PERFORMANCE EVALUATION OF LEARNING-BASED METHODS FOR MULTISPECTRAL SATELLITE IMAGE MATCHING

N. Jovhari [1]*, N. Farhadi [2], A. Sedaghat [1], N. Mohammadi [1,]

[1]Department of Geomatics Engineering, University of Tabriz, Tabriz, Iran - negarjovhari77@gmail.com
[2]Department of Geomatics and Remote Sensing, K. N. Toosi University of Technology, Tehran, Iran-
farhadinima75@email.kntu.ac.ir

Commission IV, WG IV/3

KEY WORDS: Multispectral Images, Image Registration, Feature Descriptors, Deep Learning, Geometric Differences, Illumination Variations

ABSTRACT:

Multispectral image registration is one of the most critical requirements to achieve reliable remote sensing goals such as change detection, image fusion, etc., due to providing complementary knowledge of the scene. On the one hand, this issue has always been a hot topic of research according to significant appearance differences, including geometric and nonlinear radiometric distortions. On the other hand, developing deep learning methods promises precise results in image processing and, in particular, image registration. It is no longer limited to low-level information structures, such as intensity and gradients. However, it is possible to provide more reliable results by extracting various high-level features and removing feature engineering. Therefore, we need extensive experiments in multispectral image registration to determine an efficient and robust method. To this end, this paper evaluates six well-known recently proposed learning-based feature descriptors, including LOFTR, TFeat, HardNet8, HardNet, SosNet, and HyNet, against geometric distortions within real multispectral images. Evaluations demonstrate the general superiority of the HardNet8 descriptor due to extracting high-level features within eight convolution layers.

1. INTRODUCTION

Deep learning methods have demonstrated promising results in the remote sensing and photogrammetry communities (Hosseiny et al., 2021), particularly image registration, which identifies corresponding entities within two or more images (Jiang et al., 2021). Accordingly, multispectral image registration as a specific case has attracted the attention of many researchers due to the diverse representation of the same scene and yielding reach information (Zhu et al., 2019). Different spectral responses of terrestrial objects in different spectral bands cause nonlinear radiometric differences between these images. In addition, researchers tend to employ high-resolution images due to advanced image technologies (Yan et al., 2022). Despite the valuable detail provided by multispectral images, multiple serious factors challenge the image registration performances and subsequent remote sensing products. Accordingly, high-level structural measurements and highly robust and distinct features are required. Otherwise, a slight error prevents achieving reliable results. Therefore, the more accurately the matching process is performed, the more likely it is to achieve an optimal control points network, and the more efficient and practical future processing will be (Liu et al., 2018; Ma et al., 2019; Yan et al., 2022).

Image matching methods are categorized into area-based and feature-based approaches. Generally, the area-methods require an approximate corresponding location and a predefined window for each point to be matched. However, a drawback of these methods is the lack of geometric robustness. The inlier numbers will be dramatically decreased in the presence of rotation and scale distortions. Moreover, noise and complex radiometric differences could also diminish the registration performance (Li et al., 2019).

The feature-based framework consists of three stages: local feature detection and description, correspondence identification, and blunder detection (Sedaghat & Mohammadi, 2018). A feature detector is applied to each image in the first step to extract robust local features. The most popular approach relates to scale-invariant methods, such as SIFT, which identifies three-dimensional extremes in space and scale by creating an image pyramid based on the DOG function (Lowe, 2004). Accordingly, circular features with the same content can be computed within images with scale differences.

Next, a descriptor vector is created to describe the area around the features extracted in a vector format. The designed descriptor created must be highly robust and distinctive. In other words, we need methods that effectively cope with geometric and radiometric distortions(Sedaghat & Mohammadi, 2019). Traditional methods employ intensity and gradient information. However, they fail to deal with significant illumination distortions (Alcantarilla et al., 2012; Ojala et al., 1994).

Recently, several studies have been conducted to increase the stability against radiometric variations of satellite images (Li et al., 2019; Ye et al., 2019). Despite the successful results in resampled images, the performance of many of these methods diminishes in the presence of significant geometric differences. These methods generally consider a universal criterion for encoding structural features and only within the primary scale image. Even in the case of circular detectors, if the designed descriptor is not scale-invariant, the shape property of the extracted features is practically ignored.

The integration of these problems has attracted the attention of many researchers to utilize deep neural networks to achieve stable and accurate matching. The main principle is to deal with dramatic geometric and radiometric differences by taking advantage of high-level features and identifying sufficient and

* Corresponding author

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-4-W1-2022-335-2023

reliable correspondences, especially in cases where handcrafted methods do not work appropriately.

