Multi-Modal Remote Sensing Image Matching Method Based on Deep Learning Technology

Hao Han¹,², *, Canhai Li², Xiaofeng Qiu¹,²

¹Liaoning Technical University, Fuxin 123000, China;
²Land Satellite Remote Sensing Application Center, Ministry of Natural Resources of the People's Republic of China, Beijing 100048, China

*Corresponding author: hanhao@lasac.cn

Abstract. Remote sensing is a scientific technology that uses sensors to detect the reflection, radiation or scattering of electromagnetic wave signals from ground objects in a non-contact and long-distance manner. The images are classified by the extracted image feature information. Recognition is a further study of obtaining target feature information, which is of great significance to urban planning, disaster monitoring, and ecological environment evaluation. The image matching framework proposed in this paper matches the depth feature maps, and reversely pushes the geometric deformation between the depth feature maps to between the original reference image and the target image, and eliminates the geometric deformation between the original images. Finally, through feature extraction of the corrected image, the extracted local feature image blocks are input into the trained multi-modal feature matching network to complete the entire matching process. Experiments show that the negative sample set construction strategy that takes into account the sample distance proposed in this experiment can effectively deal with the problem of neighboring point interference in RSI matching, and improve the matching performance of the network model.

Keywords: Deep Learning, Multimodal Remote Sensing, Convolutional Neural Network, Image Matching.

1. Introduction

With the rapid progress of computer technology, the use of computer-assisted classification of RSIs is currently a widely used classification method [1-2]. The application of fusion analysis of multi-modal RSIs has become a hot topic that has great prospects but needs to be solved urgently. The registration of multi-modal RSIs, as the first step of multi-modal image fusion analysis, lays the foundation for the success or failure of subsequent target detection, target recognition, change detection and other applications [3-4]. Different imaging modalities have different advantages, and the combined application of RSIs of different modalities is extremely attractive [5-6].

Due to the broad application prospects of multi-modal RSIs, many institutions at home and abroad have carried out research on multi-modal RSI registration methods, and many institutions have achieved important research results [7-8]. The pixel-based method uses a single pixel as the basic processing unit and uses the spectral characteristics of the pixel as the main basis to classify images.
However, the spectral information of high-resolution RSIs is relatively scarce, and spatial information such as texture and geometry are more abundant; at the same time, the pixel-based method has a slower interpretation speed and is prone to salt and pepper effect, so this method is not suitable for high-resolution RSI classification [9]. The key to the image segmentation stage is the selection of the segmentation method and segmentation parameters. The more widely used segmentation method is the regional segmentation method, represented by the multi-scale segmentation method based on the evolution of fractal networks. This method is also the segmentation method used by the Congnion software, which involves go to the problem of segmentation parameter selection [10].

This paper focuses on the feature matching of multi-modal RSIs based on deep learning, and on the basis of previous studies, conducts in-depth research on the difficulty of matching caused by nonlinear gray-scale changes and geometric deformations in feature matching of multi-modal RSIs.

2. Multi-Modal RSI Matching Method Based on Deep Learning Technology

2.1. CNN Technology and Image Matching Strategy

(1) CNN technology

Among many neural network models, CNN (Convolutional Neural Network) is the most widely used one. Since Hinton and his student Alex proposed Alex Net in 2012. At present, CNNs have achieved excellent results in many tasks in image fields. Use \( w_b \) to denote the bias term of the convolution kernel; use \( f \) to denote the activation function, and the convolution calculation formula is as follows:

\[
\alpha_{i,j} = f \left( \sum_{m=0}^{2} \sum_{n=0}^{3} w_{m,n} x_{i+m,j+n} + w_n \right)
\]

(1)

The formula for calculating the size of the feature map is as follows:

\[
H_{out} = \left\lfloor \frac{H_{in} + 2 \times \text{padding} - \text{dilation} \times (\text{kernel}_\text{size} - 1) - 1}{\text{stride}} + 1 \right\rfloor
\]

(2)

\[
W_{out} = \left\lfloor \frac{W_{in} + 2 \times \text{padding} - \text{dilation} \times (\text{kernel}_\text{size} - 1) - 1}{\text{stride}} + 1 \right\rfloor
\]

(3)

In the formula, \( H_{in} \) represents the height of the input image, \( W_{in} \) represents the width of the input image, \( H_{out} \) represents the height of the output feature map, \( W_{out} \) represents the width of the output feature map, padding represents the boundary filling range, dilation represents the expansion coefficient of the convolution kernel, kernel_size represents the size of the convolution kernel, stride represents the convolution step size [11-12].

