is not just about rotation and scale changes, displacement transformation must also be taken into consideration. This section constructs a five-dimensional descriptor to simultaneously describe the angle, scale, and displacement changes of multimodal images.

Any linear transformation can be expressed as a product of the displacement transformation matrix M_D, rotation transformation matrix M_R, and scale transformation matrix M_S, and it has been proven in section 2.3 that M_R and M_S can be estimated simultaneously. So, the process of simultaneously estimating three types of changes is essentially estimating matrix M_R and matrix M_S under the most suitable matrix M_D. Since matrix M_R and matrix M_D are bound, this is actually two consecutive translation estimates. In general, the most suitable rotation and scale changes are estimated by sliding pixel by pixel in the displacement plane. However, the time complexity is O(n^2), which greatly reduces the matching efficiency. It is not difficult to see that even with pixel by pixel sliding matching, the processes of two matches share the maximum value of correlation peak. And essentially, the translation amount can be estimated through phase correlation, so we can combine the two matching processes into one by using the shared maximum value of correlation peak, which eliminates the need for repeated pixel by pixel sliding matching processes and reduces the time complexity to O(1).

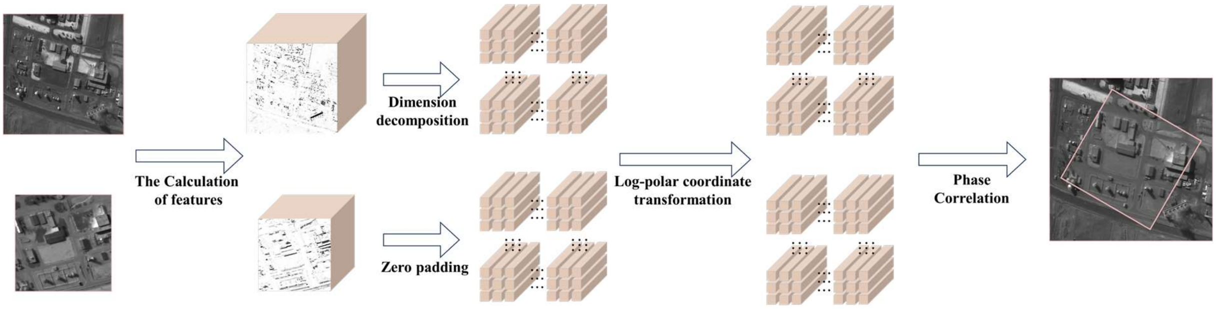

Dimension decomposition
The Calculation of features
Zero padding
Log-polar coordinate transformation
Phase Correlation

Figure 6. The construction of five-dimensional descriptor and the basic process of RIPC.

Based on this idea, we can crop the image according to the template image size (this process can be quickly achieved through pointers), then construct three-dimensional descriptors separately and perform logarithmic polar coordinate transformation, and finally expand the three-dimension descriptors into five-dimensions along two sliding directions. Among them, the 4th and 5th dimensions respectively simulate sliding in the X and Y directions. The construction process of the five-dimensional descriptor is shown in Figure 6. Since the 4th and 5th dimensions simulate sliding processes, the five-dimensional descriptor is only used for reference images. The template image is still three-dimension. However, if phase correlation is used to calculate the optimal matching position, the data dimension must be consistent. Therefore, it is necessary to add zeros in the three-dimensional descriptor of the template image to align the data to ensure consistency between the two sets of data. At the same time, adding zeros can also separate the overlapping frequencies and reduce folding errors in the fast Fourier transform process.

3. Experiment and Analysis

In this section, to verify the superiority of the proposed RIPC method in matching performance, we tested it on a large number of publicly available datasets and compared it with five state of the art methods – SIFT (Lowe, 2004), RIFT (Li et al., 2019), Superpoint (DeTone et al., 2018) + Superglue (Sarlin et al., 2020), SRIF (Li et al., 2023), and ReDFeat (Deng et al., 2022).

