THE JOINT SPATIAL AND RADIOMETRIC TRANSFORMER FOR REMOTE SENSING IMAGE RETRIEVAL

Y. Wang [1], D. Yu [1], S. Ji [1, *], Q. Cheng [2], M. Luo [2]

[1] School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan, China - (ymw, yudawen, jishunping)@whu.edu.cn
[2] Wuhan Geomatics Institute, Wuhan, China - (chengqi19800725, luominghai)@163.com

Commission III, WG III/1

KEY WORDS: Remote Sensing Image Retrieval, Convolutional Neural Network, Spatial Transformation, Radiometric Transformation, Parameter Generation Network.

ABSTRACT:

Content-based remote sensing image retrieval refers to searching interested images from a remote sensing image dataset that are similar to a query image via extracting features (contents) from images and comparing their similarity. In this work, we come up with a lightweight network structure, which we call the joint spatial and radiometric transformer, which is composed of three modules: parameter generation network (PGN), spatial conversion and radiometric conversion. The PGN module learns specific transformation parameters from input images to guide subsequent spatial and radiometric conversion processes. With these parameters, the spatial conversion and radiometric conversion transform the input images with spatial and spectrum perspectives respectively, to increase the intra-class similarity and inter-class difference, which are attached great importance to CBRSIR. In comparative experiments on multiple remote sensing image retrieval datasets, our proposed joint spatial and radiometric transformer combined with the backbone network ResNet34 has achieved optimal performance.

1. INTRODUCTION

Content-based image retrieval (CBIR) is a hot research topic both in computer vision and remote sensing (Du *et al*., 2016). A query process of CBIR consists of three steps: calculate the features of the query image and of the images in the chosen database; compare the similarity of the features; rank the images in the database according to the similarity score. As for remote sensing images, geometric deformation caused by various camera angles from overhead platforms and complex radiometric distortions caused by a dynamic atmosphere both impose higher requirements on the retrieval technology.

In recent years, convolutional neural network, which has been widely used in many fields, has shown excellent performance in the domain of remote sensing image retrieval. In the classic convolution neural network structures (e.g. Vgg, ResNet) (Krizhevsky *et al*., 2018; Simonyan and Zisserman, 2014; Szegedy *et al*., 2015; He et al., 2016), the simple and straight convolution and maxpooling operation can indeed achieve some translation invariance to a certain extent. However, using the fixed size of the convolution window and pooling unit, the geometric translation invariance may not be fully achieved when processing remote sensing images. On the other hand, the commonly used data augmentation for color transformation hardly handle with radiometric distortions completely.

In this paper, we propose a lightweight network structure performing spatial and radiometric conversion simultaneously on remote sensing images, which is robust to the diversity of perspective angles and radiometric situations of input images without extra supervision. This main part of the structure, called transformer, learns dedicated spatial and radiometric transformations for each individual image. According to the classification loss at the training stage, the transformer tends to learn the conversion parameters to perform spatial and radiometric correction on the input image, and generate a corrected image that is more conducive to the subsequent feature extraction. On one hand, spatial correction mainly aims at making the foreground more prominent, which can be regarded as an attention mechanism, and also achieves affine and some non-rigid deformation through the defined transformation model. On the other hand, the radiometric transformation will reduce the intra-class variability through spectral correction, which is particularly important for retrieval. Our method, combined with the backbone network ResNet34, demonstrates excellent performance on multiple popular remote sensing image retrieval datasets.

2. METHOD DESCRIPTION

The model of spatial and radiometric transformation consists of three modules: parameter generation network (PGN), spatial conversion and radiometric conversion.

Parameter generation network is composed of two convolutional layers with the maxpooling layer following respectively and a multi-layer perceptron with one hidden layer , through which a fixed number of parameters that are based on the selected spatial and radiometric transformers are obtained by the last regression layer. Ten parameters are regressed in our experiment.

