SCIREA Journal of Geosciences

http://www.scirea.org/journal/Geosciences

August 7, 2022

Volume 6, Issue 2, April 2022

https://doi.org/10.54647/geosciences17182

An Improved Receptive Fields Network for Matching Remote Sensing Images

Wannan Zhang

School of Computer, Central South University, China.

Abstract

We present a new network combining Residual Network (ResNet) and Receptive Fields Network (RF-Net) for matching remote sensing images. Firstly, a new remote sensing image datasets are setup, which consist of images and homograph matrices. The images are obtained by cropping, illumination changing and affine transforming of the original remote sensing images. The matrices are obtained by calculating the homograph between different images of one sequence. Next, a dual-channel network structure is proposed for keypoints detection. The network consists of Receptive Fields Detection (RF-Det) and ResNet for extracting receptive feature maps with detail information and the deep layer maps with semantic information. Then descriptors of these keypoints are generated using a L2-Net. Finally, the strategies of the nearest neighbor, nearest neighbor with a threshold and nearest neighbor distance ratio are used for matching descriptors. Experimental results demonstrate its superior matching performance with respect to the original RF-Net.

Index Terms： Remote sensing images, Registration, ResNet, RF-Net.

I. INTRODUCTION

Image matching is a key process for achieving geometric alignment of multiple images from the same or different sensors with different viewing angles and different phases [1]. The methods for image matching can be generally divided into three categories: grayscale and template-based methods, domain transformation-based methods and feature-based methods. Among them, the feature-based matching method is robust and suitable for more application scenarios. Typical feature-based methods include two parts: key point detection and feature descriptor extraction. Recently, feature-based deep learning image matching is a very important matching technology. During the deep learning image matching framework, some methods are only used to detect key points, and some methods are only used to extract feature descriptors. The key point detection part constructs a response map, a response graph with rich feature information is conducive to detecting more key points; the feature descriptor extraction part trains the feature descriptor end-to-end, and the feature descriptor is a feature vector used to describe a key point in the image. Key point detection and feature descriptor extraction are optimized based on different targets. The gains of the two parts cannot be directly superimposed. It is difficult to train these two parts in the same channel to achieve better results [2]. Therefore the problem that, how to train jointly the key point detection and feature descriptor extraction for better collaborate with each other, needs to be solved during deep learning image matching framework.

LIFT [3] is one of the first networks to realize the joint training of key point detection and feature descriptor extraction. The network uses the image block of the feature point in Scale Invariant Feature Transform (SIFT) [4] as the input. The effect of key point detection is similar to the SIFT algorithm with good robustness, but it cannot accurately extract the feature points of smooth edges. Unlike the LIFT method, SuperPoint [5] uses self-supervised training feature point detection and feature descriptor extraction, the feature is extracted through VGG [6], the key point detection part needs to be pre-trained on the synthetic image data set, and the entire network needs to be trained on the synthetically transformed image. LF-Net [2] uses the Siamense structure [7] without the need of any manual method, it generates a feature map through the deep feature extraction network, which can extract deep features of larger receptive field from the input image, but the shallow features will be lost. RF-Net [8] proposed a new key point extraction method based on receptive fields based on LF-Net, which retains shallow features and obtains a more informative scale space and response map. In the feature descriptor extraction module, it adopts the network structure which is consistent with Hard-Net [9], and proposes a general loss function term to solve the negative impact caused by the pixel position shift after the rigid transformation of the image.

The depth of the network is essential for learning features with stronger representing ability. Deep features contain rich semantic information, but as the depth of the network increases, it will cause gradients to disappear or explode and network degradation. ResNet [10] is one of the most popular feature extraction networks. In order to solve the degradation problem of the network, the basic residual learning is introduced; in order to overcome the problems of gradient disappearance and gradient explosion, batch normalization [13] is used, and the activation function is replaced with a linear rectification function [12] Rectified Linear Unit (ReLU). Deep features with strong characterization capabilities can better locate the key points of salient targets, and the deep feature maps obtained by ResNet help key point detection. RF-Net uses public data sets for training, which can only detect fewer key points when matching remote sensing images, and the mismatch rate is high. And the key point detection module has a shallow network, limited receptive fields, and lacks high-level semantic information.