Various learning-based methods have been proposed for image registration, generally divided into area-based and feature-based frameworks (Jiang et al., 2021). Similar to handcrafted methods, the framework-based framework is more popular due to its high robustness against geometric differences. A popular idea could be detecting local features by handcrafted detectors such as the well-known SIFT. Next, the images will be encoded via deep features to calculate the descriptor vector (Mishchuk et al., 2017). Also, (Yang et al., 2018) correct matches are identified gradually with the help of a multiscale descriptor and the VGG retrained network. Due to the lack of accessible sufficient training data, some researchers have employed GAN networks to generate additional samples (Quan et al., 2018). Another application of these networks is to create simulated samples of multimodal images to reduce illumination variance (Wang et al., 2018).

In addition to the lack of sufficient training data, geometric and illumination differences between satellite image pairs, especially multispectral pairs, are significant challenges for many algorithms. Many features have completely different spectral responses in various ranges of the electromagnetic wave spectrum; another issue is geometric distortions. Unlike conventional images, satellite images mainly consist of often have different directions and spatial resolutions. Such an issue defeats many traditional handcrafted methods and reduces the robustness and accuracy of many designed neural networks. In particular, many learning-based algorithms are defined for same-size images. Another issue is processing time. A proper tradeoff between efficiency and robustness must be provided, considering that co-registered images are not the ultimate products in many applications. To this point, this paper evaluates various state-of-the-art local feature descriptors against geometric distortions for multispectral image registration. Such a comprehensive evaluation is critical and helpful to carry out subsequent processing steps reliably and accurately.

2. PERFORMANCE EVALUATION

The major concentration of this study is to evaluate deep descriptors for multispectral image registration against geometric differences. Figure 1 illustrates the input images. All pairs are real multispectral images. They contain different land covers from urban areas with high buildings to the countryside, from low spatial resolution to high spatial resolution, and from the optical region of the electromagnetic spectrum to the infrared area.

To provide a comprehensive evaluation, we employed a set of well-known state-of-the-art deep descriptors, including LOFTR, TFeat, HardNet8, HardNet, SosNet, and HyNet. Figure 2 provides an overview of the designed pipeline.

Except for the first descriptor, the remaining ones must be integrated with feature detectors. Accordingly, we first encoded the input images by SIFT detector based on the Gaussian scale space. The Gaussian scale-space, as a linear approximation of the diffusion function, extracts smoothed circular features. Therefore, in addition to scale invariance, it provides partial robustness against illumination variations. In the next step, the target image is rotated at angles of 30 and 60 degrees to reveal the capabilities of the selected descriptors against geometric distortions. The dominant orientation is assigned to each extracted feature employing the weighted directional histogram to achieve rotation-invariant matching and increase geometric stability. Also, we resized it to 0.5, 1.5, and 2 times the primary scale.

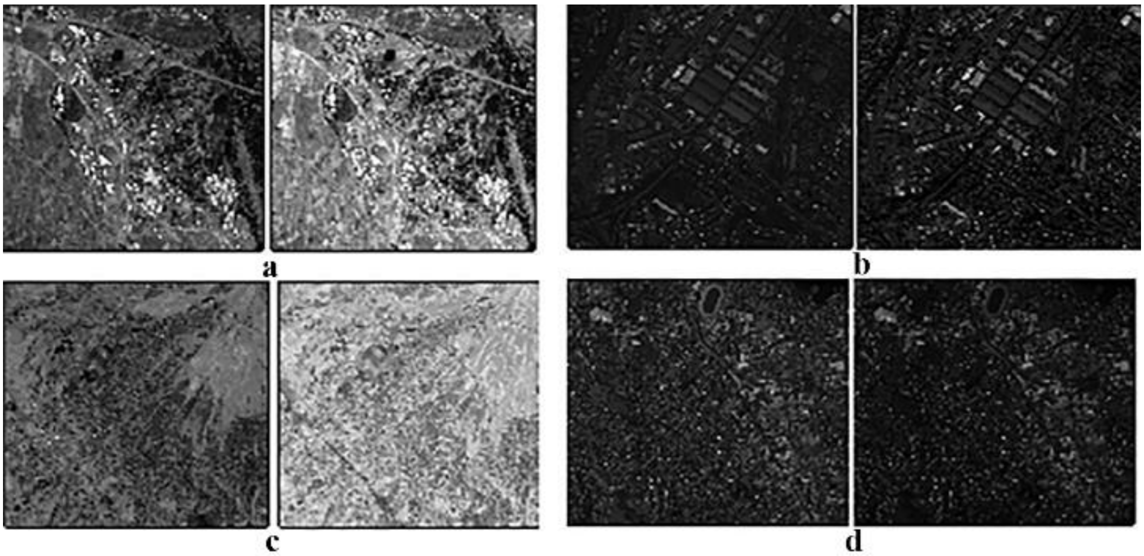

Figure 1. Input image pairs. We utilized different real images to explore the employed methods' capabilities. Case a, b, c, and d are acquired by Sentinel2, Quickbird, ETM$^+$, and IKNOS, respectively.

