ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/373951077

Multimodal image matching: A scale-invariant algorithm and an open dataset

Article *in* ISPRS Journal of Photogrammetry and Remote Sensing · September 2023

DOI: 10.1016/j.isprsjprs.2023.08.010

CITATIONS
2

READS
320

3 authors:

Jiayuan Li
Wuhan University
53 PUBLICATIONS **1,189** CITATIONS
SEE PROFILE

Qingwu Hu
Wuhan University
109 PUBLICATIONS **1,751** CITATIONS
SEE PROFILE

Yongjun Zhang
Wuhan University
93 PUBLICATIONS **1,156** CITATIONS
SEE PROFILE

All content following this page was uploaded by Jiayuan Li on 16 September 2023.

The user has requested enhancement of the downloaded file.

ISPRS Journal of Photogrammetry and Remote Sensing 204 (2023) 77–88

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Multimodal image matching: A scale-invariant algorithm and an open dataset

Jiayuan Li, Qingwu Hu *, Yongjun Zhang *

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

ARTICLE INFO

Keywords:
Image matching
Feature descriptor
Dataset
SAR-optical
Multimodal images

ABSTRACT

Multimodal image matching is a core basis for information fusion, change detection, and image-based navigation. However, multimodal images may simultaneously suffer from severe nonlinear radiation distortion (NRD) and complex geometric differences, which pose great challenges to existing methods. Although deep learning-based methods had shown potential in image matching, they mainly focus on same-source images or single types of multimodal images such as optical-synthetic aperture radar (SAR). One of the main obstacles is the lack of public data for different types of multimodal images. In this paper, we make two major contributions to the community of multimodal image matching: First, we collect six typical types of images, including optical-optical, optical-infrared, optical-SAR, optical-depth, optical-map, and nighttime, to construct a multimodal image dataset with a total of 1200 pairs. This dataset has good diversity in image categories, feature classes, resolutions, geometric variations, etc. Second, we propose a scale and rotation invariant feature transform (SRIF) method, which achieves good matching performance without relying on data characteristics. This is one of the advantages of our SRIF over deep learning methods. SRIF obtains the scales of FAST keypoints by projecting them into a simple pyramid scale space, which is based on the study that methods with/without scale space have similar performance under small scale change factors. This strategy largely reduces the complexity compared to traditional Gaussian scale space. SRIF also proposes a local intensity binary transform (LIBT) for SIFT-like feature description, which can largely enhance the structure information inside multimodal images. Extensive experiments on these 1200 image pairs show that our SRIF outperforms current state-of-the-arts by a large margin, including RIFT, CoFSM, LNIFT, and MS-HLMO. Both the created dataset and the code of SRIF will be publicly available in https://github.com/LJY-RS/SRIF.

1. Introduction

Image matching plays an important role in the fields of remote sensing and photogrammetry. It is a fundamental problem for visual understanding and interpretation such as image fusion (Ma et al., 2019; Li et al., 2022a), change detection (Tewkesbury et al., 2015; Parente et al., 2021), and image localization and navigation (Mur-Artal et al., 2015). However, due to the complexity of high-level remote sensing applications, the information richness of single-modality data is insufficient. It is necessary to comprehensively utilize data of different modalities to achieve complementary strengths, thereby improving the accuracy and reliability of image understanding. Fortunately, with the rapid development of sensor technology, imaging devices such as visible cameras, infrared cameras, synthetic aperture radar (SAR), and lasers are emerging, providing a variety of data sources for earth observation. Therefore, how to effectively integrate multi-sensors, multi-resolution, multi-temporal data and conduct in-depth analysis has become a research hotspot, and multimodal image matching is one of the core problems that need to be addressed urgently (Sui et al., 2022).

Multimodal image matching generally refers to the matching between multi-sensor images with different imaging mechanisms such as optical-SAR and optical-depth (Li et al., 2020a). There are severe nonlinear radiation differences (NRDs) and complex geometric differences such as scale, rotation, and perspective changes between images. These differences make the matching a challenging task. In recent years, many efforts have been made to try to solve this problem and many multimodal matching algorithms have been proposed. These methods can be grouped into two categories, i.e., area-based methods (e.g., histogram of orientated phase congruency (HOPC) Ye et al. (2017) and channel features of orientated gradients (CFOG) Ye et al. (2019)) and feature-based ones (e.g., radiation-variation insensitive feature transform (RIFT) Li et al. (2020a) and locally normalized image feature transform (LNIFT) Li et al. (2022b)). However, these methods mainly focus on NRDs of multimodal images and are sensitive to complex geometric variances such as scale changes. Although co-occurrence filter space matching (CoFSM) (Yao et al., 2022) and multi-scale histogram of local main orientation (MS-HLMO) (Gao et al., 2022) claim that they

* Corresponding authors.
E-mail addresses: ljy_whu_2012@whu.edu.cn (J. Li), huqw@whu.edu.cn (Q. Hu), zhangyj@whu.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.isprsjprs.2023.08.010
Received 6 December 2022; Received in revised form 22 July 2023; Accepted 26 August 2023
Available online 12 September 2023
0924-2716/© 2023 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

achieve rotation and scale invariance, these two types of geometric changes do not occur simultaneously on the same image pair in their experiments.

Thanks to the great success of deep learning technology in the field of computer vision, learning-based methods also have shown their potential in image matching task. For instance, HardNet (Mishchuk et al., 2017) and SuperPoint (DeTone et al., 2018) achieve higher matching performance on same-source images than traditional hand-crafted methods. However, these methods are limited by their generalization ability and cannot be directly applied to multimodal images. Siamese CNN show potential on optical-SAR matching task (Zhang et al., 2020; Zhou et al., 2021). However, they are area-based matching methods rather than feature-based ones. In addition, they cannot be applied to other types of multimodal images such as optical-depth, optical-map, and nighttime. An important factor hindering the successful application of deep learning techniques in multimodal matching is the lack of public data on different types of multimodal images. Each current publicly available dataset usually contains only one type of multimodal images, such as optical-SAR dataset (Huang et al., 2021; Xiang et al., 2020) and optical-infrared dataset (Brown and Süsstrunk, 2011; Jia et al., 2021).

To promote the development of multimodal image matching, especially learning-based techniques, we create and open a multimodal image dataset with six typical types of images, i.e, optical-optical, optical-infrared, optical-SAR, optical-depth, optical-map, and nighttime. This dataset contains a total of 1200 image pairs captured by three different types of imaging platforms including aerial, satellite, and close range. The images contain a rich set of features (e.g., buildings, mountains, farmland, lakes, etc.) and the resolutions range from 0.04 m to 30 m. Moreover, different geometric variations (rotation and scale) with ground truth transformation are added to these images.

We also propose a scale and rotation invariant feature transform (SRIF) method for multimodal image matching. First, an experiment is performed to study the sensitivity of different methods (methods with/without scale space) to small scale change factors. We present a simple strategy to achieve scale invariance based on the conclusion that methods with/without scale space have similar performance under small scale change factors. We obtain the scales of keypoints by projecting them into a simple pyramid scale space, which largely reduces the complexity. We then propose a local intensity binary transform (LIBT) to enhance the structure information inside multimodal images, so that feature descriptors have good distinguishability. We compare our SRIF with seven baseline and state-of-the-art methods on 1200 image pairs. The results show that SRIF outperforms them by a large margin.

Our contributions are summarized as follows:

- We create an open multimodal image dataset with 1200 pairs, which covers six typical types of images.
- We observe that methods with/without scale space have similar performance under small scale change factors. Based on this observation, we present a projection-based pyramid scale space construction strategy, which largely reduces the complexity compared to traditional Gaussian scale space.
- We propose a novel local intensity binary transform (LIBT) for structure feature map generation, which can largely enhance the structure information inside multimodal images. LIBT outperforms current state-of-the-art methods such as the local normalized image (LNI) transformation (Li et al., 2022b).
- Based on the projection-based pyramid scale space and LIBT, we propose a new scale and rotation invariant feature transform (SRIF) method for multimodal image matching.

2. Related work

Multimodal image matching methods are usually classified into two groups, i.e., area-based matching (also known as template matching or patch matching) and feature matching (Li et al., 2020a; Bas and Ok, 2021; Mohammadi et al., 2022). Each category can be further subdivided into hand-crafted methods and learning-based methods.

2.1. Area-based methods

Area-based matching generally finds the optimal location of a template image in the reference one via calculating their similarity based on a sliding window strategy.

Hand-crafted: One of the most important things for hand-crafted methods is to define the similarity measure. The sum of square differences (SSD), cross-correlation, normalized cross-correlation (NCC), and mutual information are commonly used measures, among which mutual information and its variants are more robust to NRDs and have been widely applied in multimodal image matching (Viola and Wells III, 1997; Liang et al., 2013; Öfverstedt et al., 2022b). Apart from measures in the spatial domain, phase correlation methods calculate the similarity in the frequency domain via Fourier transform (Foroosh et al., 2002), which have also been shown to be robust to non-uniform illumination changes. To enhance the performance of Fourier transform methods, several methods first convert the original images into feature space before Fourier correlation, such as HOPC (Ye et al., 2017), improved phase congruency model (Xiang et al., 2020), CFOG (Ye et al., 2019), and angle-weighted oriented gradients (AWOG) (Fan et al., 2021).