In this paper, we present a new network combining ResNet and RF-Net for matching remote sensing images. Firstly, a new remote sensing image datasets are setup, which consist of images and homograph matrices. The images are obtained by cropping, illumination changing and affine transforming of the original remote sensing images. The matrices are obtained by calculating the homograph between different images of one sequence. Next, a dual-channel network structure is proposed for keypoints detection. The network consists of RF-Det and ResNet for extracting receptive feature maps with detail information and the deep layer maps with semantic information. Then descriptors of these keypoints are generated using a L2-Net. Finally, the strategies of the nearest neighbor, nearest neighbor with a threshold and nearest neighbor distance ratio are used for matching descriptors. Experimental results demonstrate its superior matching performance with respect to the original RF-Net.

II. METHODOLOGY

The network based on ResNet and RF-Net is composed of two parts: key point detection and feature descriptor extraction. The key point detection part builds a dual-channel key point detection network fused by two sets of convolutional neural networks, and acquires rich content through deep network channels. The high-level features of semantic information are fused with the low-level features of the increasing receptive field in the shallow network channel; the feature descriptor extraction part uses L2-Net [13] to train the input image pairs and extract the feature descriptors. During the training period, the key point network inputs the data set image, and outputs the score map, the direction map and the scale map, which represent the spatial position, direction attribute and scale information of the key points. The key points are intercepted by these images. Point the image pair and send it to the feature description sub-network to generate a fixed-length feature vector for matching.

A. Structure of Dual-channel Network

In the key point detection part of this paper, the network is constructed in a dual-channel way. The two channels conclude the shallow feature extraction channel with increasing RF-Det receptive field and the deep global information feature extraction channel of ResNet. The shallow features extracted by the former have rich details, the deep features extracted by the latter contain more representative global information. These two different features are connected to extract the most visually distinguishable feature map. The first channel consists of 3*3 convolution, instance normalization regular function and the ReLU activation function are formed by the convolution method of the feature pyramid FPN [14]. The second channel is the ResNet-50. After the shallow features and deep features are fused, 1*1 convolution and instance regularization function are used for the side output to generate the required multi-scale response map. The response map obtained by the improved network proposed in this paper requires a small receptive field, but contains rich deep semantic information and shallow detailed information for detecting key points with more salient targets.

B. Key points Detection and Description

Select high-response pixels as key points, and the response map h_n represents pixel responses on multiple scales. In this paper, we design a key point detection structure which is similar to RF-Net and LF-Net. Two Softmax logistic regressions are performed to generate score maps. The first Softmax uses a sliding window of 15*15*N to generate an undistorted response map h_n, and then merges all h_n into the final score map S through the second Softmax.

$$P_n = Soft\max(h_n) \tag{1}$$

$$\mathrm{S} = \sum_n h_n \Theta P_n \tag{2}$$

In the formula, Θ is the Hada code matrix, and P_n represents the possibility that the pixel is the key point [8]. For the direction map {θn}, the values represent the sine and cosine of the direction, and the angle is calculated by the arctan function. Similarly, θn is fused into the final direction map M.

$$M = \sum_n \theta_n \Theta P_n \tag{3}$$

The same operation is also used to generate the scale map U. In the formula, U_n represents the receptive field range of h_n.

$$\mathrm{U} = \sum_n U_n \Theta P_n \tag{4}$$

In this paper, L2-Net is used to extract feature descriptors. The network consists of six 3*3 convolutional layers and one 8*8 convolutional layer. After the convolutional layer, ReLU activation function will be added and batch normalized. The filter size of the last convolutional layer of the descriptor extraction network is too large which will burden the network, but it can obtain a 128-dimensional tensor containing rich information and convert it into a feature vector to describe the key points.