Figure 3 illustrates the extracted features in different conditions. The feature-based methods are generally fully automated. As a result, the major drawback of the area-based pipeline will be solved. To achieve such a goal, an advanced strategy is needed that employs only specific image structures instead of all positions. Thus, in addition to fully eliminating the human operating effect, the matching efficiency will be increased. By taking advantage of such particular areas (i.e. the local features), only the image structures with high information content are taken into account. On this bases, the feature-based pipeline performs independently of the image intensities and achieves high robustness against complex radiometric differences. Another advantage is the possibility of extraction rotation and scale-invariant features leading to conducting the image matching robustly (Sedaghat et al., 2011; Ye et al., 2018).

In the next step, we computed the feature descriptors. The applied methods are as follows:

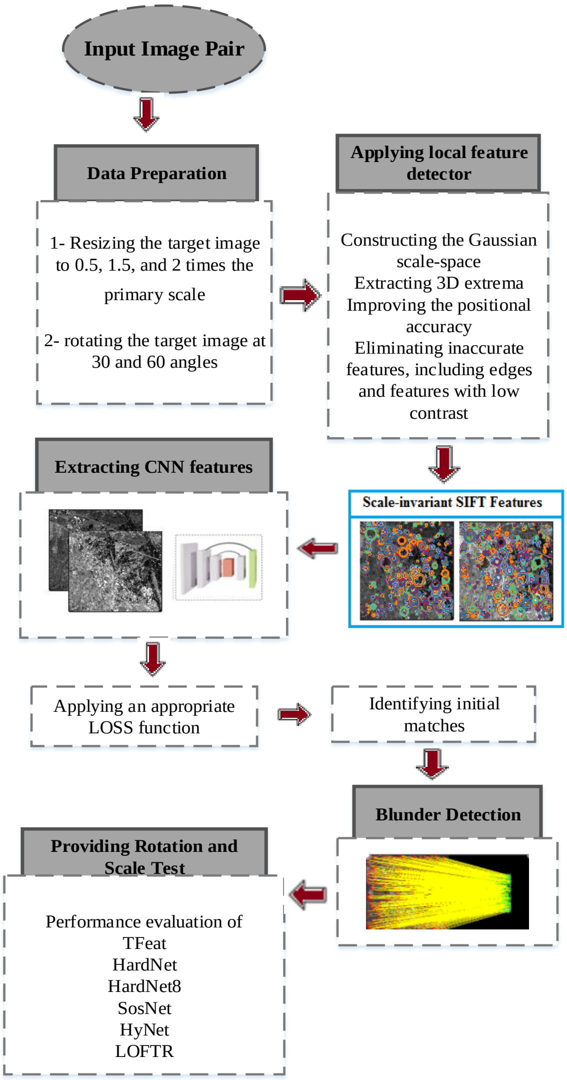
Input Image Pair
Data Preparation
1- Resizing the target image to 0.5, 1.5, and 2 times the primary scale
2- rotating the target image at 30 and 60 angles
Applying local feature detector
Constructing the Gaussian scale-space
Extracting 3D extrema
Improving the positional accuracy
Eliminating inaccurate features, including edges and features with low contrast
Scale-invariant SIFT Features
Extracting CNN features
Applying an appropriate LOSS function
Identifying initial matches
Blunder Detection
Providing Rotation and Scale Test
Performance evaluation of
TFeat
HardNet
HardNet8
SosNet
HyNet
LOFTR

Figure 2. An overview of the designed pipeline.

TFeat employs a triple cost function to identify local correspondences using convolutional neural networks. In this cost function, a strict mismatch patch selection strategy was considered within the cost function and used to train a shallow neural network. The proposed method promised an adequate computational time compared to conventional methods (Balntas et al., 2016).

HardNet introduced a new cost function that outperforms the L2Net model by selecting the most similar mismatch patch within a training batch. The critical point is to reduce the calculation burden required to make the descriptor by 33%, which eliminates the descriptor vector calculation for the mismatch patches (Mishchuk et al., 2017).

SosNet modifies the improved HardNet by adding the Second Order Similarity value Second-Order Similarity value (Tian et al., 2019).

HardNet8 extends L2Net, HardNet, and SosNet designed convolutional networks. The designed network employs eight convolutional layers to extract higher-level features (Pultar, 2020).