Learning-based: These methods automatically learn high-level information from a large amount of data without manual feature extraction. Considering the powerful feature extraction capabilities of deep learning, some methods only utilize deep neural networks (DNN) to obtain features and then apply traditional similarity metrics to search for the best match (Zhang et al., 2020; Zhou et al., 2021; Fang et al., 2021). In contrast, some methods perform the matching in an end-to-end manner, which compute a similar score between two patches (Zhang et al., 2019; Merkle et al., 2017; Hughes et al., 2018). Recently, several approaches directly regress the transformation parameters for registration, which generally consist of a transformation prediction DNN, a spatial registration network, and an optimizer to backpropagate the DNN (Zhao et al., 2021; Ye et al., 2022). The limitations of learning-based methods are that they rely on a wide variety of training datasets and require a lot computational resources (Xiang et al., 2021).

Area-based matching is generally sensitive to geometric transformations. Although some methods have achieved rotation and scale invariance based on a transformation optimizer (Öfverstedt et al., 2022a,b), they are sensitive to local extrema and have high computational complexity.

2.2. Feature-based methods

Feature-based matching generally consists of three major stages, i.e., keypoint detection, keypoint description, and feature vector matching. First, keypoints such as corner points with high repeatability are extracted by feature detectors (e.g., features from accelerated segment test (FAST) (Rosten and Drummond, 2006), Harris detector (Harris et al., 1988), SuperPoint (DeTone et al., 2018), etc.). These keypoints are then encoded to feature vectors via descriptors (e.g., scale-invariant feature transform (SIFT) (Lowe, 2004), RIFT (Li et al., 2020a, 2023), HardNet (Mishchuk et al., 2017), etc.), so that the features have better distinguishability. Finally, a one-to-one matching relationship between two feature sets is established and outliers are removed by a robust estimation technique or matching strategy (e.g., random sample consensus (RANSAC) family (Fischler and Bolles, 1981; Li et al., 2017), robust estimators (Li et al., 2021a,b, 2020b, 2016, 2023b), SuperGlue (Sarlin et al., 2020), etc.)

Hand-crafted: Traditional same-source image matching methods have achieved great success and become standard methods in many commercial software, such as SIFT (Lowe, 2004), SURF (Bay et al., 2008), and ORB (Rublee et al., 2011). However, the NRDs of multimodal images pose great challenges to these methods. Many efforts have been made to tackle this problem. For example, local self-similarity descriptor (LSS) and its variants improve the robustness to

illumination differences (Shechtman and Irani, 2007; Sedaghat and Mohammadi, 2019; Xiong et al., 2021); partial intensity invariant feature descriptor (PIIFD) is designed for retinal image matching (Chen et al., 2010); position-scale-orientation SIFT (Ma et al., 2016), histograms of directional maps (Fu et al., 2018), and LGHD (Aguilera et al., 2015) are suitable for multispectral image matching; improved SIFT (Fan et al., 2012), optical-SAR SIFT (OS-SIFT) (Xiang et al., 2018), and rotation-invariant amplitudes of log-Gabor orientation histograms (RI-ALGH) (Yu et al., 2021) are proposed to solve the optical-SAR matching problem. However, these methods usually only work well on specific image types and are not applicable to other types of multimodal images, so they are not generalizable. Recently, Li et al. (2020a) proposed a general multimodal feature matching method, called RIFT, which achieves good performance on different types of images. Several variants improve the RIFT by adding a scale space stage, an area-based fine-registration step, or modifying the maximum index map (Cui et al., 2020; Fan et al., 2022; Yao et al., 2022; Gao et al., 2022). Further, the LNIFT proposes a local image transform in the spatial domain to achieve near real-time processing performance (Li et al., 2022b). However, these methods mainly focus on the NRDs, while the robustness to complex geometric differences has not been fully evaluated.

Learning-based: Learning-based feature matching has achieved great progress in same-source image matching, such as learned invariant feature transform (LIFT) (Yi et al., 2016), HardNet (Mishchuk et al., 2017), SuperPoint (DeTone et al., 2018), D2-Net (Dusmanu et al., 2019), LoFTR (Sun et al., 2021), etc. For multimodal images, Hughes et al. (2020) developed a three-stage convolutional neural network framework for optical-SAR registration; Quan et al. (2022) proposed a self-distillation feature learning network called SDNet. However, each of these methods can only applicable to one specific image type. As aforementioned, the main obstacle is the lack of public data.

In this paper, we focus on both radiometric and geometric differences of multimodal images via the SRIF algorithm, and collect an open dataset with different types of multimodal images to promote the development of learning-based matching.

3. Multimodal image dataset

As aforementioned, an important factor hindering the successful application of deep learning techniques in multimodal matching is the lack of public data on different types of multimodal images. Here, we collect and create a multimodal image dataset with six typical types of images, i.e, optical-optical, optical-infrared, optical-SAR, optical-depth, optical-map, and nighttime, and make it open to the community. We hope it can promote the development of multimodal image matching.

3.1. Optical-Optical

We use the WHU building dataset (Ji et al., 2018) to produce our Optical-Optical dataset, which consists of a set of pre-registered multi-temporal aerial images covered over the Christchurch, New Zealand. Since these images were captured in 2012 and 2016, the objects, textures, and colors red changed dramatically between the two images of a matching pair. We found that the pre-registration is not very accurate. Thus, we use a coarse-to-fine strategy, i.e., LNIFT + CFOG, to refine the registration. These images are then cropped into subimages of 512 × 512 pixels. We randomly generate a ground-truth transformation with a rotation angle $\mathrm{a} \in [0°, 90°)$ and a scale factor $s \in [0.5, 2)$. This transformation is applied to the target image to obtain our Optical-Optical dataset.

3.2. Optical-Infrared

We produce the Optical-Infrared dataset based on Ye et al. (2022), in which the original images are obtained from Landsat-8 satellite images covered over Chengdu Plain and surrounding hills and mountains. In addition to the band differences (optical is band 2 and infrared is band 5), there are also temporal differences in this dataset since optical images were captured in 2020 while infrared ones were acquired in 2021. Optical and infrared images are accurately aligned based on the geometric correction technique. For each pair, we add a random rotation transformation to the target image, i.e., rotate the target image counterclockwise by an angle $\mathrm{a} \in [0°, 90°)$ with the midpoint of the target image as the rotation center.

3.3. Optical-SAR

The Optical-SAR dataset is created based on the dataset 2 of LNIFT, in which the SAR image is acquired by the GaoFen-3 SAR satellite and the optical image is obtained from Google Earth. This dataset covers 15 cities including Beijing, Rennes, Omaha, Dwarka, etc. Since there is already a rotation change between each image pair, we only add a random scale factor $s \in [0.5, 2)$ to each pair to get the final Optical-SAR dataset.

3.4. Optical-Depth

The indoor DIML/CVL RGB-D (Cho et al., 2021) dataset is used to produce our Optical-Depth dataset, which is acquired by a Microsoft Kinect v2 camera. Image acquisition scenes mainly include offices, bedrooms, shopping malls, and exhibition centers in South Korea. The same as the Optical-Infrared dataset, we also add a random rotation with an angle $\mathrm{a} \in [0°, 90°)$ to the target image.

3.5. Optical-Map

This dataset is collected by Ye et al. (2022), which is obtained from the Google map service. The location is in Tokyo and the object features are mainly buildings and streets. We also use a strategy of LNIFT + CFOG to refine the registration. Then, the images are resized to 400 × 400 pixels to get the final dataset. We do not add rotation and scale differences due to the extremely large NRDs between optical images and maps. Thus, the ground truth transformation of each image pair is an identity matrix.

3.6. Nighttime

The Nighttime dataset is constructed based on the LLVIP (Jia et al., 2021) dataset, whose images are captured by a binocular camera from 26 different scene locations. It is not only affected by low-light conditions (acquired at night), but also by sensor differences (a visible camera and a thermal infrared camera). We also add random rotations to this dataset.

Table 1 summarizes detailed information about each dataset, including dataset size, image size, resolution, geometric changes, radiation changes, etc. Fig. 1 shows example data for these six datasets.