3.1 Datasets

To verify the effectiveness of our proposed RIPC method, we selected nearly a thousand pairs of multimodal images from five datasets for experiments. This includes Optical-to-Optical, Optical-to-Infrared, Optical-to-SAR, Optical-to-Depth maps, and Optical-to-Labels. To ensure the diversity of test data, our testing is not limited to remote sensing data, but also includes rich close range photogrammetric data, even artificial raster data. These data have inconsistent lighting conditions, including normal exposure, overexposure, and underexposure data. The resolution of an image is variable, ranging from close range images at the decimeter level to remote sensing images at the meter level. These images have severe distortion, especially radiative distortion. The introduction of these images will pose a huge challenge to the matching algorithm and can better test the robustness of our method.

Optical-to-Optical(Cai et al., 2018): The Optical-to-Optical dataset is close range photogrammetric images with high resolution and rich texture, but the exposure levels between image pairs are different.

Optical-to-Infrared(Xu et al., 2020): The Optical-to-Infrared dataset is also a close range photogrammetric image, belonging to the autonomous driving dataset. Some of the images are road images collected at night, and some are affected by streetlights and the Tindar effect.

Optical-to-SAR(Xiang et al., 2020): The Optical-to-SAR dataset is a standard pair of remote sensing images, containing images of various scenes, and this type of dataset has significant nonlinear radiative differences. Due to the large size of the dataset, we extracted 200 pairs of images for experimentation.

Optical-to-Depth maps(Silberman et al., 2012): The Optical-to-Depth maps dataset is indoor data, and due to the fact that depth maps are a mode of manually expressing depth of field, the structural richness of depth maps is low, resulting in limited texture information. There are significant differences in radiation between images. Due to the large size of the dataset, we extracted 200 pairs of images for experimentation.

Optical-to-Labels(Silberman et al., 2012): The Optical to Labels dataset is consistent with the Optical to Depth maps dataset, but Labels have a more pronounced geometric structure but a more monotonous texture.

Due to the inconsistent resolution of the data, in order to facilitate batch processing, we sampled the size of all data to 512 * 512. All examples are shown in Figure 8. Meanwhile, in order to prevent substantial errors caused by significant differences in initial conditions, the images used in this study were corrected by physical models and resampled. The initial images were aligned between pixels without any differences in rotation, scale, or translation. The rotation, scale, and displacement changes used in this experiment are all simulated by manually rotating, scaling, and cropping the image. However, this method often reduces image quality and causes black edges due to rotation, further complicating matching.

3.2 Implementation Details and Evaluation Criteria

In our proposed method, there are a total of four parameters that need to be set, namely the circular neighborhood radius of descriptor *r*, the number of descriptor directions *N*, the polar radial sampling spacing of log-polar transformation N_ρ, and the angular sampling spacing of log-polar transformation N_θ. The parameter *r* determines the size of the local circular patches used for feature description, which reflects the richness of local information. Appropriate parameters can reasonably describe features. However, excessively large image patches are highly susceptible to local geometric distortion. The parameter *N* represents the number of sampling directions in the multi-directional feature description process. Usually, the higher the number of directions, the richer the information content of the constructed descriptor, but the higher the computational complexity. The parameters N_ρ and N_θ represent the sampling spacing in the scale and angle directions. If these parameters are small, that is, the sampling is dense, excessive interpolation will lead to local information inflation, resulting in extremely high computational complexity and changes in local features. On the contrary, sparser sampling results in more loss of image features and greater errors. Taking all factors into consideration, we will set *r* as 2, *N* as 9, N_ρ to 1.022, and N_θ to 1.41°. Under this condition, theoretically, when the scale difference is less than 1.6 times, the scale error will not exceed 2%. Meanwhile, the angle error will never exceed 0.7 °.