C. Loss Function

The key point detection network predicts the position, direction and scale of the key point. Its loss function is composed of the score map loss and the image pair loss. The score map loss refers to the image pair I_i and I_j input into the network. The score graphs S_i and S_j are used to generate a reference standard image G_j through S_j, and the mean square error (MSE) between Si and G_j is calculated. G_j refers to extracting key points from the deformed S_j, and use Gaussian convolution (σ = 0.5) to get a clean GT.

$$G_j = g[t\{w(S_j)\}] \tag{5}$$

In the formula, w, t, g represent the deformation, the process of selecting key points and Gaussian convolution. The calculation formula of the score graph loss is as follows:

$$\mathrm{L}_{score\text{-}loss}(S_i, G_j) = \left|S_i - G_j\right|^2 \tag{6}$$

The image pair loss is used to optimize the detection network so that the image pairs under the relevant response area are as similar as possible. Key points are selected from G_j, and then their spatial coordinates are deformed to the input image I_j. The key points are composed by the spatial position coordinates, the obtained direction and scale information which are obtained from prediction. Send the corresponding image pair to the feature descriptor extraction network to generate E_i^k, $\bar{E}_j^k$.

$$L_{patch-loss}=\frac{1}{K}\sum_{k=1}^{K}\sqrt{2-2E_i^k\bar{E}_j^k} \tag{7}$$

In summary, the key point detection network loss function is:

$$\mathrm{L}_{\det}=L_{score-loss}+L_{patch-loss} \tag{8}$$

Description loss introduces the feature descriptor loss function L_{des} from Hard-Net [9], which is used to maximize the distance between the nearest positive example and the nearest negative example, so that the feature descriptor training is more stable.

$$\mathrm{L}_{des}(\tau_{pos},\tau_{ng})=\frac{1}{K}\sum_{k=1}^{K}\max(0,1+\tau_{pos}-\tau_{ng}) \tag{9}$$

$$\tau_{\mathrm{pos}}(E_i^k\bar{E}_j^k)=d(E_i^k\bar{E}_j^k)=\sqrt{2-2E_i^k\bar{E}_j^k} \tag{10}$$

$$\tau_{\mathrm{ng}}=\min[d(E_i^k\bar{E}_j^n),d(E_i^m\bar{E}_j^k)] \tag{11}$$

In the formula, $\bar{E}_j^n$ is the nearest non-matching feature descriptor of E_i^k, and E_i^m is the nearest non-matching feature descriptor of $\bar{E}_j^k$.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Training Datasets and Evaluation

In view of the lack of remote sensing data for training in the existing matching network which results in poor test results on remote sensing images, this paper constructs a remote sensing image data set which performs cropping, illumination transformation and affine transformation on real remote sensing images. Another 9 images including brightness changes and viewpoint changes are generated, and the homography matrix between the original image and each generated image is calculated, and all the images and the homography matrix form a sequence. The data set consists of 50 different sequences and 500 images.

The matching standard depends on the matching strategy. In this paper, three matching strategies are used to calculate the matching score for quantitative evaluation [15]. Strategy 1: Nearest Neighbor (NN). Under this standard, each descriptor can only have one match. Two regions A and B, if and only if their descriptors D_B and D_A are nearest neighbor descriptors then A and B are matched; Strategy 2: Nearest Neighbor with a Threshold (NNT). Two regions A and B, when the descriptors D_B and D_A are nearest neighbor descriptors and the distance between the two is less than the threshold t then A and B are matched; Strategy 3: Nearest Neighbor distance ratio (NNR). Two regions A and B , D_B and D_C are the nearest neighbor and second nearest neighbor descriptor of D_A respectively. When $||D_A-D_B|| / ||D_A-D_C|| <t$, A matches with B. These three matching strategies are used to measure the accuracy of matching and the number of key points. During the training process, all the learned descriptors are L2 regularization, and the distance range is [0, 2]. The threshold t of NNT and NNR are set to 1 and 0.7, respectively.