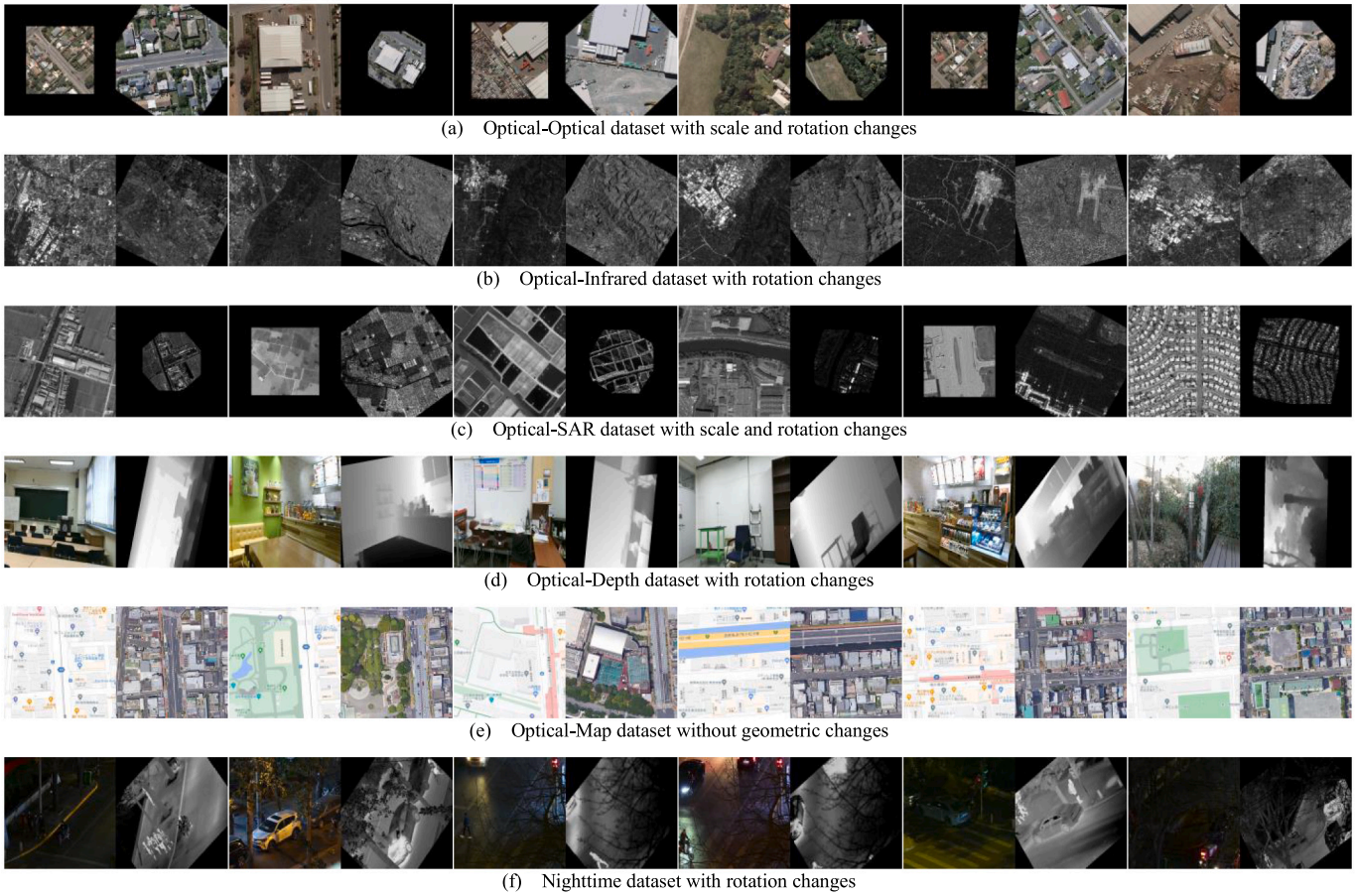
4. Our SRIF

Before describing our SRIF algorithm in detail, we first introduce the pyramid scale space strategy used by SRIF and provide experimental support for this choice. Then, we give the definition and calculation of LIBT. Because these two points are the key differences between the proposed SRIF and existing methods such as RIFT and LNIFT.

4.1. Pyramid scale space

4.1.1. Small scale sensitivity

We conduct an experiment on our Optical-Optical dataset to reveal the sensitivity of different matching methods to small scale changes. For each optical image, we use itself as the reference image and its scaled



(a) Optical-Optical dataset with scale and rotation changes

(b) Optical-Infrared dataset with rotation changes

(c) Optical-SAR dataset with scale and rotation changes

(d) Optical-Depth dataset with rotation changes

(e) Optical-Map dataset without geometric changes

(f) Nighttime dataset with rotation changes

Fig. 1. Sample data of our collected multimodal image dataset. Our dataset consists of six typical types of multimodal images, including optical-optical, optical-infrared, optical-SAR, optical-depth, optical-map, and nighttime images.

Table 1
Detailed information of our multimodal image dataset.

Dataset	Description	Dataset size	Image size	Resolution	NRDs	Rotation change	Scale change
Optical-Optical	Aerial image	200 pairs	256 × 256 to 1024 × 1024	4 cm to 16 cm	✓	✓	✓
Optical-Infrared	Satellite image	200 pairs	256 × 256	30 m	✓	✓	×
Optical-SAR	Satellite image	200 pairs	128 × 128 to 512 × 512	0.5 m to 2 m	✓	✓	✓
Optical-Depth	Close range image	200 pairs	512 × 288	Unknown	✓	✓	×
Optical-Map	Satellite image	200 pairs	400 × 400	1 m	✓	×	×
Nighttime	Close range image	200 pairs	320 × 240	Unknown	✓	✓	×

image as the target one to construct a matching pair. The scale factor is set to be {1, 1.1, 1.2, 1.3}. That is to say, any matching pair only has a small scale difference, excluding other geometric and radiation differences. We choose the SIFT, RIFT, and LNIFT for comparison, in which SIFT has a scale space while others do not. SIFT is the most widely used image matching algorithm. RIFT and LNIFT are the state-of-the-art SIFT-like methods for multimodal image matching. These three methods also offer access to the source code, thereby facilitating experimental procedures. To eliminate the influence of other factors, we first disable their dominant orientation calculation modules and set the main orientation to 0, which is the true value; then, we remove the nearest neighbor distance ratio strategy since it may discard true matches. We use the correct matching ratio (CMR) as the evaluation metric, which is the ratio of correct matches to total matches. The results are displayed in Fig. 2.

As shown, when the scale varies between 1 and 1.2, SIFT performs comparable to LNIFT, while RIFT even outperforms SIFT. However, once the scale difference reaches 1.3, the CMRs of RIFT and LNIFT are lower than that of SIFT. Therefore, we can infer a conclusion that when the scale factor is small, such as less than 1.2, the Gaussian scale space has very limited improvement in matching performance. This also motivates us to achieve scale invariance through the downsampling process without multi-scale Gaussian filtering. It is just that the downsampling factor must be small. In fact, this conclusion coincides with the scale space idea of ORB (Rublee et al., 2011) algorithm. Moreover, the default downsampling factor of the ORB algorithm is also 1.2. The differences between our projection-based scale space and the one of ORB are: First, we only construct a scale space for the target image while keeping the reference image unchanged. Second, we only detect keypoints in the original images and project them onto the layers of the pyramid scale space. Moreover, we provide the rationality of projection-based scale space based on experiments. While the above conclusion is important, it is also important to understand the reasons behind the conclusion. For example, artificial intelligence scholars regard the interpretability of deep learning as one of the important problems to be solved.

4.1.2. Projection-based scale space

Fig. 3 shows the details of the scale invariance strategy of our SRIF. It can be seen that SRIF only constructs the pyramid scale space for

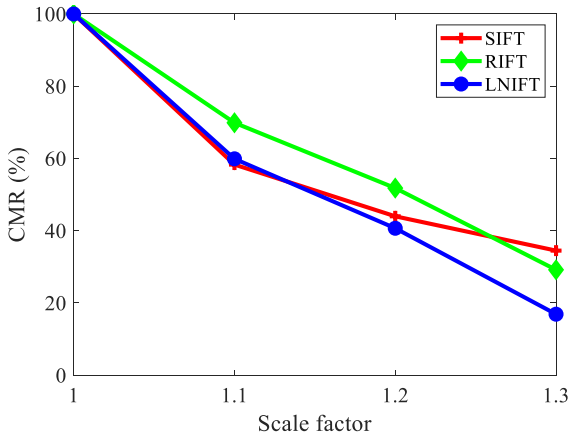


Fig. 2. The sensitivity of SIFT, RIFT, and LNIFT to small scale variances.