* Corresponding author

The spatial conversion module, which performs grid generating and sampling in turn, borrows ideas from Jaderberg et al (Jaderberg *et al*., 2016). The grid generator transforms the spatial coordinates of the input image by using the parameters obtained by the PGN and the defined spatial transformation model. The sampling module resamples the input images with the transformation model and a specific interpolation method. In this paper, we choose affine transformation as the spatial transformation model, which requires 6 transformation parameters. Affine transformation is a linear transformation from one 2D coordinate to the other, which can be divided into a series of single transformations, including translation, scale, flip, rotation and shear. The interpolation method we choose is the bilinear interpolation.

Compared with spatial transformer, which has appeared as similar versions such as attention mechanism in previous works, radiometric conversion has not attached much significance in image retrieval tasks. For remote sensing image retrieval tasks, the radiometric correction of the image makes the network having the ability to actively learn to increase intra-class similarity and inter-class difference at the spectral level. In this study, we apply four transformation parameters obtained by PGN on the input image for radiometric correction. We observed that a variety of different satellite images covering the same area, including illumination change, under- or over-exposure, color cast, can be largely modeled and repaired by adjusting different spectral channels. Therefore, we set the four parameters as a linear stretching coefficient respectively to the R, G and B channels with the same translation bias. The input image is then transformed pixel by pixel according to the stretching parameters.

After the original image is adjusted by the transformer model, the corrected image is inputted into the shortcut structures of the ResNet to extract features. In the training phase, which is the same as a common classification task, the features are processed by two full connected layers to output predictions. The outputs are compared with the ground truths to optimize the whole network consisting of the transformer and the Resnet. During the retrieval phase, the last fully connected layer is replaced with principal component analysis (PCA), which outputs a feature vector with a fixed length. Then, the normalized correlation coefficients (NCC) are calculated between the feature vectors extracted from the query image and from any image in the database to be retrieved. The NCC score ranks the images in database.

The whole process of our proposed mothed is shown in Figure 1.

3. EXPERIMENTS AND RESULTS

3.1 Data Used

We use PatternNet (Zhou *et al*., 2018) as the fine-tuning dataset to transfer a model pretrained on close range ImageNet dataset adaptive to overhead images, and RS19 (Xia *et al*., 2010), UCM (Yang and Newsam, 2010) and RSSCN (Zou *et al*., 2015) as the test dataset for retrieval.

PatternNet is a large-scale high-resolution remote sensing dataset specifically designed for RSIR. The dataset has a total of 30400 images each of which size 256×256.

WHU-RS19 contains 19 categories, total of 1005 remote sensing images, which can be used for scene classification and retrieval. This dataset has around 50 images of each type, and each image is 600 × 600 pixels in size.

The UC Merced Land-Use Dataset contains 21 types of scenes, each of which is composed of 100 images. The size of each image is 256 × 256 pixels.

RSSCN7 consists of 7 typical scene categories and 2800 images. Each category contains 400 images of size 400 × 400, averagely sampled from 4 different scales.

3.2 Setting

The proposed network was pre-trained on the ImageNet dataset for weight initialization. The input images for fine-tuning and retrieval were all resized to 224×224 pixels. In the fine-tuning phase, 40 epochs were conducted, among which the learning rates of 1st to 15th epoch were set to 10^{-3}, those of 16th to 30th epoch were 10^{-4}, and those of 31st to 40th epoch were 10^{-5}. The batch size was set to 64 and the optimizer was SGD (Adam for the compared compact bilinear pooling (CBP) method(Wang *et al*., 2020)). A Linux PC with an NVIDIA GeForce GTX 1060 6G GPU and the PyTorch deep learning environment was used.

We use mean Average Precision (mAP), Precision at k (P@k) where k indicates the top k retrieval results in a query, to evaluate the retrieval performances of different methods. The mean Average Precision (mAP) is the average of AP where AP means the average of the correct rates on different recalls in a query.

