Heterogeneous remote-sensing image matching method based on deep learning

Huitai Hou*, Qing Xu, Chaozhen Lan, Zhixiang Cui, Jianqi Qin, and Lianxia Wang

Information Engineering University Zhengzhou 450052, China

* Corresponding author: houhuitai@163.com

Abstract—In this paper, a deep learning matching method is proposed to address the difficulty in matching heterogeneous remote sensing images, which is caused by their differences in imaging modes, time phases, and resolutions. A heterogeneous image is inputted into the convolutional neural network (CNN) to extract deep features and, then, a graph neural network (GNN) is used for matching. Finally, correct matching points are retained while ensuring the effective elimination of mismatches. The algorithm adopted in this study was tested using multiple sets of heterogeneous remote-sensing images and compared with D2-Net+NN+RANSAC and Superpoint+SuperGlue algorithms. The results show that the algorithm used in this study possesses strong adaptability and robustness and is an optimal algorithm for the robust matching of remote-sensing images with different sources and large differences.

1. INTRODUCTION

Owing to the rapid development of remote-sensing technology, ground observation images of visible light, infrared, synthetic aperture radar (SAR), and other sensors are becoming more abundant. Heterogeneous images acquired by different platforms and sensors have a certain complementarity, thereby providing a massive data source for the in-depth mining of remote sensing information and big data analysis. The matching between images is a core issue for further processing and analysis of heterogeneous images. Their differences in imaging mechanism, waveband, and time suggest that heterogeneous images have huge differences in radiation and geometric characteristics. Therefore, matching them has always been a difficult task in image matching research.

Image feature matching usually extracts local feature information descriptors from a certain neighborhood of key points and compares the descriptors to determine the matching points. The most famous descriptor is the scale invariant feature transformation (SIFT) descriptor [1]. SIFT descriptors can efficiently resist the rotation and scale differences between images; however, owing to the gradient distribution based on the local neighborhood of the image, the matching effect on heterogeneous images is poor. In recent years, deep learning methods, especially convolutional neural network (CNN), have made tremendous progress and performance improvements in computer vision tasks such as image classification, target detection, and segmentation. With the continuous layers of CNN, complex image features have increasingly become more accessible and specific high-level features can be learned. Since its first introduction in [2] in 2014, scholars have been applying CNN to the image feature extraction process, gradually shifting from SIFT features to CNN features [2-6]. In 2019, Dusmanu et al. [5] proposed a method (D2-Net) that involves detecting feature points and extracting feature descriptors simultaneously, which have made important progress in solving the problem of road sign recognition in varying scenes and have shown great potential.
However, after image feature extraction is completed, feature-based matching is required. The matching task aims to establish the correct image pixel or point correspondence between two images with or without feature detection and/or description. This task plays a significant role in the entire image matching pipeline. Feature-based matching is generally performed by matching descriptors with the nearest neighbor (NN) search and estimating a geometric transformation to filter incorrect matches. However, heterogeneous images are not similar in terms of imaging mechanism, waveband, and time. It is difficult to overcome these differences using traditional matching methods, thereby leading to mismatching. In recent years, GNN has become increasingly popular in various fields. It has demonstrated obvious advantages in the processing of non-European spatial data. The graph neural network can be used to estimate the geometric transformation between heterogeneous images. In 2020, for matching among large viewpoint and lighting changes, occlusion, blur, and lack of texture images, Sarlin et al. proposed a feature matching method based on a graph neural network, SuperGlue [7]. The combination of SuperGlue and Superpoint [6] feature extraction methods can handle ground close-up visible optical images with large viewing angle changes. However, for heterogeneous remote sensing images with huge differences in various aspects, the feature extraction network used by Superpoint cannot extract stable deep learning features and performs poorly in processing the images.

Therefore, based on D2-Net deep learning features, this paper attempts to introduce the basic ideas of the D2-Net feature extraction and SuperGlue feature matching algorithms, and proposes a new heterogeneous remote sensing image matching algorithm based on deep learning features, as shown in Fig. 1. It is expected that the high-level features extracted by CNN can be matched with features based on the graph neural network to form a robust and new heterogeneous remote-sensing image matching method.

2. METHOD AND PRINCIPLE

2.1. Feature Extraction and Detection

The CNN trained with a large number of specific sample data has excellent feature expression capabilities. To make the network suitable for feature extraction, this study selects and adapts the classic VGG16 network model. The VGG16 model has 5 convolutional networks. Generally, the receptive fields of the first few layers of the network are very small, and the features obtained are local features relative to the bottom layer. The extracted features are mostly edges, corners, and other features, with higher positioning accuracy; the higher the number of network layers, the more abstract the extracted features. The more global the information is, the more resistant the information becomes to interference caused by heterogeneous images; however, the positioning accuracy becomes worse.