(2) RSI matching strategy based on deep features of CNN

The RSI matching strategy based on depth features includes four steps:

1) Model training: Use training samples to train the CNN to obtain the optimal CNN model;
2) Feature point extraction Use a certain feature point extraction method to extract feature points on the reference image and the image to be matched, and use the neighborhood information of feature points to form image blocks;
3) Depth feature description Use the trained optimal CNN model to extract the depth feature;
4) Feature matching judges the similarity measurement based on the depth feature description of each image block, so as to determine the matching image block on the reference image and the image.
to be matched, that is, realize the one-to-one correspondence between the points of the same name and obtain the matching result.

2.2. Multi-Modal RSI Matching Method Based on Deep Learning Technology

CNNs have achieved excellent results in the field of computer vision, but if they are directly transplanted to the RSI matching problem, the ideal results are often not obtained. CNNs can mine the deep semantic information of images well and overcome the non-linear gray-scale difference between multi-modal images. However, for RSIs with widespread geometric deformation, the existing methods are difficult to take into account the effects of nonlinear grayscale differences and obvious geometric deformation. In addition, the neighboring points around some matching points can cause mismatching. In response to the above problems, this paper proposes a sample set construction strategy that takes into account the sample distance, extracts the depth features of multimodal images through CNNs, and then proposes an image matching based on deep learning technology for the problem of geometric deformation between images method.

The multi-modal image matching method proposed in this paper deals with the non-linear gray-scale difference and geometric deformation respectively in turn. The multi-modal image feature extraction network is used to alleviate the non-linear gray-scale difference between the images, and the BoF algorithm-based image retrieval algorithm is used to accelerate the matching of deep feature maps, and then the SFT algorithm is used to match the feature maps and transfer them back to the original image. Partial image blocks of feature points cropped based on the target image after rough correction are sent to the similarity measurement network for similarity calculation to complete the entire matching process.

(1) Multimodal image depth feature extraction and matching network construction

The image feature matching network proposed in this paper is composed of two sub-networks, a multi-modal feature extraction network (CS Net) and a similarity measurement network (CS Net+FC). Each branch of the multi-modal feature extraction network is composed of several convolution modules, and the network parameters are reduced by sharing weights between the two branches. Each convolution module adopts the "Conv + BN + ReLU" construction strategy, that is, after each layer of convolution, the batch normalization layer (BN layer) is used to accelerate the convergence during network training, and then the ReLU function is added to the neuron activate. Compared with the traditional Siamese network, the matching network proposed in this paper eliminates the pooling layer structure. Because in the RSI matching task, it is necessary to obtain the coordinates of the points with the same name as accurately as possible, and while the pooling layer reduces the amount of network parameters by means of dimensionality reduction, it is easy to cause incorrect matching of the adjacent feature points of the points with the same name.

(2) Robust multi-modal image depth feature extraction with nonlinear grayscale changes

In the network training process of this article, CS Net receives positive and negative sample pairs and corresponding labels, and extracts image depth features layer by layer; the similarity measurement network measures the similarity of image pairs and trains the model, continuously through gradient descent and back propagation update the parameters such as the convolution kernel and the fully connected layer in the network, so that the network model can correctly identify whether the input image pair matches or not.

3. RSI Matching Experiment Based on Deep Features of CNN

3.1. Experimental Data

The multi-modal image sample training set constructed in this paper contains three data sets of visible light-near infrared, optical-SAR, and optical-LiDAR. The image sources include Landsat8 satellite image, Gaofen-3 GF3 satellite image, Ziyuan-3 ZY3 image, Tera SAR-X image and LIDAR point cloud elevation rendering image. Due to the deviation of the sample ratio and the weighting of the samples, the network parameters will shift to the negative sample, which is not conducive to correctly
identifying the similarity of the sample pair. Therefore, this paper randomly selects N/2 pairs from all N pairs of normal negative samples and N pairs of negative samples taking into account the sample distance, and matches with N pairs of positive samples. This article N takes 150,000.