LOFTR learns the essential features without employing any detectors. In the first step, both reference and target images are fed to an encoder-decoder network, which extracts (Coarse Level) and small scale (Fine Level) features. Next, a confidence matrix is computed by applying a matching module to the transformed features. The initial correspondences are identified using the L2 norm and a defined threshold. Eventually, final correspondences with sub-pixel accuracy are determined by considering local windows surrounding the initial correspondences within fine-level feature maps (Sun et al., 2021).

HyNet improves the L2Net architecture by adding a hidden layer and converting the common similarity to hybrid similarity (Tian et al., 2020).

3. EVALUATION RESULTS

This study considers precision and inlier numbers as matching criteria. The initial matches are identified by the mutual nearest neighbor method. Obviously, all the initial correspondences

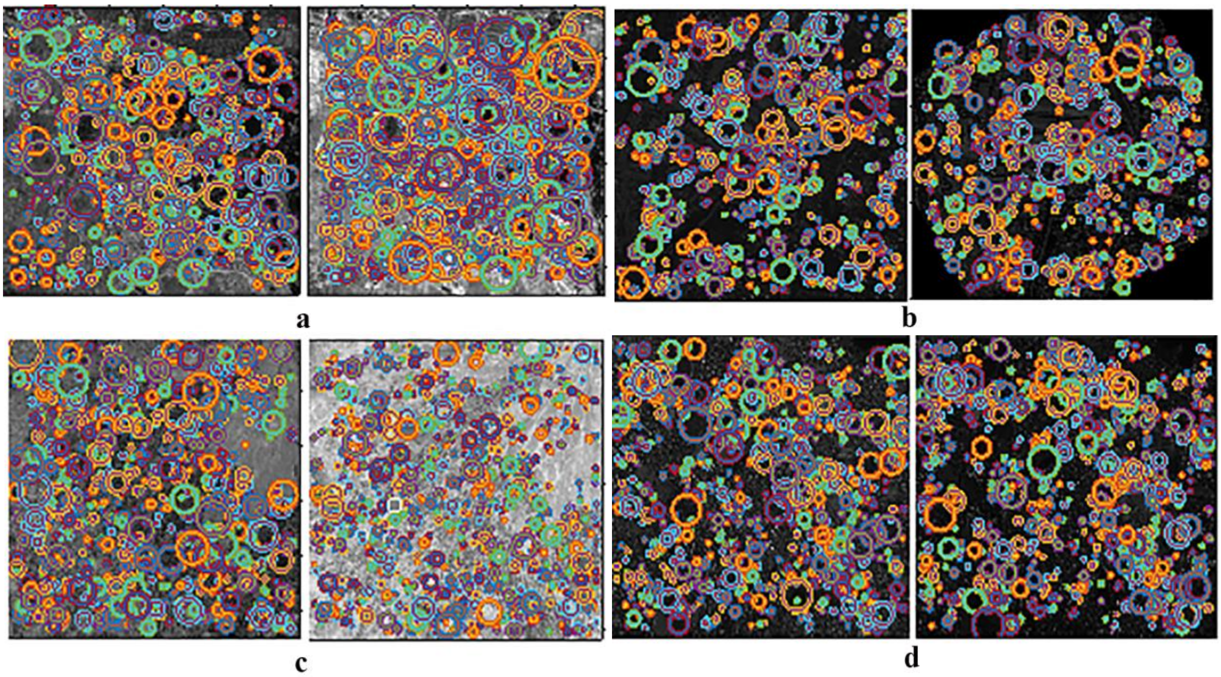

Figure 3. extracted local features in various conditions within pairs a to d

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-4-W1-2022-335-2023 | © Author(s) 2023. CC BY 4.0 License.

could not contribute to the registration procedure. The extracted features must be refined as much as possible. Various methods have been proposed to refine the initial on this basis. The most well-known of which is RANSAC. In this method, the geometric relationship is estimated using random points. Eventually, the model with the most inliers will be considered.

Here, the extracted features are initially refined due to the detector responses. Also, in the current stage, the mismatches are excluded using the fundamental matrix and DEGENSAC algorithm (Jin et al., 2021). The correct matches for LOFTR are identified according to (Sun et al., 2021). Eventually, the mismatches will be excluded using the fundamental matrix and DEGENSAC algorithm (Jin et al., 2021). The Implementation is carried out using the Kornia library. Figure 4 illustrates the correct matches for the rotation test.