To make the feature points not only abstract enough but also to obtain higher positioning accuracy, the last (third) convolutional layer, Conv4_3, in the middle fourth layer is selected as the feature map. Using this network structure, the original input image, its size, and the number of channels are \( I, w \times h \), and \( n = 512 \), respectively. Additionally, the network output feature map is a 3D tensor, \( F = F(I), F \in \mathbb{R}^{w \times h \times n} \).
The feature key point location and descriptor extraction are denoted by $F$ at the same time.

2.2. Dimensionality Reduction and Keypoint Encoder

The feature descriptor extracted by the feature extraction algorithm has a dimension of 512, and its feature dimension is very high, and the calculation is very large, which is not suitable for SuperGlue. In contrast, to make the full use of the context information of feature points, it is necessary to process feature descriptors and location information of feature points simultaneously. The next step is to reduce the dimensions and encode the features.

This study uses the principal component analysis (PCA) method to reduce the feature descriptor to 256 dimensions. We embed the key point position into a vector with a multilayer perceptron (MLP) [8] as follows:

$$^0x_i = d_i + \text{MLP}_{\text{enc}}(p_i)$$

where $d_i$ is the feature descriptor after dimensionality reduction and $p_i$ is the coordinate of the feature point.

2.3. Matching

Considering matching as an optimization problem, the cost is predicted by a graph neural network, and relevant experiences are directly learned from the data. Feature matching is based on two constraint principles: 1) a key point can have a maximum of a single correspondence with the other image and 2) some key points will be unmatched due to the occlusion and failure of the detector. Based on these two principles, a partial soft assignment matrix is designed $P \in [0,1]^{M \times N}$ as follows:

$$P_{1N} \leq 1_M \text{ and } P_{1M} \leq 1_N.$$  

We consider a single complete graph whose nodes are the key points of both the images. This graph is a multiplex graph with two types of undirected edges. This study uses a message passing formulation to transmit information along these two sides. The resulting multiplex graph neural network starts with a high-dimensional state for each node and computes an updated representation by aggregating messages across all given edges for all nodes simultaneously [9].

After the node information is aggregated, the next step is to perform matching optimization to construct a partial assignment matrix $P$. As in the standard graph matching formulation, we first calculated the score matrix $S \in \mathbb{R}^{M \times N}$ of all possible matches, under the constraint of equation (2), by maximizing the overall score (equation (3)) to obtain the best $P$. The process is solved using the Sinkhorn algorithm [10]:

$$\sum_{i,j} S_{ij} P_{ij}.$$  

3. Results and Discussion

To comprehensively evaluate the performance and adaptability of the heterogeneous remote sensing image matching algorithm proposed in this paper, we established a test data set containing multi-source remote sensing images. The deep learning model was constructed under the PyTorch 1.3.0 framework. The computer used for the test was an MSI P65 notebook (i9-9880H CPU, GeForce RTX graphics card, 2070Max-Q 8 GB video memory, and 32 GB of memory). The implementation language used was Python and the operating system was Ubuntu 16.04.

3.1. Test Data

Table 1 shows the test data. For ease of description, the data are named DS1-DS8; the image data source, band type, resolution, cropped pixel size, and imaging time are listed in detail in the table.

| Group | Image | Data source | Image type | Resolution (m) | Pixel   | Imaging time |
|-------|-------|-------------|------------|---------------|---------|--------------|
| 1     | DS1   | Google Earth | visible    | 0.5           | 1000x1000 | 2009.12.27   |
|       | DS2   | UAV         | visible    | 0.5           | 1000x1000 | 2019.06.07   |
The test data sources include space-borne sensors, drone sensors, and Google Earth images. The bands include visible light, SAR, and thermal infrared bands, with different resolutions, time periods, and seasonal spans. The performance of the algorithm can be completely tested. The data are shown in Fig. 2–5.

|   | Data Source | Sensor | Band Type     | Resolution | Date       |
|---|-------------|--------|---------------|------------|------------|
| 2 | DS3         | UAV    | visible       | 0.1        | 640x512    | 2017.09.18 |
|   | DS4         | UAV    | thermal infrared | 0.1        | 640x512    | 2019.05.16 |
| 3 | DS5         | Google Earth | visible | 0.5        | 1500x1500  | 2018.10.29 |
|   | DS6         | UAV    | SAR           | 0.5        | 1500x1500  | 2019.10.26 |
| 4 | DS7         | ZY-3 PAN | visible   | 2.0        | 1000x1000  | 2018.06.07 |
|   | DS8         | GF-3 SL | SAR           | 2.0        | 1000x1000  | 2016.10.14 |

(a) DS1 Google Earth image (2009)  
(b) DS1 UAV visible image (2019)

Figure 2. Group 1 dataset.