In order to describe the matching effect more reasonably, the successful matching rate (*SMR*) is used as an evaluation criterion, which calculates the proportion of the number of successful matches (*NSC*) in the total number of matches (*TN*), which $SMR = NSC / TN \times 100\%$. Here, to clarify the criteria for successful matching, as the image itself has true values and there is no non-rigid deformation inside, we define that a successful matching is achieved when the RMSE is less than 3 when the matching result is projected onto the original image according to a linear transformation matrix.

3.3 Evaluation of matching performance

This section compares our RIPC with five SOTA matching methods: SIFT, RIFT, Superpoint + Superglue, SRIF, and ReDFeat, and evaluates them from both quantitative description and qualitative visualization perspectives. In addition, all data undergoes random angle, scale, and displacement transformations, with an angle range of [0, 2π] and a scale range of [1, 2]. The size of the artificially simulated images used for matching is fixed at 384 × 384. To ensure the fairness in the comparison process, we obtained specific implementation codes for various methods used for comparison on the corresponding author's personal website. To ensure optimal matching performance, all hyperparameters and implementation details depend on the best parameters and details provided in the original text. Among them, the deep learning method uses the pre-trained model provided in the paper.

3.3.1 SMR and Matching accuracy.

Figure 7 shows the SMR metrics for mixed rotation, scale, and displacement changes using different methods on multiple datasets. Among them, the SMR of RIFT is very low because the method itself does not have scale invariance and can only be applied to situations with no scale changes or small scale changes. The SIFT and Superpoint + Superglue methods have excellent performance on optical datasets, but they perform poorly on multimodal datasets, especially on SAR, Depth map, and Label datasets with significant nonlinear radiation differences, where SMR approaches 0. This is because these methods do not have radiation invariance and are only effective for visual images. SRIF, ReDFeat, and our proposed RIPC all exhibit good performance on multimodal datasets. Our RIPC method achieved SMR of 92.78%, 88.07%, 75.5%, 73%, and 77.5% on five datasets, respectively. On multimodal datasets, the SMR of our method is approximately 5% higher than SRIF and approximately 7% higher than ReDFeat.

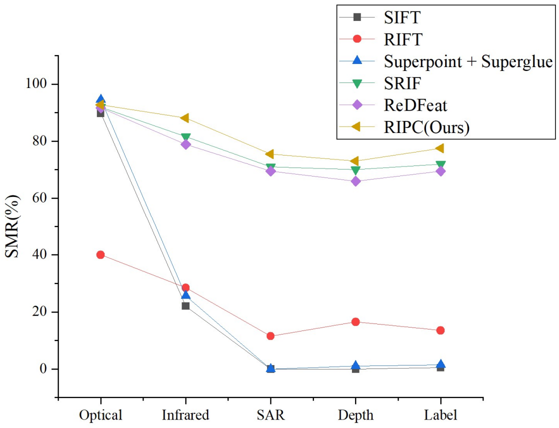


Figure 7. The SMRs of SIFT, RIFT, Superpoint + Superglue, SRIF, ReDFeat and our RIPC.

Figure 8 visualizes the registration performance of the proposed RIPC method using checkboard. In these example images, it can be seen that the edges of each checkboard can be aligned well without obvious misalignment. This further validates the robustness of our proposed RIPC method. Overall, these analyses demonstrate that our RIPC is more effective than state-of-the-art matching methods in resisting significant radiation differences and achieving estimation of multimodal image rotation and scale.

Meanwhile, in order to evaluate the accuracy of matching, we calculated the average value of root mean square error (RMSE) of the image pairs correctly matched by the proposed method on each dataset. The test results are shown in Tabel 1.

Dataset	RMSE (pixels)	Dataset	RMSE (pixels)
Optical	1.67	Infrared	1.88
SAR	1.84	Depth	1.75
Label	1.77		

Table 1. The RMSE of RIPC on Each Datasets

As can be seen, our method has an average RMSE of less than 2 on each dataset, indicating high accuracy and stability.