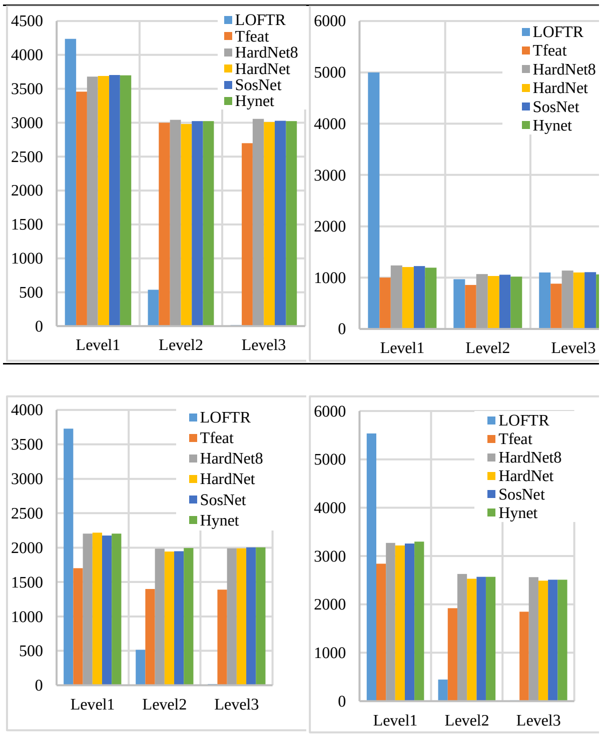

Figure 4. The number of identified inliers for the rotation test. The first row from left to right relate to cases a and b, and the second row from left to right relates to cases c and d, respectively.

As can be seen, LOFTR has achieved better results in the original images, which indicates the highest robustness against radiometric differences. However, despite employing an advanced approach for encoding the images, its performance has rapidly deteriorated in the presence of rotation due to the lack of local feature detectors. HardNet8 surpasses other methods by using eight convolution layers and extracting more complex features among the remaining descriptors. Table.1 compares the architecture of HardNet and HardNet8.

HardNet (1.3M pram)	HardNet8(4.7M pram)
Block1: spatial size 32×32	
Conv 3×3×32 Conv 3×3×32	Conv 3×3×32 Conv 3×3×32
Block2: spatial size 16×16	
Conv 3×3×64/2 Conv 3×3×64/2	Conv 3×3×64/2 Conv 3×3×64/2
Block3: spatial size 8× 8	
Conv 3×3×128 Conv 3×3×128	Conv 3×3×128/2 Conv 3×3×128 Conv 3×3×256
Global pooling block	
Dropout(0.3)	
Conv 8×8×128	Conv 8×8×256
Flatten, L2 Norm	

However, a high number of identified matches is not still satisfying. Inadequate stability is another factor that deteriorates subsequent processing. One may work with UAV images by software such as Pix4D or PhotoScan and tend to extract an extensive number of tie points. Although this may align images with insufficient overlapping, the accuracy of further processing will not be improved. A fundamental reason is the lack of robust identified features. To this point, we used the precision criterion, which is the number of correct correspondents to the total number of correspondences found by the matching algorithm. The higher precision, the higher the amount of stability. Figure 5 demonstrates the highest precision of HardNet8, implying the highest achieved stability and distinctiveness.

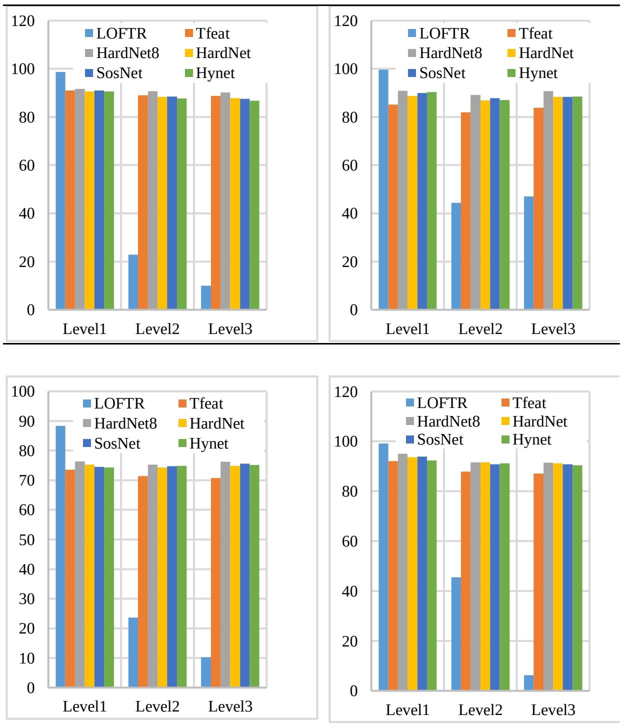

Figure 5. The precision values for the rotation test. the first row from left to right relates to cases a and b, and the second row from left to right relate to cases c and d, respectively