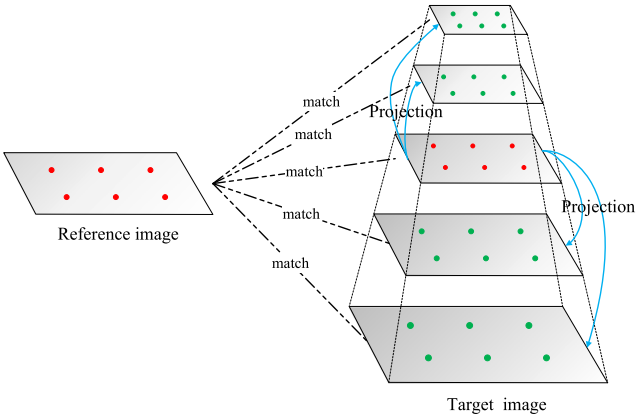


Fig. 3. The projection-based scale space of SRIF.

the target image. The pyramid layers are obtained via simultaneously upsampling and downsampling the original target image. To some extent, our method has some similarities with SI-PIIFD (Du et al., 2018), MS-PIIFD (Gao and Li, 2021), and MS-HLMO (Gao et al., 2022), because these methods are all improved on the basis of Gaussian scale space framework. Different from MS-PIIFD (Gao and Li, 2021) and MS-HLMO (Gao et al., 2022), we omit the multiscale Gaussian filtering and difference operation steps, while they preserve the full Gaussian scale space. Although both our method and SI-PIIFD (Du et al., 2018) can be considered as sampling-based methods, SI-PIIFD is based on additive operations while our method is based on multiplicative operations. In addition, our method only constructs pyramid scale space on the target image while these three methods construct scale spaces on both the reference and the target images. Assuming that the number of upsampling/downsampling operations is K, the scale factor is s, and the image size of the original target image is $[h, w]$, then, the pyramid contains a total of $2K + 1$ layers and the image size of the ith ($i \in \{1, 2, \ldots, 2K + 1\}$) layer is $[w_i, h_i]$,

$$[w_i, h_i] = s^{K+1-i}\,[w, h] \tag{1}$$

Note that the first layer is at the bottom of the pyramid. Different from the ORB and SIFT, we only detect keypoints in the original target image and project these keypoints onto the pyramid layers to obtain multiscale keypoints. For downsampled pyramid layers, we only project some randomly selected features. The purpose of this is to avoid that the distances between features are too small, resulting in large overlaps between local image patches used for feature description and thus interfering with the subsequent matching process. The number of projected keypoints of each layer N_i is,

$$N_i = \begin{cases} N & i \le K + 1 \\ \frac{N}{s^{2(i-K-1)}} & i > K + 1 \end{cases} \tag{2}$$

where N is the number of keypoints in the original target image.

To adapt to our scale space strategy, SRIF also modifies the traditional nearest neighbor matching strategy. Traditional matching strategy first merges the feature points on each pyramid layer to obtain the total feature set. Then, for a feature in the reference image, the matching strategy searches for the best match in the total set. In contrast, we do not perform a merge operation. We first search for the best match for the feature in each layer of the pyramid scale space of the target image, and then search for the best match among these $2K + 1$ features as the correspondence for that feature. This two-level matching strategy can effectively reduce the matching search space. At the same time, due to the reduced search space, the possibility of match ambiguity is also reduced.

4.2. Local Intensity Binary Transform (LIBT)

Compared with conventional matching, the bottleneck of multimodal image matching lies in the severe NRDs. Although commonly used intensity and gradient information are sensitive to NRDs, fortunately, numerous studies have shown that structural and shape features are very important information for multimodal image matching (Heinrich et al., 2011; Li et al., 2015; Ye et al., 2017, 2019; Li et al., 2020a), since they are preserved across different modalities and relatively independent of radiation changes. Therefore, the core idea of this paper is to enhance the structural information in the image through some transformations, and then use a structural descriptor such as the HOG-like or SIFT-like descriptor for feature description.

4.2.1. LIOT

Recently, Shi et al. (2022) proposed a local intensity order transformation (LIOT) to enhance the structures of an image. LIOT uses the relative intensity order to characterize the intrinsic properties of the curvilinear structures without relying on absolute intensity values. Thus, it not only enhances the structure information, but also has good contrast invariance. Experimental results show that LIOT can effectively improve the performance of state-of-the-art methods for tasks such as retinal vessel segmentation and crack segmentation.

Here, we briefly describe the basic idea of LIOT. As shown in Fig. 4(a), for a pixel p of an image I (the red pixel in the figure), LIOT compares the intensity $I(p)$ with intensities of its 8 neighboring pixels $\{q_i^c | i = 1, \ldots, 8\}$ along one of the four directions $c \in \{l, r, u, b\}$, where l, r, u, b represent left, right, up, and below, respectively. Each direction can generate an 8-bit 2D image $\bar{I}^c$, where the intensity $\bar{I}^c(p)$ is calculated based on the binary code obtained by the intensity order between $I(p)$ and $I(q_i^c)$. Formally, the calculation equation is as follows,

$$\bar{I}^c(p) = \sum_{i=1}^{8} \mathbb{1}[I(p) > I(q_i^c)] \cdot 2^{i-1} \tag{3}$$

where $\mathbb{1}[x]$ is an indicator function that returns 1 if event x is true and 0 otherwise.

It can be seen that LIOT is highly correlated with orientations, while rotation invariance is an important property for feature matching. Furthermore, LIOT generates a four-channel image, while only a single-channel image is required for feature description. These prevent LIOT from being used for the image matching task.

4.2.2. LIBT

To address the above two issues, we propose a variant of LIOT, called LIBT. The illustration of LIBT is shown in Fig. 4(b). We discard the orientation-dependent pixel order strategy of LIOT, and directly

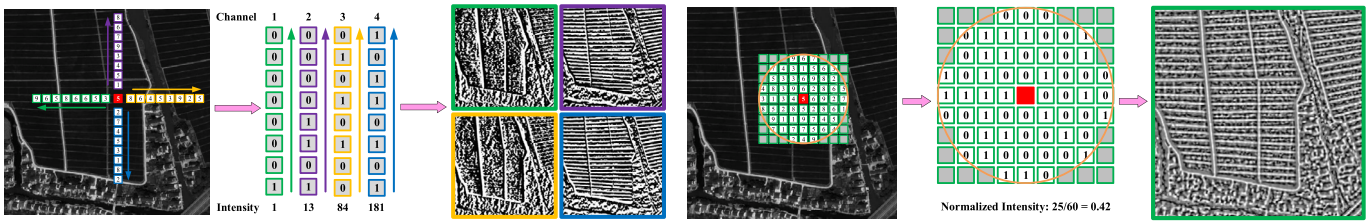


Fig. 4. Illustration of LIOT and LIBT. Both LIOT and LIBT use relative intensity to encode a new image, which are robust to contrast changes and can effectively enhance the structure information. LIOT converts an image into a 4-channel one and is sensitive to rotation changes. LIBT overcomes the limitations of LIOT and converts an image into a single-channel normalized one. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Image similarity comparison based on the LPIPS↓ measure. **Bold** fonts denote the best.

Dataset	Method		
	Original image	LNI	LIBT
Optical-Optical	0.37	0.35	**0.26**
Optical-Infrared	0.47	0.39	**0.22**
Optical-SAR	0.65	0.51	**0.36**
Optical-Depth	0.63	**0.43**	0.46
Optical-Map	0.78	0.61	**0.49**
Nighttime	0.51	0.44	**0.33**

compare the intensity $I(p)$ with all pixels in its circular region $\boldsymbol{\Phi}_p$ to generate a binary code. Then, the proportion of non-zero elements is counted as the normalized intensity value of $\bar{I}(p)$. Obviously, both the circular region and the proportion of non-zero elements are rotationally invariant. Thus, the formula of our LIBT is as follows,

$$\bar{I}(p) = \frac{1}{\left|\boldsymbol{\Phi}_p\right|} \sum_{i \in \boldsymbol{\Phi}_p} \mathbb{1}[I(p) > I(q_i)] \tag{4}$$

where $\left|\boldsymbol{\Phi}_p\right|$ is the total number of pixels in the region $\boldsymbol{\Phi}_p$.

As aforementioned, we hope to enhance the structural information through LIBT, thereby reducing the difficulty of multimodal image matching. Generally, the similarity between images is inversely proportional to the matching difficulty. The matching difficulty here mainly refers to the severity of NRDs. Therefore, if LIBT can effectively reduce the matching difficulty, the similarity of the transformed images should be higher. To verify this conclusion, an experiment on our collected multimodal dataset with 1200 image pairs is conducted. In this experiment, we use the ground-truth transformation to accurately register each image pair for a more convenient image similarity comparison. We use the learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018) as an evaluation metric, which is used to measure the difference between two images. A lower value of LPIPS means that the two images are more similar. LPIPS is more in line with human perception than traditional methods, such as the structural similarity index measure (SSIM) (Wang et al., 2004), correlation coefficient, feature similarity index measure (FSIM) (Zhang et al., 2011), etc. Table 2 reports the LPIPS results of original images without any transformation, with local normalized image (LNI) transformation (Li et al., 2022b), and with the proposed LIBT transformation. From the results, we can draw the following conclusions: (1) the LPIPS metric can truly reflect the similarity between images. For example, the LPIPS values of Optical-SAR, Optical-Depth, and Optical-Map are larger than others, which is very consistent with our human perception, because the differences between these three types of images are significantly larger than those of the other types. (2) Both LNI and our LIBT can improve the similarity of original images, and LIBT is much better than LNI. As can be seen, LIBT achieves the best results in five out of the six datasets and is only slightly worse than LNI in the Optical-Depth dataset. The average LPIPS values of original images, LNI, and our LIBT are 0.57, 0.46, and 0.35, respectively.

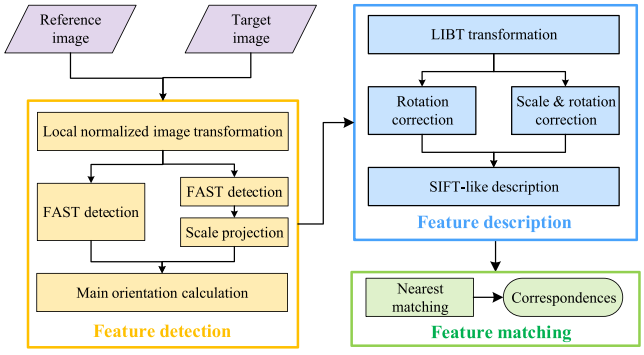


Fig. 5. The main framework of our SRIF.