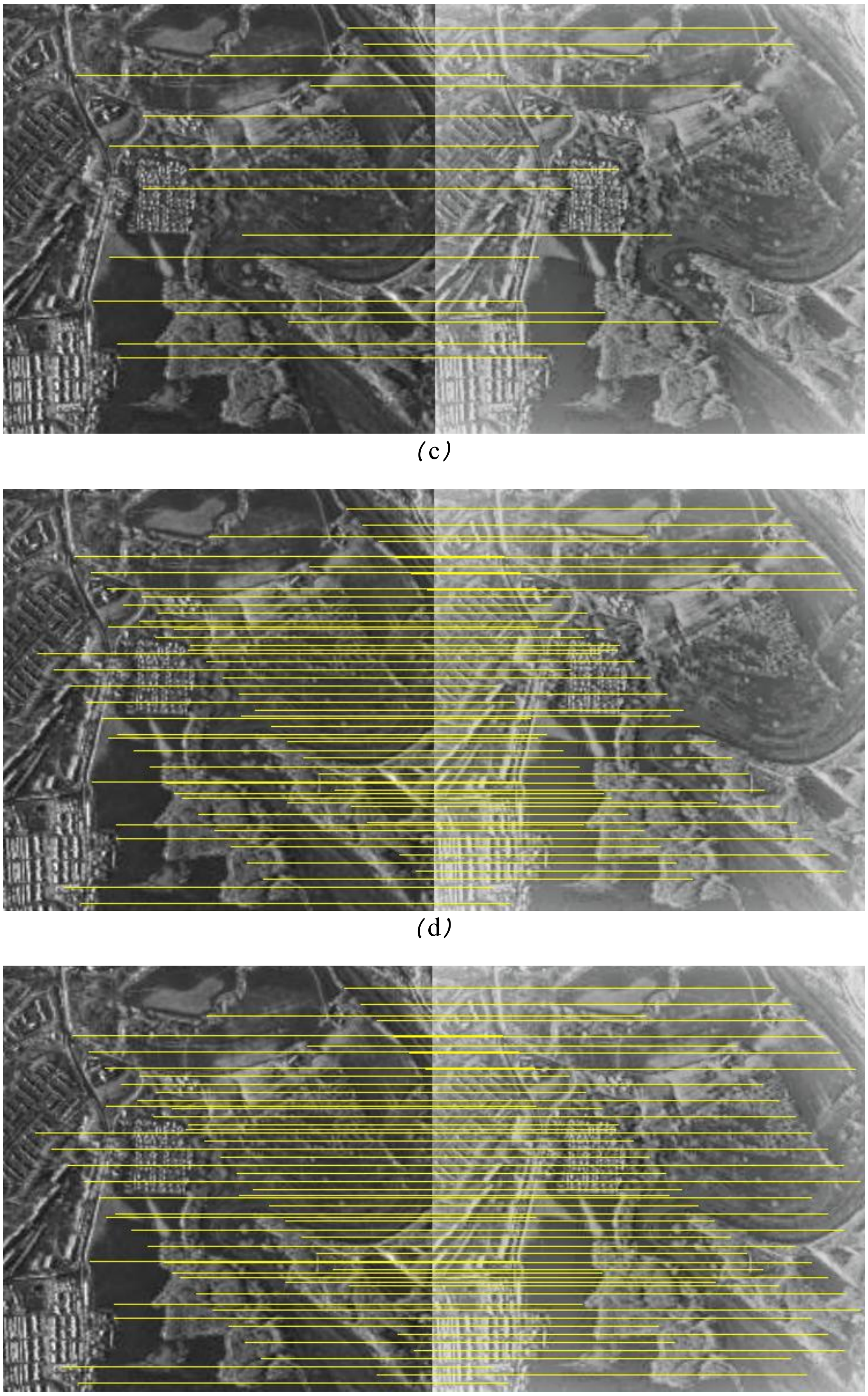
B. Training Process

This paper conducts training and testing on the constructed remote sensing image data set which is divided into training set and test set at a ratio of 9:1. 500 remote sensing images are used for training and 60 remote sensing images are used for testing. During the training stage, change the image scale to 320×240 and convert grayscale transformation. For the descriptor extraction network part, the 32×32 size image is cropped around the key points and input to the network for training. Extract 512 key points in the training phase, but in the testing phase the number of key points can be chosen, and 512 128-dimensional feature vectors are obtained through the descriptor sub-network. Use adaptive moment estimation Adam [16] for optimization, and the initial learning rate set to 0.1, and train two descriptor sub-networks and one key point detection network at the same time.

C. Matching Results

According to the SIFT algorithm, RF-Net and the improved network of this paper, two remote sensing images with the same scene and different affine transformations are matched. Matching results are shown in Figure.1 and Table 1..