As expected, assignment and applying the SIFT feature detector reveal rotation robustness effectively and prevent the matching performance from being diminished. Also, we provided a scale robustness test. Figure 6 indicates the inlier numbers. As can be seen, LOFTR is more resistant to scale differences compared to the rotation test. In many image pairs, LOFTR has extracted more inliers. Especially within high-resolution images of the second pair. The coarse-to-fine approach of extracting features at 1/8 and 1/2 of the initial scale and feeding them to the transformer module provided extensive correct match numbers. However, according to Figure 7, other descriptors still achieved acceptable accuracy in many cases despite identifying lower correct correspondences. This issue implies the importance of feature

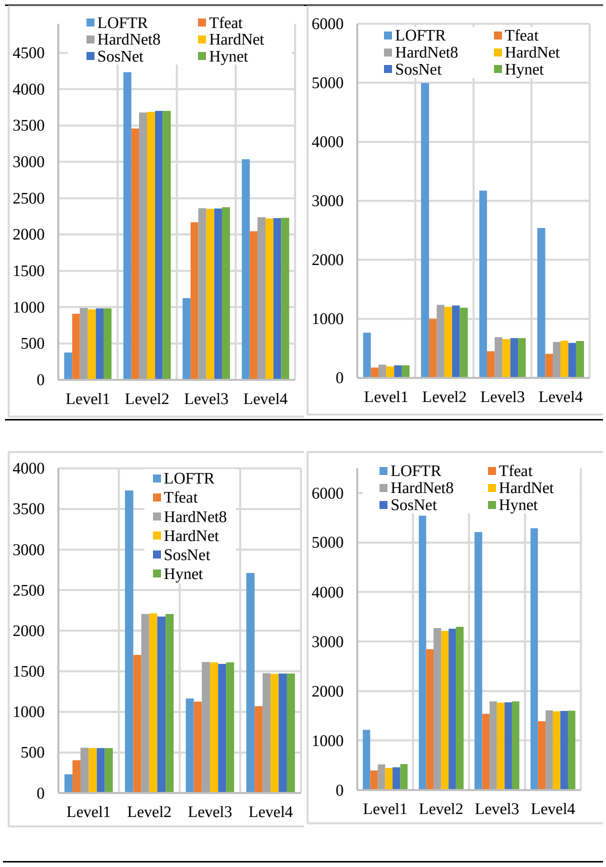

Figure 6: the number of identified inliers for the scale test. The first row from left to right relate to cases a and b, and the second row from left to right relates to cases c and d, respectively.

detectors. As expected, employing distinct structural regions instead of all positions improved the matching performance. In pixel-based methods, although a large number of correspondences are usually identified, many of these correspondences are unstable features due to geometrical

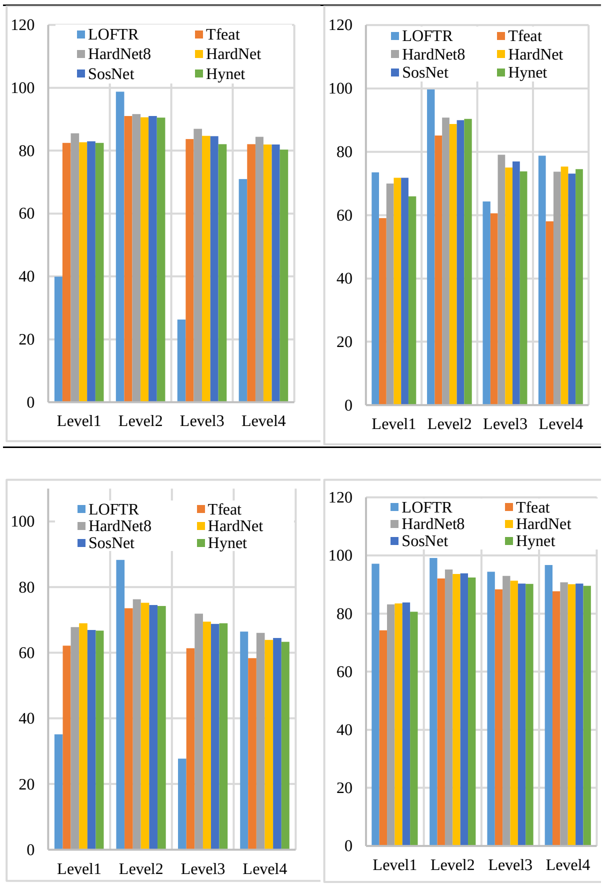

Figure 7. The precision values for the scale test, the first row from left to right relates to cases a and b, and the second row from left to right relate to cases c and d, respectively.

