Journal of Computer and Communications, 2018, 6, 185-193
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Learning Multi-Modality Features for Scene Classification of High-Resolution Remote Sensing Images

Feng'an Zhao*, Xiongmei Zhang, Xiaodong Mu, Zhaoxiang Yi, Zhou Yang

Department of Computer Science and Technology, Xi'an Research Institute of High-Tech, Xi'an, China
Email: *zhao_flying123@163.com

How to cite this paper: Zhao, F.A., Zhang, X.M., Mu, X.D., Yi, Z.X. and Yang, Z. (2018) Learning Multi-Modality Features for Scene Classification of High-Resolution Remote Sensing Images. *Journal of Computer and Communications*, **6**, 185-193.
https://doi.org/10.4236/jcc.2018.611018

Received: November 5, 2018
Accepted: November 12, 2018
Published: November 19, 2018

Abstract

Scene classification of high-resolution remote sensing (HRRS) image is an important research topic and has been applied broadly in many fields. Deep learning method has shown its high potential to in this domain, owing to its powerful learning ability of characterizing complex patterns. However the deep learning methods omit some global and local information of the HRRS image. To this end, in this article we show efforts to adopt explicit global and local information to provide complementary information to deep models. Specifically, we use a patch based MS-CLBP method to acquire global and local representations, and then we consider a pretrained CNN model as a feature extractor and extract deep hierarchical features from full-connection layers. After fisher vector (FV) encoding, we obtain the holistic visual representation of the scene image. We view the scene classification as a reconstruction procedure and train several class-specific stack denoising autoencoders (SDAEs) of corresponding class, *i.e.*, one SDAE per class, and classify the test image according to the reconstruction error. Experimental results show that our combination method outperforms the state-of-the-art deep learning classification methods without employing fine-tuning.

Keywords

Feature Fusion, Multiple Features, Scene Classification, Stack Denoising Autoencoder

1. Introduction

With the development of remote sensing instrument technologies, more and more high-resolution remote sensing (HRRS) images, which contain detailed spatial information, are now available. To automatically labeling an image from a set of semantic categories is the main purpose of HRRS scene classification, and has become one of the most important applications of the HRRS. However, there exist semantic gap between raw visual data and its semantic category. Constructing discriminating feature representation of HRRS image is one of the most necessary steps to bridge the gap in HRRS scene classification. Low-level handicraft features methods, such as Local Binary Patterns (LBP) [1], capture different properties like texture, spatial global or local structure information of the HRRS scene. Mid-level methods such as the bag-of-visual-words (BOVW) model, probabilistic latent semantic analysis, and Fisher kernel vector [2] is the classical feature representation methods which are used to bridge the semantic gap. However improvements over these methods grow slowly in recent years because of the specificities of HRRS data. In other words, although some of the low-level or mid-level feature methods have performed well on some homogeneous structures, the classification results decrease a lot when the dataset shows more complex structures and spatial layouts.

In recent years, deep learning methods, which construct learning models with multiple processing layers, have shown its great ability of representing high-level features from raw data hierarchically. These deep-learning based methods learn hierarchical feature representation and give a fine classification result when the size of the training sample is sufficiently large. A large number of convolutional neural network (CNN) based methods have been proposed in the field of HRRS scene classification.

Generally speaking, both engineered low-level methods and deep model based methods have their own superiority. The former has advantage in classifying the simple geospatial objects such as the farm land, airports and so on, while the latter prefers the complex geographic images, owing to the generated generic robust deep features. However, these methods are not well-suited for all kinds of geographic images. Besides, most existing approaches use only single-modality features which are insufficient in reflecting various characteristics of the HRRS scene.

Different from existing methods of HRRS scene classification which focus on improving the network architecture or feature encoding method, we investigate how to fuse the hierarchical features and the low-level complementary features. More specifically, we use a low-level feature method, *i.e.*, patch based MS-CLBP, to acquire local representations, and then we extract features from the convolution layers of a pretrained CNN model, which contain rich hierarchical structural information. Both low-level and high-level features are encoded through Fisher Kernel encoding. Thus, we obtain holistic hierarchical and local visual representations. We also compare the performance of the proposed with the state-of-the-art methods. The superiorities of our method in classification accuracy are shown at the end of this study.