3.3 Experimental Results

The retrieval results on dataset RS19, UCM and RSSCN are shown in Tables 1, 2 and 3 respectively. The content in parentheses represents the modules (decoder) after the shortcut structures (encoder) in the retrieval network, in which FC is the fully connected layer, and PCA stands for principal component analysis. Respectively, ST and RT are abbreviations of spatial and radiometric transformation.

It can be seen that in the three remote sensing image retrieval datasets, our transformer model all achieved the highest accuracy on mAP, surpassing the newest classification network NTS-Net, SENet, SKNet and that attention boosted bilinear pooling (ATT + CBP). The groups replacing the last FC layer with PCA get better results in all the controlled experiments. To prove the effectiveness of our proposed joint transformer, we tested the case of adding a single spatial transformer or a single radiometric transformer, respectively. Their retrieval results are both better than those of the simple resnet34 network, but inferior to those of the network that added the joint spatial and radiometric transformer, which indicates that the single spatial or radiometric transformer can indeed learn the transformation parameters that are conducive to retrieval, what's more, the joint spatial and radiometric transformer can effectively integrate the advantages of the single spatial and radiometric transformer.

The retrieval results on different datasets are shown in Figure 2. In each figure, the query image is shown on the first row, and the result of ResNet34(PCA) and ST+RT+ResNet34(FC1+PCA) are shown on the second and third row respectively. The red box indicates that the image is irrelevant to the query image and

wrongly predicted by algorithm, and the green one means relevant and correctly predicted.

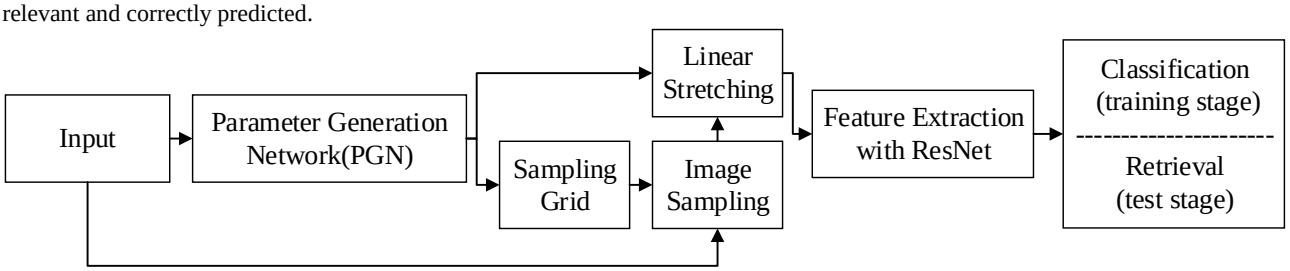

Figure 1. The flowchart of the proposed method.

Method	mAP	P@5	P@10	P@50	P@100
ResNet34(FC)	0.8867	0.8469	0.8087	0.5755	0.3854
ResNet34(PCA)	0.9131	0.8745	0.8291	0.5474	0.3475
ResNet34(ATT+CBP+PCA)	0.8951	0.8490	0.7974	0.6026	0.3901
NTS-Net(PCA) (Yang et al. 2018)	0.8994	0.8562	0.8040	0.6187	**0.4758**
SENet(FC1+PCA) (Hu et al. 2018)	0.9318	0.9051	0.8816	0.6593	0.4209
SKNet(FC1+PCA) (Li et al. 2018)	0.8901	0.8480	0.8015	0.6301	0.4081
ST+ResNet34(FC1+FC2)	0.8933	0.8541	0.8179	0.6043	0.3985
ST+ResNet34(FC1+PCA)	0.9323	0.9092	0.8827	0.6889	0.4316
RT+ResNet34(FC1+FC2)	0.8686	0.8153	0.7828	0.5650	0.3838
RT+ResNet34(FC1+PCA)	0.9181	0.8908	0.8673	0.6579	0.4204
ST+RT+ResNet34(FC1+FC2)	0.9099	0.8755	0.8413	0.6129	0.4019
ST+RT+ResNet34(FC1+PCA)	**0.9432**	**0.9194**	**0.8964**	**0.6974**	0.4354

Table 1. The performance of different methods on RS19.