(a) DS3 UAV visible image (2017)  
(b) DS4 UAV thermal infrared

Figure 3. Group 2 dataset.

(a) DS5 Google Earth image (2018)  
(b) DS6 UAV SAR image

Figure 4. Group 3 dataset.
3.2. Experimental Results and Analysis

Based on the test data set of multi-source remote sensing images, this study conducted a comparative evaluation of three matching methods: D2-Net+NN+RANSAC, Superpoint+SuperGlue, and D2-Net+SuperGlue. Among them, NN+RANSAC is the traditional matching method in close proximity to the next search and RANSAC mismatch elimination. The results are shown in Table 2 and Fig. 6.

It can be observed from the results that, compared with the two methods, D2-Net+NN+RANSAC and Superpoint+SuperGlue, the D2-Net+SuperGlue method greatly improved the matching effect and its number of correct matches is the highest in the set. Particularly for the third group of images with the largest difference, D2-Net+NN+RANSAC can extract seven sets of correct matching point pairs, whereas Superpoint+SuperGlue returns as completely invalid. D2-Net+SuperGlue can extract 182 correct matching point pairs. For heterogeneous images with large differences, Superpoint cannot extract deep learning features effectively. Although D2-Net can extract effective features, after using traditional methods for matching, a large number of effective features are eliminated, and the matching effect is not ideal. The D2-Net+SuperGlue method proposed in this paper can extract valid features and match them correctly with the best effect.

| Group | Method                             | Correct number of matches | Elapsed time |
|-------|------------------------------------|---------------------------|--------------|
| 1     | D2-Net+NN+RANSAC                   | 455                       | 5.25         |
|       | Superpoint+SuperGlue               | 587                       | 1.89         |
|       | D2-Net+SuperGlue                   | 1045                      | 6.13         |
| 2     | D2-Net+NN+RANSAC                   | 183                       | 0.7          |
|       | Superpoint+SuperGlue               | 368                       | 1.55         |
|       | D2-Net+SuperGlue                   | 422                       | 1.82         |
| 3     | D2-Net+NN+RANSAC                   | 7                         | 4.69         |
|       | Superpoint+SuperGlue               | 0                         | 1.55         |
|       | D2-Net+SuperGlue                   | 182                       | 5.38         |
| 4     | D2-Net+NN+RANSAC                   | 291                       | 4.22         |
|       | Superpoint+SuperGlue               | 477                       | 1.14         |
|       | D2-Net+SuperGlue                   | 928                       | 5.07         |

Figure 5. Group 4 dataset.
4. CONCLUSIONS
Image matching is a huge challenge owing to the huge differences in imaging mode, time phase, resolution, etc., among different heterogeneous remote sensing images. This study proposes a matching method based on deep learning, inputs heterogeneous images into a CNN network to extract deep features, and then uses a graph neural network for matching. By ensuring that mismatches are effectively eliminated, correct matching points can be retained. Experimental results show that the heterogeneous image matching algorithm proposed in this paper has strong adaptability and robustness and is an optimal method for the robust matching of heterogeneous remote sensing images. There are a few limitations to the proposed method. First, this method requires a long time for matching. Second, owing to the characteristics of CNN, the accuracy of feature positioning is not high enough. Improving the efficiency and positioning accuracy of the algorithm is the next major challenge that requires a solution.

REFERENCES
[1] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.
[2] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, “Discriminative unsupervised feature learning with exemplar convolutional neural networks,” IEEE transactions on pattern analysis machine intelligence, vol. 38, no. 9, pp. 1734–1747, 2015.
[3] S. Zagoruyko and N. Komodakis. “Learning to compare image patches via convolutional neural networks,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4353–4361, 2015.
[4] F. Ye, Y. Su, H. Xiao, X. Zhao, and W. Min, “Remote sensing image registration using convolutional neural network features,” IEEE Geoscience Remote Sensing Letters, vol. 15, no.
2, pp. 232–236, 2018.

[5] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, et al., “D2-Net: A Trainable CNN for joint description and detection of local features,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8092–8101, 2019.

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

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

[8] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, “Convolutional sequence to sequence learning,” ICML, p. 3, 2017.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, et al., “Attention is all you need,” Advances in neural information processing systems, pp. 5998–6008, 2017.

[10] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” Advances in neural information processing systems, pp. 2292–2300, 2013.