3.2. Description of Network Training and Training Parameter Settings

The training environment configuration is shown in Table 1:

| Software and hardware | Configuration       |
|----------------------|---------------------|
| System               | Ubuntu 16.04        |
| RAM                  | 32GB                |
| CPU                  | Intel(R) Core i7-6850K @ 3.60GHz 3.60GHz |
| GPU                  | NVIDIA GTX 1080Ti * 2 |
| Deep learning framework | TensorFlow     |

In the training process, the sample batch size of each iteration is 32 pairs of samples. The initial learning rate of the network is 0.001, and the momentum is 0.9. When the average training loss value is lower than 0.001, the network terminates training.

4. Experimental Analysis of RSI Matching Based on Deep Features of CNN

4.1. Experimental Results and Comparative Analysis of Sample Set Construction Taking into Account Sample Distance

The test data of this comparative experiment comes from 5 Landsat8 images, which are cropped into image blocks with a size of 97×97 pixels, with a total of 4797 pairs of test samples. In the experiment, the number of correct recognitions of visible light-near-infrared neighboring negative samples in the above five test sets of the network model 1 and network model 2 obtained from the training of the two training sets is counted, and the recognition accuracy of each test set is calculated separately. The results of this comparative experiment are shown in Table 2:

| Serial number | Total number of samples | Model 1 | Model 2 | Model 1 | Model 2 |
|---------------|------------------------|---------|---------|---------|---------|
|               |                        | Number of recognitions | Number of recognitions | Correct rate | Correct rate |
| 1             | 1021                   | 875     | 984     | 85.70%  | 93.44%  |
| 2             | 1158                   | 724     | 947     | 62.52%  | 81.78%  |
| 3             | 1124                   | 647     | 912     | 57.56%  | 79.36%  |
| 4             | 1354                   | 941     | 1012    | 69.50%  | 84.34%  |
| 5             | 543                    | 354     | 402     | 65.19%  | 85.08%  |
5

Figure 1. Comparative experiment results of sample set construction considering sample distance

Visualize the model recognition accuracy rate in Table 2 to get Figure 1. Combining the statistical results of the two to summarize, we can get the following conclusion: Compared with the network model 1 obtained by training with the ordinary training set, the training set with the training set taking into account the sample distance is obtained. In the network model 2, the number of identifications and the accuracy of identification of the negative samples in the visible-near-infrared neighborhood are greatly improved on the five test sets.

Combined with the comparative experimental plan, the above experimental results are comprehensively analyzed. In this comparative experiment, aiming at the interference of neighboring points around the point with the same name in RSI matching, a negative sample set construction strategy that takes into account the sample distance is proposed. The search image corresponding to each local image block on the reference image has the same name. Point position, by randomly shifting 5 pixels in 8-direction to simulate neighboring points. Considering that the ideal network model needs to deal with two non-matching situations of non-matching points and neighboring points, half of the normal negative samples are randomly replaced with neighboring negative samples in the training set, so that the trained network model 2 has the characteristics of neighboring non-matching points. However, the network model 1, which uses ordinary negative samples for training, lacks consideration of neighboring point interference, and performs poorly in each test set.

Based on the above experimental analysis, it can be seen that the negative sample set construction strategy that takes into account the sample distance proposed in this experiment can effectively deal with the problem of neighboring point interference in RSI matching, and improve the matching performance of the network model.