Method	mAP	P@5	P@10	P@50	P@100
ResNet34(FC)	0.9096	0.8605	0.8162	0.6669	0.5436
ResNet34(PCA)	0.8994	0.8562	0.8040	0.6187	0.4758
ResNet34(ATT + CBP + PCA)	0.9056	0.8638	0.8367	0.7227	**0.5939**
NTS-Net(PCA)	0.8352	0.7610	0.6950	0.4560	0.3438
SENet(FC1+PCA)	0.9167	0.8786	0.8419	0.7003	0.5771
SKNet(FC1+PCA)	0.9018	0.8610	0.8167	0.6895	0.5774
ST+ResNet34(FC1+FC2)	0.9028	0.8657	0.8202	0.6791	0.5554
ST+ResNet34(FC1+PCA)	0.9233	**0.8919**	**0.8579**	0.7116	0.5861
RT+ResNet34(FC1+FC2)	0.8832	0.8405	0.8031	0.6594	0.5384
RT+ResNet34(FC1+PCA)	0.9163	0.8800	0.8479	0.7022	0.5706
ST+RT+ResNet34(FC1+FC2)	0.9089	0.8729	0.8333	0.6874	0.5634
ST+RT+ResNet34(FC1+PCA)	**0.9290**	0.8871	0.8536	**0.7118**	0.5900

Table 2. The performance of different methods on UCM.

Method	mAP	P@5	P@10	P@50	P@100
ResNet34(FC)	0.8297	0.7754	0.7562	0.6755	0.6266
ResNet34(PCA)	0.8562	0.8129	0.7884	0.6862	0.6188

ResNet34(ATT + CBP + PCA)	0.8132	0.7550	0.7312	0.6542	0.6045
NTS-Net(PCA)	0.8437	0.7914	0.7686	0.6808	0.6281
SENet(FC1+PCA)	0.8757	0.8407	0.8187	0.7423	0.6918
SKNet(FC1+PCA)	0.8636	0.8229	0.7991	0.7214	0.6727
ST+ResNet34(FC1+FC2)	0.8400	0.7929	0.7764	0.6943	0.6452
ST+ResNet34(FC1+PCA)	0.8707	0.8339	0.8155	0.7421	0.6904
RT+ResNet34(FC1+FC2)	0.8235	0.7657	0.7384	0.6589	0.6079
RT+ResNet34(FC1+PCA)	0.8640	0.8225	0.7973	0.7074	0.6533
ST+RT+ResNet34(FC1+FC2)	0.8406	0.7907	0.7696	0.6964	0.6488
ST+RT+ResNet34(FC1+PCA)	**0.8889**	**0.8475**	**0.8307**	**0.7658**	**0.7156**

Table 3. The performance of different methods on RSSCN.

Figure 2. The top ten result images from RS-19 returned by ResNet34(PCA) and ST+RT+ResNet34(FC1+PCA).

Figure 3. The top ten result images from UCM returned by ResNet34(PCA) and ST+RT+ResNet34(FC1+PCA).