4.3. Main framework of SRIF

The main framework of our SRIF is given in Fig. 5, which also contains three main stages, including feature detection, feature description, and matching. In this paper, we only focus on the first two stages.

4.3.1. Feature detection

We first convert original images into transformed ones based on the proposed LIBT and detect FAST features. As pointed out by Li et al. (2022b), FAST features are prone to aggregation. Thus, to obtain evenly distributed keypoints, we also use an adaptive non-maximal suppression strategy to suppress clustered features. We then project the features of the target image into a pyramid scale space to achieve scale invariance. It can be seen that we first detect feature points and then perform scale-space projection, which is completely opposite to the steps of traditional methods. Finally, we achieve rotation invariance in the same way as SIFT. Specifically, we use the gradient orientation histogram technique to obtain the histogram maximum and local extrema, and all local extrema greater than 80% of the maximum are taken as the main orientation of a feature. Based on the experimental conclusion in Section 4.2.2, we can replace LIBT with LNI to achieve better results on the Optical-Depth dataset.

4.3.2. Feature description

Similarly, we also perform feature description on the LIBT-transformed images. First, SRIF computes the gradient maps of the LIBT images and normalizes the orientations into $[0°, 180°)$ since multimodal images often have reversed orientations. Then, local image patches corresponding to the feature points are cropped. The reference patches only need to be rotated, while the target patches need to be rotated and scaled at the same time. All local patches are resized into the same size to facilitate subsequent feature description. Then, we use a SIFT-like descriptor for feature vector encoding. The descriptor first divides a local patch with $J \times J$ pixels into $N_{grid} \times N_{grid}$ grids. We compute a

Table 3
Detailed settings of compared algorithms.

Method	Main parameters	Implementations	Rotation invariance	Scale invariance
SIFT	Keypoint number: 5000; patch size: $15\sqrt{2}$ scale; contrast threshold: 0.001; number of grids: 4 × 4; orientation bins: 8	C++ code: https://www.vlfeat.org/overview/sift.html	✓	✓
OS-SIFT	Harris function threshold: 0.001; Keypoint number: 5000; patch size: 24 scale; number of circle grids: 8; orientation bins: 8	MATLAB code: https://sites.google.com/view/yumingxiang	✓	✓
RIFT	Keypoint number: 5000; patch size: 96; FAST threshold: 0.001; number of grids: 6 × 6; orientation bins: 6	MATLAB code: https://ljy-rs.github.io/web/	✓	×
3MRS	Template window size: 101; standard deviation of 2D Gaussian kernel: 0.5	C++ code: https://github.com/ZhongLi-Fan/3MRS	×	×
CoFSM	Co-occurrence filter window size: 5; scale layers: 4; feature threshold: 500; number of grids: 19; orientation bins: 8	MATLAB code: https://github.com/yyxgiser/CoFSM	✓	✓
LNIFT	Keypoint number: 5000; patch size: 96; FAST threshold: 0.001; number of grids: 8 × 8; orientation bins: 4	C++ code: https://ljy-rs.github.io/web/	✓	×
MS-HLMO	Keypoint number: 5000; patch size: 96; pyramid octaves: 3; pyramid layers: 4; number of grids: 12 × 12; orientation bins: 12	MATLAB code: https://github.com/MrPingQi	✓	✓
Our SRIF	Keypoint number: 5000; patch size: 96; FAST threshold: 0.001; $K = 3$; number of grids: 8 × 8; orientation bins: 4	C++ code: https://github.com/LJY-RS/SRIF	✓	✓

N_{bin}-histogram for each grid and obtain a total of $N_{grid} \times N_{grid}$ histograms. These histograms are then concatenated together to get a feature vector with length $N_{grid} \times N_{grid} \times N_{bin}$, which is then normalized to improve the robustness to illumination changes. Actually, our description method is the same as LNIFT, but the layers used for description are different. We use LIBT while LNIFT uses LNI. Therefore, we use the same parameters as LNIFT, i.e., $N_{grid} = 8$ and $N_{bin} = 4$.

5. Experiments

Here, we comprehensively evaluate the proposed SRIF on our collected multimodal datasets with a total of 1200 pairs. Our SRIF is compared with seven baseline or state-of-the-art algorithms, i.e., SIFT (Lowe, 2004), OS-SIFT (Xiang et al., 2018), RIFT (Li et al., 2020a), 3MRS (Fan et al., 2022), CoFSM (Yao et al., 2022), LNIFT (Li et al., 2022b), and MS-HLMO (Gao et al., 2022). The official implementation of each method is used in the experiments. For a fair comparison, we set the maximum number of features to 5000 and apply the same matching strategy (using brute force searching to establish one-to-one correspondence without a nearest neighbor distance ratio (NNDR) test) for all compared methods except 3MRS, CoFSM, and MS-HLMO, since 3MRS, CoFSM, and MS-HLMO only provide binary code and are difficult to modify. Table 3 summarizes the parameter settings, implementation details, and invariance properties of each method.

Three measures are used for quantitative evaluation, i.e., correct matching number n, root mean square error (RMSE) r, and success rate γ. Note that we do not apply RANSAC-like methods or local geometric constraints to filter outliers before evaluation for all methods, since our goal is to evaluate local descriptors while these outlier removal methods may discard some true inliers. The definitions of these three measures are as follows:

- **Correct match number** n: The number of correct correspondences in an image pair. If the residual of a correspondence under ground truth transformation is smaller than ε pixels ($\varepsilon = 3$), it is accepted as a correct one.
- **Success rate** γ: If the correct match number of an image pair satisfies $n \geq 10$, the image pair is considered to have been successfully matched, since a too small n will cause subsequent model fitting failure.
- **RMSE** r: Suppose $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_1^n$ are correct correspondences of an image pair, $T(\cdot)$ represents its ground truth transformation, the RMSE is computed by,

$$r = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i - T(\boldsymbol{x}_i)\right)^2}, \tag{5}$$

For those images that failed to match ($n < 10$), we set their RMSEs to 20 pixels.

5.1. Qualitative evaluations

The first pair of each type of multimodal images displayed in Fig. 1 is used for comparison. The first image pair suffers from temporal, scale, and rotation changes; the second pair has band and rotation differences; the third pair suffers from severe speckle noise, scale, and rotation changes; the fourth one has a huge difference in imaging mechanism and a rotation change; the map of the fifth pair is not really an image; and the last one suffers from low-light condition, sensor difference, and a rotation change. Fig. 6 shows the result of each compared algorithm.

As can be seen, SIFT and OS-SIFT achieve successful matching only on the Optical-Optical pair, but the correct match number is very low. SIFT uses gradients for description, which has been shown to be very sensitive to NRDs. OS-SIFT is designed for Optical-SAR matching and not suitable for other types of multimodal images with severe NRDs. RIFT and LNIFT perform well on image pairs without scale changes since they do not achieve scale invariance. 3MRS has the worst results, i.e., it fails to match on all pairs. 3MRS uses a coarse-to-fine strategy for registration, which is not robust to either scale or rotation change. Hence, 3MRS performs very badly when an image pair suffers from large geometric changes such as rotation, scale, perspective, etc. CoFSM and MS-HLMO perform well only on the first pair. Although these methods achieve good results on some datasets, they perform poorly on our collected dataset. One possible reason is that our dataset is more difficult. For example, the scale and rotation changes do not appear simultaneously on the same image pair in their experiments. In contrast, our SRIF achieves the best results on all image pairs. Our correct match number n is much larger than other methods such as RIFT and LNIFT.