4.2. Experimental Results and Comparative Analysis of Feature Extraction Network Structure

The multi-modal image feature matching network proposed in this paper is composed of two sub-networks, a feature extraction network and a measurement network. The structure of the feature matching network adopts the Siamese double-branch architecture, and the specific structure of each branch can be flexibly customized. Based on different network models, the test set in the comparison experiment is divided into 6 sub-test sets, the number of correct matching features NCM of the network model is counted, and the matching accuracy MP of each pair of images is calculated separately.
Figure 2. Matching accuracy results of network combined optimization experiment

Based on the above experimental analysis, it can be seen that the use of a small-size convolution kernel for multi-modal image feature extraction can obtain more accurate image local features, thereby improving the matching accuracy of the model. When taking into account the time efficiency and matching accuracy, the mixed-size convolution kernel scheme can be considered first. The small-size convolution kernel is first used to extract local accurate features, and then the larger-size convolution kernel is used to reduce the model complexity and improve the calculation efficiency.

5. Conclusions

This paper focuses on the feature matching problem of multi-modal RSIs, and studies the existing multi-modal image matching methods in dealing with the problems of non-linear gray-scale difference and geometric deformation robustness that exist simultaneously among multi-modal RSIs. A multi-modal RSI feature matching method based on image depth features is proposed to achieve robust multi-modal RSI matching with nonlinear gray-scale differences and geometric deformation. The work of this paper is mainly divided into the following four points: the construction of a multi-modal RSI feature matching network, the construction of a sample set taking into account the distance of the sample, the image matching method based on the depth feature of the multi-modal image, and the comparative analysis of experiments. Existing image feature matching methods, whether they are based on artificially designed feature descriptors or deep learning methods, take into account the nonlinear gray-scale differences between images, but often fail to take into account significant image geometric deformation, and vice versa. The above two image differences are common in multi-modal RSIs, which brings challenges to existing matching methods.

References

[1] QI Bing-jie, LIU Jin-guo, ZHANG Bo-yan, et al. Research on matching performance of SIFT and SURF algorithms for high resolution RSI [J]. Chinese Optics, 2017, 10(3):331-339.

[2] Chen M, Qin R, He H, et al. A Local Distinctive Features Matching Method for RSIs with Repetitive Patterns [J]. Photogrammetric Engineering and Remote Sensing, 2018, 84(8):513-524.

[3] Huang L, Yang N, Zhang Y. A Point Cloud Optimization Method of Low Altitude RSI Based on Multi-channels [J]. IOP Conference Series: Earth and Environmental Science, 2020, 428(1):012033 (7pp).

[4] Sedaghat A, Mohammadi N. Illumination-Robust RSI matching based on oriented self-similarity [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2019, 153(JUL.): 21-35.
[5] Huan, Liu, Gen-Fu, et al. Multi-source RSI Registration Based on Contourlet Transform and Multiple Feature Fusion [J]. International Journal of Automation and Computing, 2019, v.16(05): 15-28.

[6] He H, Chen M, Chen T, et al. Matching of RSIs with Complex Background Variations via Siamese CNN [J]. Remote Sensing, 2018, 10(3):355.

[7] Peng C, Li Y, Jiao L, et al. Densely Based Multi-Scale and Multi-Modal Fully Convolutional Networks for High-Resolution Remote-Sensing Image Semantic Segmentation [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(8):2612-2626.

[8] Zhan L, Ma J, Sang X, et al. Matching Method of Lunar RSI Based on Laplacian [J]. IOP Conference Series: Materials Science and Engineering, 2020, 768(6):062083 (10pp).

[9] Liang A, Li Q, Chen Z, et al. Spherically Optimized RANSAC Aided by an IMU for Fisheye Image Matching [J]. Remote Sensing, 2021, 13(10):2017.

[10] Jeong S, Howat I M, Ahn Y. Improved Multiple Matching Method for Observing Glacier Motion with Repeat Image Feature Tracking [J]. IEEE Transactions on Geoscience & Remote Sensing, 2017, 55(4): 2431-2441.

[11] Ye Y, Shen L, Chen M, et al. An Automatic Matching Method Based on Local Phase Feature Descriptor for Multi-source RSIs [J]. Wuhan Daxue Xuebao, 2017, 42(9):1278-1284.

[12] Liu Y, Mo F, Tao P. Matching Multi-Source Optical Satellite Imagery Exploiting a Multi-Stage Approach[J]. Remote Sensing, 2017, 9(12): 1249.