2. Method Description

2.1. Low-Level and High-Level Features Extraction and Fisher Kernel Encoding

The flowchart of the proposed method is shown in **Figure 1** and **Figure 2**. To

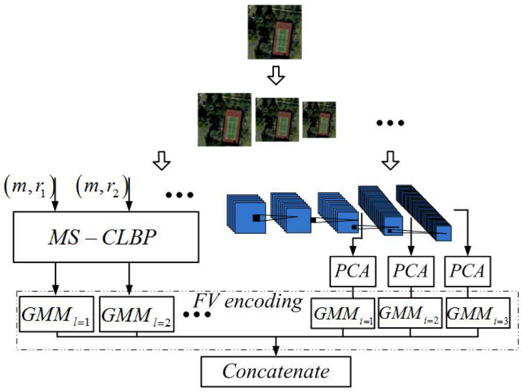

Figure 1. Procedure of low-level and high-level feature extraction and Fisher Kernel encoding.

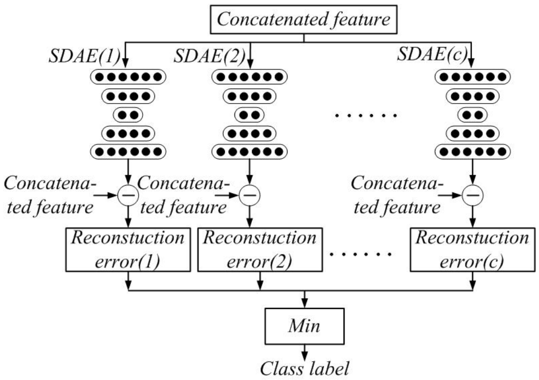

Figure 2. Procedure of classfication using reconstruction error which are output by class-specific SDAEs.

obtain multiscale Fisher vector (FV) of the HRRS image, pyramid algorithm is used to produce different observation scales. These multi-scale scene images are fed into patch based MS-CLBP and a pretrained CNN respectively for extracting multiscale local features and convolutional features, which are then stacked to be encoded by the Fisher kernel.

For the low-level features, we apply the CLBP [4] operator with a parameter pair (m, r_l), $l \in (1, 2, \ldots, t)$ to represent images with two generated CLBP component, a sign component (CLBP_S) and a magnitude component (CLBP_M). Given a center pixel $\mathbf{x}_c$, its *m* neighboring pixels $\mathbf{x}_i$ equally distributed on a circle of radius *r*. CLBP_S is equivalent to the traditional LBP and the CLBP_M is defined as

$$CLBP\_M_{m,r} = \sum_{i=0}^{m-1} p\left(\left|\mathbf{x}_i - \mathbf{x}_c\right|, \gamma\right) \times 2^i, \quad p(\delta, \gamma) = \begin{cases} 1, \delta \geq \gamma \\ 0, \delta < \gamma \end{cases} \tag{1}$$

where γ is the mean value of $\left|\mathbf{x}_i - \mathbf{x}_c\right|$ from the entire image. These two complementary components of CLBP can capture the spatial patterns and contrast of local image texture, such as edge and corners. The CLBP operator with the same parameter pair (m, r) is applied to the multi-scale images to generate patch-based CLBP histogram features [5]. For each patch i, two occurrence histograms are computed from both CLBP_S and CLBP_M. After concatenating the two histograms, we get the histogram feature vector $\mathbf{f}_i$ $j \in (1,2,\ldots,M)$ of patch *j*. Suppose *M* patches are extracted from the multi-scale sub-images of a HRRS scene image, the feature matrix of the scene image can be denote as $\mathbf{\Phi}^{(m,r_l)} = [\mathbf{f}_1,\mathbf{f}_2,\ldots,\mathbf{f}_M]$. Each $\mathbf{f}_j$ of the matrixes corresponds to a histogram feature vector of a patch.