3.3.2 Matching efficiency.

In addition to SMR and accuracy, matching time is also an important evaluation indicator. The experiment in this article was implemented on a personal computer with an Intel i7-12700KF CPU, NVIDIA RTX 3090 GPU, and 16GB RAM. We only report the average matching time for a single match here, as shown in Tabel 2.

Method	Time (s)	Method	Time (s)
SIFT (C++)	1.17	SRIF (C+MATLAB)	4.17
RIFT (MATLAB)	24.77	ReDFeat (Python)	2.12
S+S* (Python)	0.66	RIPC (MATLAB)	17.23

* S+S represents Superpoint + superglue.

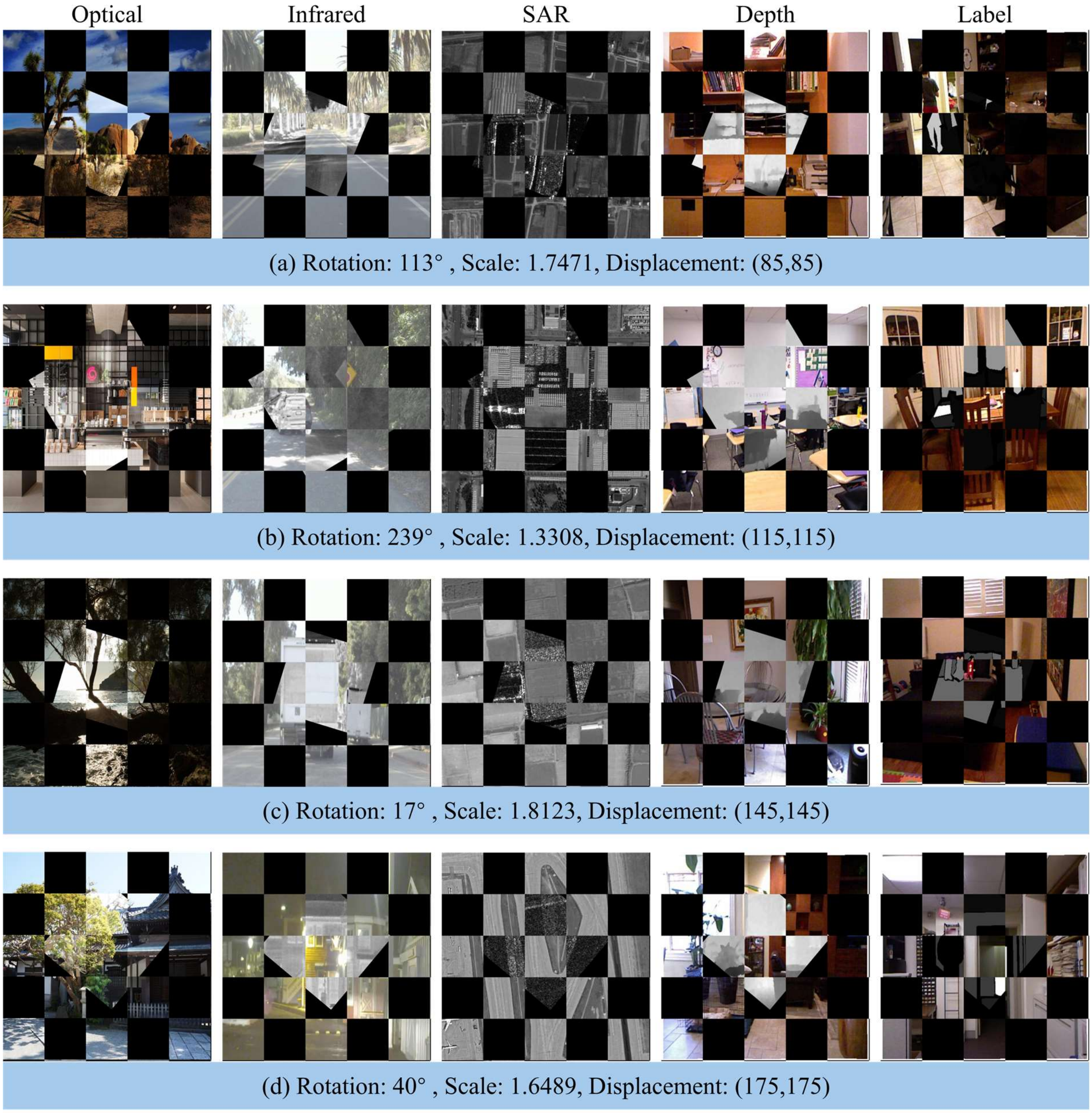
Table 2. The Running Time of Each Algorithm