differences. Applying the SIFT scale-invariant features preserves the matching distinctiveness and stability. Also, advanced image

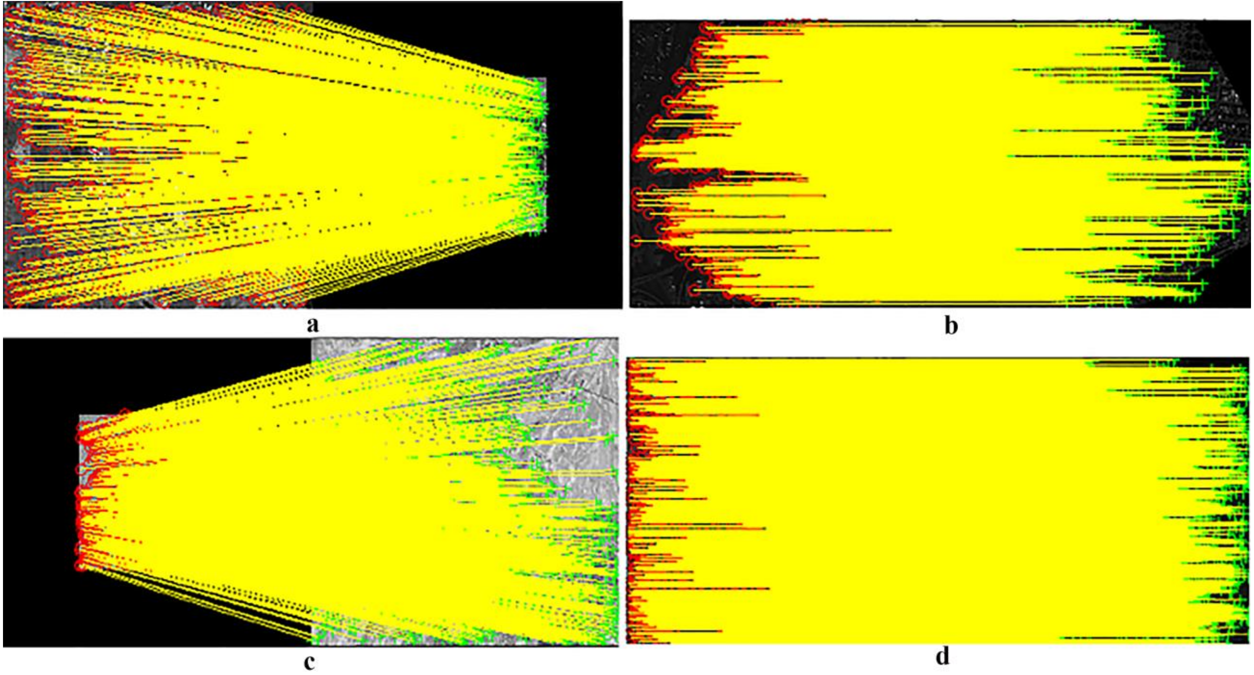

Figure 8. visualization of correct matches identified by HardNet8. The first row from left to right relate to pairs a and b, and the second row from left to right relates to pairs c and d, respectively.

simulation employing eight convolution layers at various scales caused HardNet8 to outperform the others.

As previously stated, regardless of geometric differences, another challenge of multispectral images is the variation in the spectral signature of terrestrial features. Therefore, radiometric differences could also deteriorate the descriptor. It can be seen that all descriptors achieved appropriate radiometric robustness, which implies the superiority of feature engineering elimination and encoding the images to various and robust features. Apart from the LOFTR descriptor, which has detected many matches due to its pixel-based nature, among the other detector-based methods, the Hardnet8 descriptor is the most insensitive to radiometric differences. However, the strategies used in other methods, such as the second-order similarity in SosNet and hybrid similarity in HyNet, have also dealt well with the complex distortions. Generally, in methods based on deep learning, the researcher is more able to extract high-level features, effectively improving the matching performance of multimodal images.

Figure 8 inspects the HardNet8 identified correct matches in multiple cases. As can be seen, sufficient inlier numbers are identified despite all various challenging conditions.

4. CONCLUSIONS

This paper comprehensively evaluated multiple well-known deep feature descriptors against geometric distortions for multispectral image matching. Results demonstrate the overall superiority of Hardnet8 due to extracting deep high-level structures. Moreover, employing SIFT scale-invariant detector and assigning dominant orientation to each feature maintained geometric robustness significantly. Also, in the case of lacking rotation and scale differences distortions, the LOFTR detector free descriptor obtains a great number of correspondences. However, the descriptor performance diminishes rapidly in the presence of geometric distortions.