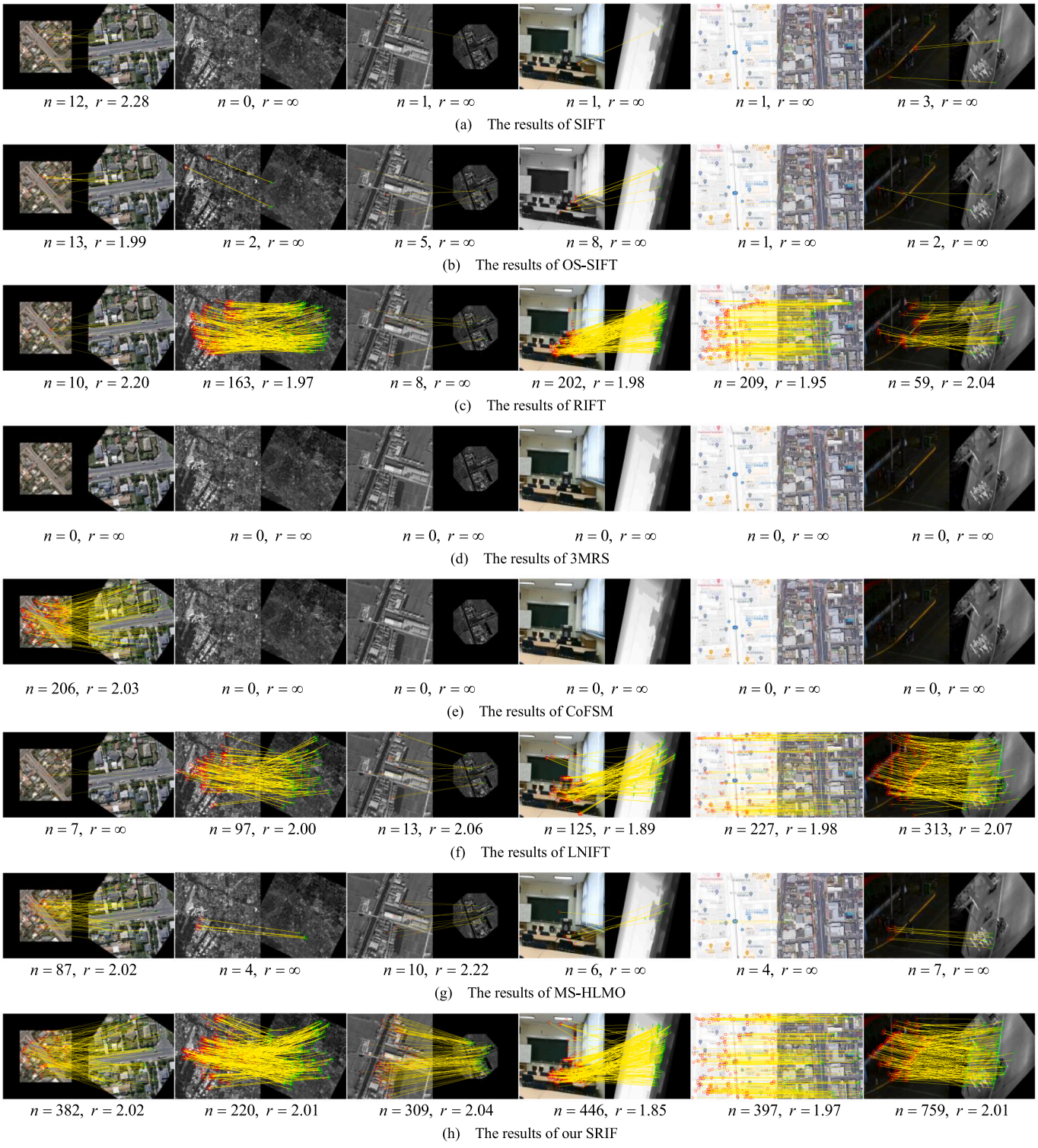


Fig. 6. Qualitative comparison results on the first pair of each type of sample data in Fig. 1. Keypoints are shown as red circles and green crosshairs; correct correspondences are represented as yellow lines. If the matching fails, its RMSE is denoted by $r = \infty$. We only display no more than 200 correspondences in each pair for better visualization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2. Quantitative evaluations

The quantitative results are reported in Table 4, including correct match number n (higher is better), RMSE r (lower is better), and success rate γ (higher is better). As shown, SIFT is only suitable for optical images, which have less NRDs. Its success rate is close to 0 on other types of multimodal datasets. This is predictable since the SIFT algorithm is commonly used for same-source image matching. OS-SIFT performs slightly better than SIFT, since it modifies the calculation of gradients to improve the robustness to speckle noise. However, it gets very poor performance on the Optical-Infrared and Optical-Map datasets since the major matching difficulty of these datasets is not speckle noise. Even on the SAR-Optical dataset, the performance of OS-SIFT is still far from good. The main reason is that OS-SIFT does not

Table 4
Quantitative evaluation results. Each value is the average result.

Data	Metric	Method							
		SIFT	OS-SIFT	RIFT	3MRS	CoFSM	LNIFT	MS-HLMO	Our SRIF
Optical-Optical	RMSE r (pixels)↓	6.03	5.03	5.65	18.04	4.54	6.56	5.03	**2.05**
	Success rate γ (%)↑	77.5	83.5	80	11	84	75	83	**100**
	Correct match number n ↑	30	44	127	39	144	56	109	**415**
Optical-Infrared	RMSE r (pixels)↓	20	20	3.20	16.62	19.55	2.52	18.94	**2.07**
	Success rate γ (%)↑	0	0	93.5	18.5	2.5	97.5	6	**100**
	Correct match number n ↑	1	2	72	91	2	77	2	**185**
Optical-SAR	RMSE r (pixels)↓	20	15.96	9.62	18.66	15.34	7.36	14.26	**2.07**
	Success rate γ (%)↑	0	22.5	58	7.5	25.5	70.5	32.5	**100**
	Correct match number n ↑	1	7	32	16	13	59	14	**309**
Optical-Depth	RMSE r (pixels)↓	19.63	13.44	2.51	16.81	15.71	2.00	13.15	**1.92**
	Success rate γ (%)↑	2	36	97	17.5	23.5	**100**	38	**100**
	Correct match number n ↑	2	10	227	265	28	194	18	**378**
Optical-Map	RMSE r (pixels)↓	20	19.19	2.30	14.10	19.73	2.07	19.28	**1.99**
	Success rate γ (%)↑	0	1.5	98	33	4.5	**100**	4.5	**100**
	Correct match number n ↑	0	3	130	15	4	78	3	**332**
Nighttime	RMSE r (pixels)↓	19.46	18.02	2.85	14.95	15.38	2.08	14.96	**2.04**
	Success rate γ (%)↑	3	11	95.5	28	25.5	**100**	28.5	**100**
	Correct match number n ↑	2	5	119	169	34	225	9	**488**
Average	RMSE r (pixels)↓	17.52	15.36	4.36	16.53	14.94	3.77	14.27	**2.02**
	Success rate γ (%)↑	13.75	25.75	87	19.25	27.58	90.5	32.08	**100**
	Correct match number n ↑	6	12	118	99	38	115	26	**351**

Table 5
The results of ablation study.

Pipeline	Baseline (BL)	Baseline* (BL*)	Rotation invariance (RI)	Scale invariance (SI)	RMSE r (pixels)↓	Success rate γ (%)↑	Correct match number n ↑
BL	✓	×	×	×	11.44	47.50	89
BL+RI	✓	×	✓	×	4.56	85.83	161
BL+SI	✓	×	×	✓	10.25	60.33	141
BL*+RI+SI (SRIF_LNI)	×	✓	✓	✓	2.83	95.50	262
BL+RI+SI (our SRIF)	✓	×	✓	✓	**2.02**	**100**	**351**

attenuate the NRDs between optical and SAR images. 3MRS does not achieve more than a 40% success rate on any one dataset. The reason is that it does not consider rotation and scale changes between images, while most image pairs in our dataset suffer from rotation variance or scale change. We also observe that despite its low success rate, 3MRS still has a high average correct match number. For example, it ranks second on the Optical-Infrared and Optical-Depth. This is because 3MRS uses a dense template matching method to refine the coarse feature matching results. CoFSM and MS-HLMO perform well on the Optical-Optical dataset and are comparable to OS-SIFT on other datasets. The unsatisfactory performance of CoFSM and MS-HLMO may be due to the high matching difficulty of our dataset. In the original papers of CoFSM and MS-HLMO, they were only tested on small-scale datasets without complex geometric changes. RIFT and LNIFT achieve very high success rates (>90%) on the datasets without scale changes. Even on datasets with scale variance such as the Optical-SAR dataset, their success rate is still higher than 50%. As shown in Section 4.2.2, although RIFT and LNIFT do not construct a scale space, they still have a certain ability to resist scale changes, especially for small scale factors. In our datasets, the scale factor is between 0.5 and 2, which is not very large. Our SIRF gets a success rate of 100% and more than 300 correct matches. There are two main reasons why our SRIF performs so well: first, our method achieves rotation and scale invariance, so it can cope with complex geometric differences; second, we propose the LIBT transformation to enhance the structural information in the images, thereby greatly reducing the NRDs. LIBT outperforms LNI, which enables SRIF to obtain better matching results than LNIFT.

The average success rates of these eight compared algorithms on these six datasets are 13.75%, 25.75%, 87%, 19.25%, 27.58%, 90.5%, 32.08%, and 100%, respectively. Our SRIF gains a growth rate of 10% compared with the second best method, i.e., LNIFT. In terms of correct match number n, the results of SIFT, OS-SIFT, RIFT, 3MRS, CoFSM, LNIFT, MS-HLMO, and our SRIF are 6, 12, 118, 99, 38, 115, 26, and 351, respectively. Our correct matches are three times that of RIFT and LNIFT, and about 10 times that of CoFSM and MS-HLMO. Our RMSE is around 2 pixels at an inlier threshold of 3 pixels. It is slightly better than LNIFT on the Optical-Depth, Optical-Map, and Nighttime datasets, on which the success rates of LNIFT are also 100%. This matching accuracy is sufficient for many remote sensing applications. As known, the matching accuracy of template-based methods is generally better than feature-based ones. If we want to use it in applications that require very high geometric accuracy, a template-based matching algorithm such as the CFOG can be applied to further refine our results.