As noted in [1], LBP features generated from single (m,r_l) may not be able to represent the intrinsic texture features. Therefore, different parameter sets $\{(m,r_1),(m,r_2),\ldots,(m,r_t)\}$ are provided by operators of varying the parameter pair (m,r). Specifically, we use fixed number of neighbors *m* and multiple radii r_i in the process of patch based MS-CLBP feature extraction. For each radii $r_l, l \in (1,2,\ldots,t)$, we can obtain a corresponding feature matrix $\mathbf{\Phi}^{(m,r_l)}$, thus we can get a set of *t* feature matrix set $\mathbf{\Phi} = \{\mathbf{\Phi}^{(m,r_1)},\mathbf{\Phi}^{(m,r_2)},\ldots,\mathbf{\Phi}^{(m,r_t)}\}$ for a HRRS scene image. The feature matrix set of the corresponding scene image is of high dimension and cannot be used as a representative feature, then we also use the effective patch aggregation mechanism IFK [2] to characterize the dense local patch descriptors. Given N_T training images, each image $I_q, q \in (1,2,\ldots,N_T)$ can be represent by $\mathbf{\Phi}_q = \{\mathbf{\Phi}_q^{(m,r_1)},\mathbf{\Phi}_q^{(m,r_2)},\ldots,\mathbf{\Phi}_q^{(m,r_t)}\}$ based on above method. For each CLBP parameter pair (m,r_l), we use $\mathbf{\Phi}^l = \{\mathbf{\Phi}_1^{(m,r_l)},\mathbf{\Phi}_2^{(m,r_l)},\ldots,\mathbf{\Phi}_d^{(m,r_l)}\}$ as the feature set of all the training set to estimate the Gaussian mixture model (GMM) [3] parameters via the Expectation Maximization (EM) algorithm. Thus for *t* CLBP parameter sets $\{(m,r_1),(m,r_2),\ldots,(m,r_t)\}$, we obtains *t* GMMs. After the GMM estimating, the concatenated low-level FV features of the testing scene image are obtained.

As for the high-level hierarchical features, suppose the pretrained CNN models contains *k* convolutional layers and denote multi-scale convolutional features on *i*th layer ($i \in (1,2,\ldots,k)$) of a given scene image I_q at scale level *s* as $\{\mathbf{I}_q^{s,i}\}_{s=0}^{n}$. Let the number and size of the filter maps be *u* and *v*. Firstly we flatten the filter maps from *i*th convolutional layer into a set of feature vectors. Then each column of the feature set represents for a *u*-dimensional local descriptor which can be regarded as the feature representation of the corresponding image region. Then PCA is performed to reduce the dimension of each modality to u'. Thus we obtain *v* u'-dimensional multi-scale features of *i*th convolutional layer for the image I_q, which can be defined by

$$\mathbf{\Phi}_q^i = \{\mathbf{I}_q^{s,i}\}_{s=0}^{n} = \{\mathbf{f}_1^i,\mathbf{f}_2^i,\ldots,\mathbf{f}_v^i\} \in \mathbb{R}^d \tag{2}$$

given N_T training images the descriptor set $\mathbf{\Phi}^i = \{\mathbf{\Phi}_1^i,\mathbf{\Phi}_2^i,\ldots,\mathbf{\Phi}_{N_T}^i\}$ are generated from *i*th convolutional layer. Similar as in the FV encoding of low-level features, *k* FVs are generated from the *k* convolutional layers, we denote the high-level hierarchical meaningful feature as $\mathbf{\Phi}_H$, and denote the low-level local and global features as $\mathbf{\Phi}_L$. Then the final feature-level fusion can be formulated as:

$$\mathbf{\Phi}(x) = \alpha\mathbf{\Phi}_H(x) + (1-\alpha)\mathbf{\Phi}_L(x) \tag{3}$$

where the α is the weight parameter, which balance the effect of the high-level and low-level features. $\Phi(x)$ is a final feature representation of test image *x*.

2.2. Classification by Stack Auto-Encoder Reconstruction Error

After obtaining the concatenated FV feature, we conduct the classification in the view of reconstruction error, which is illustrated in **Figure 2**. The denoising autoencoder (DAE) is the enhanced variant of the conventional autoencoder using criterion of denoising. It is more robust since it learns to recover an image of corrupted version. The DAE can be stacked to obtain high level features like CNNs, resulting in stacked denoising autoencoder (SDAE) [6] approach. The SDAE can learn holistic hierarchical representations of an image due to the multi-layers abstraction. For the dataset which contains c scene classes, we first train c class-specific SDAEs using the combined feature of all training samples of the corresponding class, *i.e.* one SDAE for one class. Thus we get the encoding and decoding weights each of the c SDAEs. At testing stage, we feed the combined feature of the testing scene image to each of the SDAEs to generate a c -dimensional reconstruction error vector, and each value of the vector corresponds to a scene class. The SDAE which is trained by images of the same class obtains the minimum reconstruction error. So we assign the class of combined feature to the index of the minimal one in the reconstruction error vector.