From Table 2, it can be seen that the Superpoint + Superglue method has the shortest matching time, while the RIFT method has the longest matching time. The time consumption of the other methods, sorted in descending order, is SIFT, ReDFeat, SRIF, and RIPC. Our proposed RIPC method also has a longer matching time. This is because the RIPC method requires building a large five-dimensional descriptor, which will take a lot of time. Of course, this also depends on our RIPC method not using any acceleration modules.

Based on the above experiments, it can be concluded that our proposed RIPC method has higher SMRs on all datasets, followed by SRIF. Although RIPC is more time-consuming than most SOTA methods, it is more robust to nonlinear radiative differences between multimodal images. And it is one of the few multimodal image matching methods that can handle rotation and scale changes. Therefore, considering all factors, RIPC is a more robust method for multimodal image matching.

4. Conclusion

In this paper, we introduce the radiation invariant phase correlation (RIPC) method, specifically designed to overcome the challenges associated with multimodal matching using phase correlation. This method enables us to simultaneously assess rotation, scale, and displacement variations between multimodal image pairs. Firstly, we utilize kernel canonical correlation analysis to delve into the local structure of the image, constructing a multimodal local self-correlation (MLSC) descriptor resilient to local rotation invariance. Subsequently, the logarithmic polar coordinate transformation is employed to



(a) Rotation: 113° , Scale: 1.7471, Displacement: (85,85)

(b) Rotation: 239° , Scale: 1.3308, Displacement: (115,115)

(c) Rotation: 17° , Scale: 1.8123, Displacement: (145,145)

(d) Rotation: 40° , Scale: 1.6489, Displacement: (175,175)

Figure 8. Checkboard visualization of RIPC.

effectively capture rotation and scale changes, which are often difficult to quantify in the Cartesian coordinate system. Finally, leveraging the continuity of the matching process, we develop a five-dimensional descriptor tailored for phase correlation. This method establishes a versatile multimodal image matching framework, with the flexibility to replace MLSC descriptors with any rotation-resistant descriptor. Through extensive testing on nearly a thousand pairs of multimodal images and comparisons with five state-of-the-art (SOTA) methods, the RIPC method demonstrates superior matching performance. However, it is worth noting that the RIPC method, due to the size of its descriptors, is relatively time-consuming and memory-intensive. To address these limitations, our future work will focus on rewriting the algorithm using faster computation methods. Additionally, we plan to explore the integration of traditional dimensionality reduction techniques, such as PCA, with deep learning methods to achieve a more lightweight descriptor. Furthermore, given that the method proposed in this article is inherently self-iterative, we also aim to introduce unsupervised learning strategies to enable unsupervised or self-supervised learning in future iterations.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 42271446) and Key Issue of China Railway Design Corporation's Science and Technology Research Program (No. 2023A0240104). The authors would like to thank everyone who has contributed datasets and fundamental codes to the public.

References

Cai, J., Gu, S., and Zhang, L., 2018. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Trans. Image Process.*, 27(4), 2049-2062.

Chen, J., Chen, X., Chen, S., et al, 2023. Shape-Former: Bridging CNN and Transformer via ShapeConv for multimodal image matching. *Inf. Fusion*, 91, 445-457.