(a) (b)



(c)

(d)

(e)

Fig.1. (a) Original image; (b) Image after affine transformation; Matching results using (c) SIFT, (d) RF-Net, and (e) the proposed method.

The matching effect of RF-Net and the improved network of this paper is significantly better than that of the traditional SIFT algorithm. It can be intuitively seen from the figure that the key point matching logarithm of RF-Net and the improved network of the proposed method is significantly more than that of the SIFT algorithm. Deep learning convolutional neural network has powerful feature extraction capability which is helpful for the detection of key points in the image. The RF-Net and the improved network proposed in this paper are matched on the data set using NN, NNT and NNR strategies respectively. Their matching scores and average matching scores (MeanScore) are shown in Table 1.

Table 1. TEST RESULTS

Network	Matching Strategy			
	NN	NNT	NNR	Mean Score
RF-Net	0.126	0.405	0.349	0.293
Proposed	0.347	0.421	0.405	0.391

It can be seen from Table 1 that the average score of the network in this paper is 0.098 higher than that of RF-Net, and the matching accuracy is much higher. In summary, the improved network proposed in this paper has better ability of key point extraction and matching than RF-Net. Especially, it has better robust performance against affine transformation which can be better applied to remote sensing image matching.

IV. CONCULSION

In this paper, we present a new network combining ResNet and RF-Net for matching remote sensing images. Firstly, a new remote sensing image datasets are setup, which consist of images and homograph matrices. The images are obtained by cropping, illumination changing and affine transforming of the original remote sensing images. The matrices are obtained by calculating the homograph between different images of one sequence. Next, a dual-channel network structure is proposed for keypoints detection. The network consists of RF-Det and ResNet for extracting receptive feature maps with detail information and the deep layer maps with semantic information. Then descriptors of these keypoints are generated using a L2-Net. Finally, the strategies of the nearest neighbor, nearest neighbor with a threshold and nearest neighbor distance ratio are used for matching descriptors. Experimental results demonstrate its superior matching performance with respect to the original RF-Net.

REFERENCE

[1] Barbara Zitová, Flusser J . Image Registration Methods: A Survey[J]. Image and Vision Computing, 2003, 21(11):977-1000.

[2] Ono Y , Trulls E , Fua P , et al, “LF-Net: Learning Local Features from Images”, 2018.

[3] Yi K M , Trulls E , Lepetit V , et al. LIFT: Learned Invariant Feature Transform[C]// European Conference on Computer Vision. Springer, Cham, 2016.

[4] Lowe D G . Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[5] A P Y , A Y T , A Y T , et al. Unsupervised Learning Framework for Interest Point Detection and Description via Properties Optimization[J]. Pattern Recognition, 2021.

[6] Li M , Wang Y . An Energy-Efficient Silicon Photonic-Assisted Deep Learning Accelerator for Big Data[J]. Wireless Communications and Mobile Computing, 2020, 2020:1-11.

[7] Chopra S , Hadsell R , Lecun Y . Learning a similarity metric discriminatively, with application to face verification[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005.

[8] Shen X , Wang C , Li X , et al. RF-Net: An End-To-End Image Matching Network Based on Receptive Field[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019..

[9] Wang S , Guo X , Tie Y , et al. Local Feature Descriptor Learning with a Dual Hard Sampling Strategy[C]// 2019 IEEE International Symposium on Multimedia (ISM). IEEE, 2020.

[10] He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2016.

[11] Katz G , Barrett C , Dill D , et al. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks[C]// International Conference on Computer Aided Verification. 2017.

[12] Gulcehre C , Cho K , Pascanu R , et al. Learned-Norm Pooling for Deep Feedforward and Recurrent Neural Networks[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2014.

[13] Gulcehre C , Cho K , Pascanu R , et al. Learned-Norm Pooling for Deep Feedforward and Recurrent Neural Networks[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2014.

[14] Zhao X , Li W , Zhang Y , et al. Aggregated Residual Dilation Based Feature Pyramid Network for Object Detection[J]. IEEE Access, 2019, PP(99):1-1.

[15] MIKOLAJCZYK K,SCHMID C. A performance evaluation of local descriptors[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(10):1615-1630.

[16] Kingma D , Ba J . Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.