5.3. Ablation study

To demonstrate the effectiveness of the key novel steps in the proposed SRIF algorithm, we conduct an ablation experiment on our multimodal image dataset. This experiment compares five different pipeline settings (BL, RI, SI represent baseline, rotation invariance module, and scale invariance module, respectively):

- **BL:** Remove the rotation invariance and scale invariance modules from our SRIF algorithm;
- **BL+RI:** Only remove the scale invariance module from our SRIF algorithm;
- **BL+SI:** Only remove the rotation invariance module from our SRIF algorithm;
- **BL*+RI+SI:** Only replace the LIBT in our SRIF algorithm with the local normalized image (LNI) to generate structural feature maps;
- **BL+RI+SI:** The proposed SRIF algorithm.

The average quantitative experimental results on the 1200 image pairs are reported in Table 5. As shown, (1) comparing BL+RI (or

Table 6
Running time analysis.

Method	Image size (pixel)			
	256 × 256	512 × 512	768 × 768	1024 × 1024
SIFT	0.22	0.82	1.97	3.90
OS-SIFT	0.50	3.39	11.12	28.43
RIFT	2.94	21.16	33.40	49.78
3MRS	0.76	2.45	3.46	4.91
CoFSM	4.55	9.02	24.13	54.23
LNIFT	0.36	0.39	0.44	0.48
MS-HLMO	27.86	169.10	191.54	203.71
SRIF	3.06	3.13	3.17	3.19

BL+SI) with BL, we can see that rotation invariance module and scale invariance module can improve the success rate and correct match number by a large margin; (2) comparing BL+RI+SI (our SRIF) with BL*+RI+SI (SRIF_LNI), we can see that the LIBT can generate much better structural feature map than LNI. The success rate increases by 4.5 percentage points, and the correct match number increases by more than 30%.

5.4. Computational time

An experiment is conducted to compare the computational time performance of these methods on the Optical-Depth dataset. The images are resized to 256 × 256, 512 × 512, 768 × 768, and 1024 × 1024 pixels to produce four datasets with different image sizes. This experiment is performed on a PC with a 3.6 GHz, 8 cores, i9-10850K CPU, and 64 GB of RAM, and the results are reported in Table 6.

The computational time of our SRIF is less dependent on the image size because the computational complexity of LIBT is very low compared to feature detection and description, which is very similar to LNIFT. Actually, the computational time of SRIF and LNIFT is mainly affected by the number of features, however, we fix it to 5000 and thus the computational time of SRIF is similar on images of different sizes. From the results, our SRIF runs much more efficiently than RIFT, CoFSM, and MS-HLMO when the image size is large. For example, SRIF is about 7 times faster than RIFT, 3 times faster than CoFSM, and 53 times faster than MS-HLMO on a 512 × 512 image. It ranks fourth and is only slower than SIFT, 3MRS, and LNIFT. When the image size is 1024 × 1024, SRIF is 15 times, 17 times, and 63 times faster than RIFT, CoFSM, and MS-HLMO, respectively, and ranks second among all 8 methods.

5.5. Limitations

The limitations of our SRIF mainly lie in twofold:

- **The computational complexity of SRIF is high compared with LNIFT**, which makes it unsuitable for real-time/near real-time matching tasks. Due to the way we construct the scale space, the number of features after projection on the target image is high. Furthermore, $K = 3$ can only cover the scale changes between 0.5 and 2. However, when the scale variation between images is large, we need to increase the value of K, but this will also increase the number of projected features and further reduce the efficiency of the algorithm. One possible solution is to use a coarse ground sample resolution prior to reduce the variance of scales or use GPU implementation for acceleration.
- **The correct matching ratio is low**, which is a common limitation of current multimodal feature matching methods. Although our SRIF can obtain many correct correspondences in the above experiments, it is based on extracting a large number of feature points, i.e., 5000 before projection. In fact, our correct matching ratio is still very low, mostly less than 10%. Thus, in practice, we need to increase the number of features to guarantee the matching performance, which will increase the computational complexity of SRIF. One possible solution is to learn a LIBT-like or a LNI-like image, which extracts the common features between multimodal images based on the powerful feature extraction capability of deep learning.

6. Conclusions

In this paper, we create a multimodal image dataset with six typical types of images, i.e, optical-optical, optical-infrared, optical-SAR, optical-depth, optical-map, and nighttime, and make it open to the community. This dataset contains a total of 1200 image pairs with good diversity in image categories, feature classes, resolutions, geometric variations, etc. We hope it will make a small contribution to the advancement of multimodal image matching, especially learning-based techniques. we also propose a scale and rotation invariant feature transform (SRIF) method for multimodal feature matching. We introduce a simple scale space construction strategy based on the analysis of a small scale sensitivity experiment. To enhance structural information to resist NRDs of multimodal images, we propose a local intensity binary transform (LIBT) for feature description and verify its effectiveness based on the LPIPS metric. By comparing with seven baseline and state-of-the-art algorithms on 1200 image pairs, we can see that our SRIF outperforms them by a large margin, i.e., our method gains a success rate of 10% and obtains three times of correct matches compared to the second best method. Our future work will focus on learning-based LIBT-like image generation and learning-based feature description.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 42030102 and 42271444, and the Science and Technology Major Project of Hubei Province under Grant 2021AAA010.

References

Aguilera, C.A., Sappa, A.D., Toledo, R., 2015. LGHD: A feature descriptor for matching across non-linear intensity variations. In: IEEE International Conference on Image Processing. IEEE, pp. 178–181.

Bas, S., Ok, A.O., 2021. A new productive framework for point-based matching of oblique aircraft and UAV-based images. Photogramm. Rec. 36 (175), 252–284.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-up robust features (SURF). Comput. Vis. Image Underst. 110 (3), 346–359.

Brown, M., Süsstrunk, S., 2011. Multi-spectral SIFT for scene category recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 177–184.

Chen, J., Tian, J., Lee, N., Zheng, J., Smith, R.T., Laine, A.F., 2010. A partial intensity invariant feature descriptor for multimodal retinal image registration. IEEE Trans. Biomed. Eng. 57 (7), 1707–1718.

Cho, J., Min, D., Kim, Y., Sohn, K., DIML/CVL RGB-D dataset: 2M RGB-D images of natural indoor and outdoor scenes, arXiv preprint arXiv:2110.11590, 1, 1–7.

Cui, S., Xu, M., Ma, A., Zhong, Y., 2020. Modality-free feature detector and descriptor for multimodal remote sensing image registration. Remote Sens. 12 (18), 2937.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, pp. 224–236.

Du, Q., Fan, A., Ma, Y., Fan, F., Huang, J., Mei, X., 2018. Infrared and visible image registration based on scale-invariant piifd feature and locality preserving matching. IEEE Access 6, 64107–64121.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T., 2019. D2-net: A trainable CNN for joint description and detection of local features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 8092–8101.

Fan, B., Huo, C., Pan, C., Kong, Q., 2012. Registration of optical and SAR satellite images by exploring the spatial relationship of the improved SIFT. IEEE Geosci. Remote Sens. Lett. 10 (4), 657–661.

Fan, Z., Liu, Y., Liu, Y., Zhang, L., Zhang, J., Sun, Y., Ai, H., 2022. 3MRS: An effective coarse-to-fine matching method for multimodal remote sensing imagery. Remote Sens. 14 (3), 478.

Fan, Z., Zhang, L., Liu, Y., Wang, Q., Zlatanova, S., 2021. Exploiting high geopositioning accuracy of SAR data to obtain accurate geometric orientation of optical satellite images. Remote Sens. 13 (17), 3535.

Fang, Y., Hu, J., Du, C., Liu, Z., Zhang, L., 2021. SAR-optical image matching by integrating siamese U-net with FFT correlation. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24 (6), 381–395.

Foroosh, H., Zerubia, J.B., Berthod, M., 2002. Extension of phase correlation to subpixel registration. IEEE Trans. Image Process. 11 (3), 188–200.

Fu, Z., Qin, Q., Luo, B., Wu, C., Sun, H., 2018. A local feature descriptor based on combination of structure and texture information for multispectral image matching. IEEE Geosci. Remote Sens. Lett. 16 (1), 100–104.

Gao, C., Li, W., 2021. Multi-scale PIIFD for registration of multi-source remote sensing images. J. Beijing Inst. Technol. 30 (2), 113–124.

Gao, C., Li, W., Tao, R., Du, Q., 2022. MS-HLMO: Multiscale histogram of local main orientation for remote sensing image registration. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Harris, C., Stephens, M., et al., 1988. A combined corner and edge detector. In: Alvey Vis. Conf.. 15 (50), 10–5244.

Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, J.M., Schnabel, J.A., 2011. Non-local shape descriptor: A new similarity metric for deformable multi-modal registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 541–548.

Huang, M., Xu, Y., Qian, L., Shi, W., Zhang, Y., Bao, W., Wang, N., Liu, X., Xiang, X., The QXS-SAROPT dataset for deep learning in SAR-optical data fusion, arXiv preprint arXiv:2103.08259, 1, 1–5.