Deng, Y., and Ma, J., 2022. ReDFeat: Recoupling detection and description for multimodal feature learning. *IEEE Trans. Image Process.*, 32, 591-602.

DeTone, D., Malisiewicz, T., and Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. In *IEEE Conf. Comput. Vis. Pattern Recog.*(CVPR'18) , 224-236.

Du, W.-L., Zhou, Y., Zhao, J., et al, 2020. K-means clustering guided generative adversarial networks for SAR-optical image matching. *IEEE Access*, 8, 217554-217572.

Jiang, X., Ma, J., Xiao, G., et al, 2021. A review of multimodal image matching: Methods and applications. *Inf. Fusion,* 73, 22-71.

Kuglin, C. D., 1975. The phase correlation image alignment method. In *IEEE Int. Conf. on Cybernetics and Society*, 163-165.

Li, J., Hu, Q., and Ai, M., 2019. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Trans. Image Process.*, 29, 3296-3310.

Li, J., Hu, Q., and Zhang, Y., 2023. Multimodal image matching: A scale-invariant algorithm and an open dataset. *ISPRS-J. Photogramm. Remote Sens.*, 204, 77-88.

Li, Z., Yue, J., and Fang, L., 2022. Adaptive regional multiple features for large-scale high-resolution remote sensing image registration. *IEEE Trans. Geosci. Remote Sensing*, 60, 1-13.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60, 91-110.

Qiu, C., Schmitt, M., and Zhu, X. X., 2018. Towards automatic SAR-optical stereogrammetry over urban areas using very high resolution imagery. *ISPRS-J. Photogramm. Remote Sens.*, 138, 218-231.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., et al, 2020. Superglue: Learning feature matching with graph neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*(CVPR'20) , 4938-4947.

Silberman, N., Hoiem, D., Kohli, P., et al, 2012. Indoor segmentation and support inference from rgbd images. In *Proc. Eur. Conf. Comput. Vis.*(ECCV'12), 746-760.

Sun, Y., Zhao, L., Huang, S., et al, 2015. Line matching based on planar homography for stereo aerial images. *ISPRS-J. Photogramm. Remote Sens.*, 104, 1-17.

Wang, M., Zhu, B., Zhang, J., et al, 2023. A Lightweight Change Detection Network based on Feature Interleaved Fusion and Bi-stage Decoding. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*.

Xiang, Y., Tao, R., Wang, F., et al, 2020. Automatic registration of optical and SAR images via improved phase congruency model. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, 13, 5847-5861.

Xiong, X., Xu, Q., Jin, G., et al, 2019. Rank-based local self-similarity descriptor for optical-to-SAR image matching. *IEEE Geosci. Remote Sens. Lett.*, 17(10), 1742-1746.

Xu, H., Ma, J., Jiang, J., et al, 2020. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1), 502-518.

Ye, Y., Shan, J., Bruzzone, et al, 2017. Robust registration of multimodal remote sensing images based on structural similarity. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5), 2941-2958.

Ye, Y., Tang, T., Zhu, B., et al, 2022. A multiscale framework with unsupervised learning for remote sensing image registration. *IEEE Trans. Geosci. Remote Sensing*, 60, 1-15.

Ye, Y., Yang, C., Gong, G., et al, 2024a. Robust optical and SAR image matching using attention-enhanced structural features. *IEEE Trans. Geosci. Remote Sensing*, 62, 1-12.

Ye, Y., Zhang, J., Zhou, L., et al, 2024b. Optical and SAR image fusion based on complementary feature decomposition and visual saliency features. *IEEE Trans. Geosci. Remote Sensing*, 62, 1-15.

Zhou, L., Ye, Y., Tang, T., et al, 2021. Robust matching for SAR and optical images using multiscale convolutional gradient features. *IEEE Geosci. Remote Sens. Lett.*, 19, 1-5.