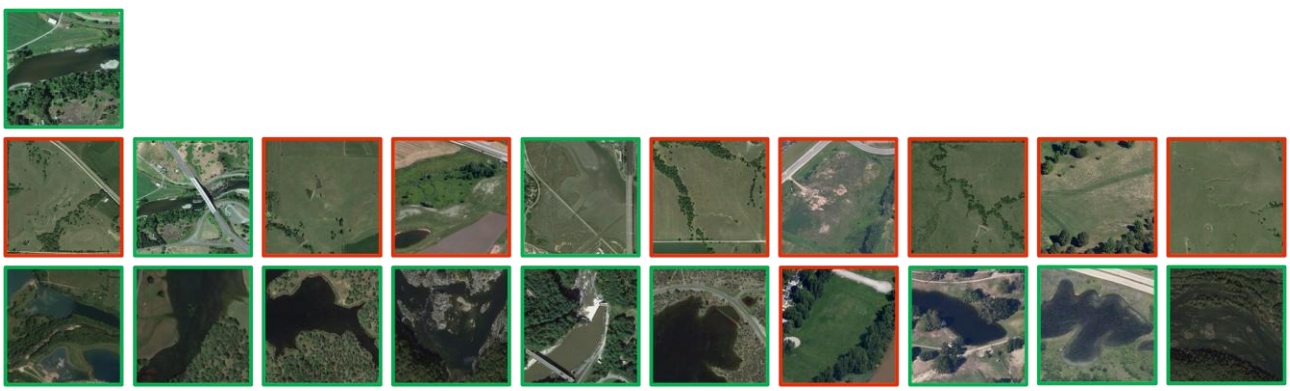

Figure 4. The top ten retrieval images from RSSCN returned by ResNet34(PCA) and ST+RT+ResNet34(FC1+PCA).

Figure 2, 3 and 4 demonstrated that the introduction of the ST and RT combination obviously improves the retrieval performance of the baseline.

CONCLUSION

In this paper, we propose a joint spatial and radiometric transformer to converse the input image for image retrieval. Specifically, a spatial conversion can be regarded as an attention mechanism to make the foreground of the image more prominent, while radiometric conversion uses the parameters obtained by actively learning to increase the intra-class similarity and inter-class difference at the spectral level. Experiments on multiple challenging remote sensing image retrieval datasets show that our joint transformer surpasses the popular latest networks such as SENet and effectively improves retrieval accuracy.

Compared with the FC layer that learns parameters from the fine-tuning dataset and output features more discriminatory, PCA which is unrelated to any specific dataset is more universal in terms of output feature vectors, therefore replacing the last FC layer with PCA in the retrieval process can achieve better results.

REFERENCES

Du, P.; Chen, Y.; Hong, T.; Tao, F., 2005. Study on content-based remote sensing image retrieval. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, Seoul, p. 4.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Amsterdam, pp. 770–778.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City (UT), pp. 7132-7141.

Jaderberg, M., Simonyan, K., Zisserman, A., 2016. Spatial transformer networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Angeles (CA), pp. 2017-2025.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. pp. 1097-1105.

Li, X., Wang, W., Hu, X., Yang, J., 2019. Selective kernel networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Angeles (CA), pp. 510-519.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556.

Szegedy, C., et al., 2015. Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston (MA), pp. 1–9.

Wang, Y., Ji, S., Lu, M., Zhang, Y., 2020. Attention boosted bilinear pooling for remote sensing image retrieval. *International Journal of Remote Sensing*. 41(7), 2704-2724.

Xia, G. S., Yang, W., Delon, J., Gousseau, Y., Sun, H., Ma ître, H., 2010. Structural High-resolution Satellite Image Indexing. *Proceedings of ISPRS TC VII Symposium – 100 Years ISPRS*. Vienna, pp. 298-303.

Yang, Y., Newsam, S., 2010. Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification. *Proceedings of the 18th ACM SIGSPATIAL international conference on advances in geographic information systems*. San Jose (CA), pp. 270-279.

Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L., 2018. Learning to navigate for fine-grained classification. *Proceedings of the European Conference on Computer Vision*. Munich (BY), pp. 420-435.

Zhou, W., Newsam, S., Li, C., Shao, Z., 2018. Patternnet: A Benchmark Dataset for Performance Evaluation of Remote Sensing Image Retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*. 145, 197–209.

Zou, Q., Ni, L., Zhang, T., Wang, Q., 2015. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. *IEEE Geoscience and Remote Sensing Letters*. 12(11), 2321–2325.