Hughes, L.H., Marcos, D., Lobry, S., Tuia, D., Schmitt, M., 2020. A deep learning framework for matching of SAR and optical imagery. ISPRS J. Photogramm. Remote Sens. 169, 166–179.

Hughes, L.H., Schmitt, M., Zhu, X.X., 2018. Mining hard negative samples for SAR-optical image matching using generative adversarial networks. Remote Sens. 10 (10), 1552.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 57 (1), 574–586.

Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W., 2021. LLVIP: A visible-infrared paired dataset for low-light vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, pp. 3496–3504.

Li, J., Ai, M., Wang, S., Hu, Q., 2022a. GRF: guided residual fusion for pansharpening. Int. J. Remote Sens. 43 (10), 3609–3627.

Li, J., Hu, Q., Ai, M., 2016. Robust feature matching for remote sensing image registration based on l_{q} -estimator. IEEE Geosci. Remote Sens. Lett. 13 (12), 1989–1993.

Li, J., Hu, Q., Ai, M., 2017. Robust feature matching for geospatial images via an affine-invariant coordinate system. Photogramm. Rec. 32 (159), 317–331.

Li, J., Hu, Q., Ai, M., 2020a. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. IEEE Trans. Image Process. 29, 3296–3310.

Li, J., Hu, Q., Ai, M., 2020b. Robust geometric model estimation based on scaled welsch q-norm. IEEE Trans. Geosci. Remote Sens. 58 (8), 5908–5921. http://dx.doi.org/10.1109/TGRS.2020.2972982.

Li, J., Hu, Q., Ai, M., 2021a. Point cloud registration based on one-point RANSAC and scale-annealing biweight estimation. IEEE Trans. Geosci. Remote Sens. 59 (11), 9716–9729. http://dx.doi.org/10.1109/TGRS.2020.3045456.

Li, J., Hu, Q., Ai, M., Wang, S., 2021b. A geometric estimation technique based on adaptive M-estimators: Algorithm and applications. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 5613–5626. http://dx.doi.org/10.1109/JSTARS.2021.3078516.

Li, Z., Mahapatra, D., Tielbeek, J.A., Stoker, J., van Vliet, L.J., Vos, F.M., 2015. Image registration based on autocorrelation of local structure. IEEE Trans. Med. Imaging 35 (1), 63–75.

Li, J., Shi, P., Hu, Q., Zhang, Y., RIFT2: Speeding-up RIFT with a new rotation-invariance technique, arXiv preprint arXiv:2303.00319, 1, 1–5.

Li, J., Shi, P., Hu, Q., Zhang, Y., 2023b. QGORE: Quadratic-time guaranteed outlier removal for point cloud registration. IEEE Trans. Pattern Anal. Mach. Intell. 45 (9), 11136–11151. http://dx.doi.org/10.1109/TPAMI.2023.3262780.

Li, J., Xu, W., Shi, P., Zhang, Y., Hu, Q., 2022b. LNIFT: Locally normalized image for rotation invariant multimodal feature matching. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Liang, J., Liu, X., Huang, K., Li, X., Wang, D., Wang, X., 2013. Automatic registration of multisensor images using an integrated spatial and mutual information (SMI) metric. IEEE Trans. Geosci. Remote Sens. 52 (1), 603–615.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110.

Ma, J., Ma, Y., Li, C., 2019. Infrared and visible image fusion methods and applications: A survey. Inf. Fusion 45, 153–178.

Ma, W., Wen, Z., Wu, Y., Jiao, L., Gong, M., Zheng, Y., Liu, L., 2016. Remote sensing image registration with modified SIFT and enhanced feature matching. IEEE Geosci. Remote Sens. Lett. 14 (1), 3–7.

Merkle, N., Luo, W., Auer, S., Müller, R., Urtasun, R., 2017. Exploiting deep matching and SAR data for the geo-localization accuracy improvement of optical satellite images. Remote Sens. 9 (6), 586.

Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J., 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. Adv. Neural Inf. Process. Syst. 30, 1–12.

Mohammadi, N., Sedaghat, A., Jodeiri Rad, M., 2022. Rotation-invariant self-similarity descriptor for multi-temporal remote sensing image registration. Photogramm. Rec. 37 (177), 6–34.

Mur-Artal, R., Montiel, J.M.M., Tardos, J.D., 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Trans. Robot. 31 (5), 1147–1163.

Öfverstedt, J., Lindblad, J., Sladoje, N., 2022a. Cross-sim-NGF: FFT-based global rigid multimodal alignment of image volumes using normalized gradient fields. In: International Workshop on Biomedical Image Registration. Springer, pp. 156–165.

Öfverstedt, J., Lindblad, J., Sladoje, N., 2022b. Fast computation of mutual information in the frequency domain with applications to global multimodal image alignment. Pattern Recognit. Lett. 159, 196–203.

Parente, L., Chandler, J.H., Dixon, N., 2021. Automated registration of sfm-MVS multitemporal datasets using terrestrial and oblique aerial images. Photogramm. Rec. 36 (173), 12–35.

Quan, D., Wei, H., Wang, S., Lei, R., Duan, B., Li, Y., Hou, B., Jiao, L., 2022. Self-distillation feature learning network for optical and SAR image registration. IEEE Trans. Geosci. Remote Sens. 60, 1–18.

Rosten, E., Drummond, T., 2006. Machine learning for high-speed corner detection. In: European Conference on Computer Vision. Springer, pp. 430–443.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF. In: International Conference on Computer Vision. IEEE, pp. 2564–2571.

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

Sedaghat, A., Mohammadi, N., 2019. Illumination-robust remote sensing image matching based on oriented self-similarity. ISPRS J. Photogramm. Remote Sens. 153, 21–35.

Shechtman, E., Irani, M., 2007. Matching local self-similarities across images and videos. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.

Shi, T., Boutry, N., Xu, Y., Géraud, T., 2022. Local intensity order transformation for robust curvilinear object segmentation. IEEE Trans. Image Process. 31, 2557–2569.

Sui, H., Liu, C., Gan, Z., Jiang, Z., Chuan, X., 2022. Overview of multi-modal remote sensing image matching methods. Acta Geod. Cartogr. Sinica 51 (9), 1848–1861.

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

Tewkesbury, A.P., Comber, A.J., Tate, N.J., Lamb, A., Fisher, P.F., 2015. A critical synthesis of remotely sensed optical image change detection techniques. Remote Sens. Environ. 160, 1–14.

Viola, P., Wells III, W.M., 1997. Alignment by maximization of mutual information. Int. J. Comput. Vis. 24 (2), 137–154.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Xiang, Y., Jiao, N., Wang, F., You, H., 2021. A robust two-stage registration algorithm for large optical and SAR images. IEEE Trans. Geosci. Remote Sens. 60, 1–15.

Xiang, Y., Tao, R., Wang, F., You, H., Han, B., 2020. Automatic registration of optical and SAR images via improved phase congruency model. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 5847–5861.

Xiang, Y., Wang, F., You, H., 2018. OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas. IEEE Trans. Geosci. Remote Sens. 56 (6), 3078–3090.

Xiong, X., Jin, G., Xu, Q., Zhang, H., 2021. Self-similarity features for multimodal remote sensing image matching. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 12440–12454.

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

Ye, Y., Bruzzone, L., Shan, J., Bovolo, F., Zhu, Q., 2019. Fast and robust matching for multimodal remote sensing image registration. IEEE Trans. Geosci. Remote Sens. 57 (11), 9059–9070.

Ye, Y., Shan, J., Bruzzone, L., Shen, L., 2017. Robust registration of multimodal remote sensing images based on structural similarity. IEEE Trans. Geosci. Remote Sens. 55 (5), 2941–2958.

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

Yi, K.M., Trulls, E., Lepetit, V., Fua, P., 2016. LIFT: Learned invariant feature transform. In: European Conference on Computer Vision. Springer, pp. 467–483.

Yu, Q., Ni, D., Jiang, Y., Yan, Y., An, J., Sun, T., 2021. Universal SAR and optical image registration via a novel SIFT framework based on nonlinear diffusion and a polar spatial-frequency descriptor. ISPRS J. Photogramm. Remote Sens. 171, 1–17.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 586–595.

Zhang, H., Lei, L., Ni, W., Tang, T., Wu, J., Xiang, D., Kuang, G., 2020. Optical and SAR image matching using pixelwise deep dense features. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Zhang, H., Ni, W., Yan, W., Xiang, D., Wu, J., Yang, X., Bian, H., 2019. Registration of multimodal remote sensing image based on deep fully convolutional neural network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (8), 3028–3042.

Zhang, L., Zhang, L., Mou, X., Zhang, D., 2011. FSIM: A feature similarity index for image quality assessment. IEEE Trans. Image Process. 20 (8), 2378–2386.

Zhao, Y., Huang, X., Zhang, Z., 2021. Deep lucas-kanade homography for multimodal image alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 15950–15959.

Zhou, L., Ye, Y., Tang, T., Nan, K., Qin, Y., 2021. Robust matching for SAR and optical images using multiscale convolutional gradient features. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

View publication stats