REFERENCES

Alcantarilla, P. F., Bartoli, A., & Davison, A. J. (2012). *KAZE features.* Paper presented at the European conference on computer vision.

Balntas, V., Riba, E., Ponsa, D., & Mikolajczyk, K. (2016). *Learning local feature descriptors with triplets and shallow convolutional neural networks.* Paper presented at the Bmvc.

Hosseiny, B., Mahdianpari, M., Brisco, B., Mohammadimanesh, F., & Salehi, B. (2021). WetNet: A Spatial-Temporal Ensemble Deep Learning Model for Wetland Classification Using Sentinel-1 and Sentinel-2. *IEEE Transactions on Geoscience and Remote Sensing*.

Jiang, X., Ma, J., Xiao, G., Shao, Z., & Guo, X. (2021). A review of multimodal image matching: Methods and applications. *Information Fusion, 73*, 22-71.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K. M., & Trulls, E. (2021). Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision, 129*(2), 517-547.

Li, J., Hu, Q., & Ai, M. (2019). RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing, 29*, 3296-3310.

Liu, X., Ai, Y., Tian, B., & Cao, D. (2018). Robust and fast registration of infrared and visible images for electro-optical pod. *IEEE Transactions on Industrial Electronics, 66*(2), 1335-1344.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision, 60*(2), 91-110.

Ma, J., Yu, W., Liang, P., Li, C., & Jiang, J. (2019). FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information fusion, 48*, 11-26.

Mishchuk, A., Mishkin, D., Radenovic, F., & Matas, J. (2017). Working hard to know your neighbor's margins: Local descriptor learning loss. *Advances in neural information processing systems, 30*.

Ojala, T., Pietikainen, M., & Harwood, D. (1994). *Performance evaluation of texture measures with classification based on Kullback discrimination of distributions.* Paper presented at the Proceedings of 12th international conference on pattern recognition.

Pultar, M. (2020). Improving the HardNet Descriptor. *arXiv preprint arXiv:2007.09699*.

Quan, D., Wang, S., Liang, X., Wang, R., Fang, S., Hou, B., & Jiao, L. (2018). *Deep generative matching network for optical and SAR image registration.* Paper presented at the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium.

Sedaghat, A., & Mohammadi, N. (2018). Uniform competency-based local feature extraction for remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing, 135*, 142-157.

Sedaghat, A., & Mohammadi, N. (2019). Illumination-Robust remote sensing image matching based on oriented self-similarity. *ISPRS Journal of Photogrammetry and Remote Sensing, 153*, 21-35. doi:10.1016/j.isprsjprs.2019.04.018

Sedaghat, A., Mokhtarzade, M., & Ebadi, H. (2011). Uniform robust scale-invariant feature matching for optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing, 49*(11), 4516-4527.

Sun, J., Shen, Z., Wang, Y., Bao, H., & Zhou, X. (2021). *LoFTR: Detector-free local feature matching with transformers.* Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Tian, Y., Barroso Laguna, A., Ng, T., Balntas, V., & Mikolajczyk, K. (2020). Hynet: Learning local descriptor with hybrid similarity measure and triplet loss. *Advances in neural information processing systems, 33*, 7401-7412.

Tian, Y., Yu, X., Fan, B., Wu, F., Heijnen, H., & Balntas, V. (2019). *Sosnet: Second order similarity regularization for local descriptor learning.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Wang, S., Quan, D., Liang, X., Ning, M., Guo, Y., & Jiao, L. (2018). A deep learning framework for remote sensing image registration. *ISPRS Journal of Photogrammetry and Remote Sensing, 145*, 148-164.

Yan, H., Yang, S., Xue, Q., & Zhang, N. (2022). HR optical and SAR image registration using uniform optimized feature and extend phase congruency. *International Journal of Remote Sensing, 43*(1), 52-74.

Yang, Z., Dan, T., & Yang, Y. (2018). Multi-temporal remote sensing image registration using deep convolutional features. *Ieee Access, 6*, 38544-38555.

Ye, Y., Bruzzone, L., Shan, J., Bovolo, F., & Zhu, Q. (2019). Fast and robust matching for multimodal remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing, 57*(11), 9059-9070.

Ye, Y., Shan, J., Hao, S., Bruzzone, L., & Qin, Y. (2018). A local phase based invariant feature for remote sensing image matching. *ISPRS Journal of Photogrammetry and Remote Sensing, 142*, 205-221.

Zhu, R., Yu, D., Ji, S., & Lu, M. (2019). Matching RGB and infrared remote sensing images with densely-connected convolutional neural networks. *Remote Sensing, 11*(23), 2836.