3. Experiment Setup and Result

3.1. Dataset

AID [7] dataset, which is a public available HRRS dataset, is adopted to evaluate the performance of our proposed methods. AID dataset is a new large-scale HRRS dataset established for advancing the state-of-the-arts. It consists of 30 scene categories including airport, bare land, baseball field, and so on. There are a total of 10,000 images with the size of 600 × 600 pixels. For increasing the intra-class variability of AID, the scene images are from several countries and acquired under different imaging conditions. **Figure 3** shows some example images of AID dataset.

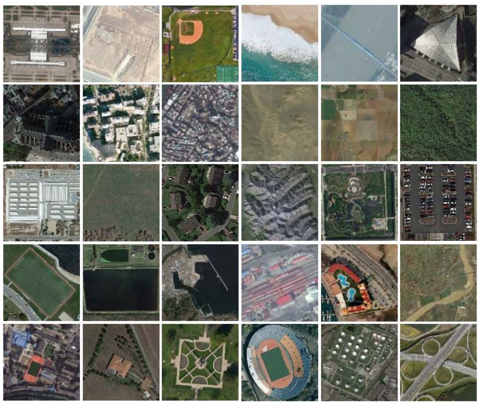

Figure 3. Some example images from AID remote sensing dataset.

3.2. Experimental Settings

We randomly separate the dataset with the ratio of 50% on AID for training, and left for testing. We evaluate our classification performance with the average accuracy over 50 runs. Multi-scale of 150 × 150, 300 × 300, 600 × 600 are set to enhance the classification performance. The number of Gaussian components in GMM which are used for encoding convolutional features and MS-CLBP are empirically set to be 100 and 16 respectively. We set the weight of high-level and low-level features to 0.85 after a series of experiments on AID dataset. For the setting of SDAE, we set three hidden layers with 800, 800 and 300 neurons. As mentioned in the methodology section, different parameter set (m,r) can grasp intrinsic texture features. We empirically set $m=8$ and $r=(1,2,\ldots,6)$. Specifically, 6 radii are used for the parameter set $\{(m=8,r_1=1),\ldots,(m=8,r_6=6)\}$. We then study the number of scales for generating multi-scale image and the patch size used in the patch-based MS-LBP. Different choice of multiple scales contains $\{1,(1,1/2),\ldots,(1,1/6)\}$. For example, $(1,1/6)$ means there are two scales of the image are used, original image and the down sampled image at 1/6 size of the original image. Note that the scale 1/6 has nothing to do with the parameter pair $(m=8,r_6=6)$.

As for the size of patch $(P\times P)$, we empirically set $P\in\{16,24,32,48,64,96\}$, and discuss the effects of different parameter setting pair of image scale and patch size on the classification results, as shown in **Figure 4**. It can be seen that, the combing of scales $(1,1/5)$ and patch size 64 achieves the best result in AID dataset.

3.3. Comparison with the State-of-the-Arts Methods

To evaluate the performance of our proposed method, we make a comparison of our method and the state-of-the-art methods performed on the AID databases

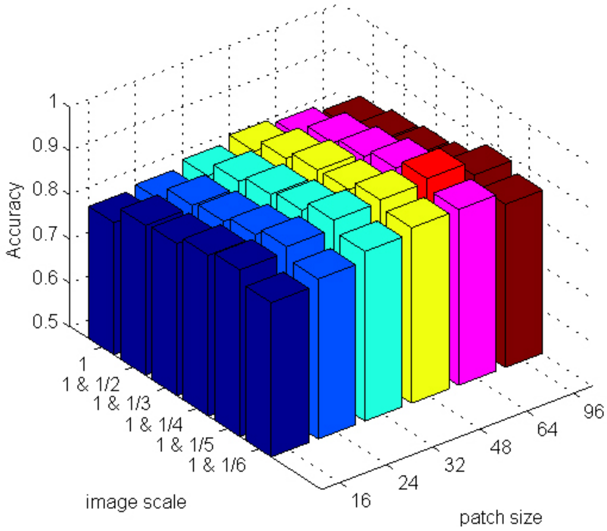

Figure 4. Effects of different parameter settings of scales and patch size on the classification results of the AID dataset.

under the same experimental settings. Due to the high intra-class variability and inter-class similarity of this dataset, the most existing scene classification methods are based on deep neural network. The researchers [7] who set up this dataset achieved accuracy of almost 90% based on CaffeNet, VGG-VD16, and GoogLeNet. [8] proposed a feature fusion strategy based on discriminant correlation analysis (DCA), and achieved accuracy of 89.71% with the smallest feature size of 58. Furthermore, they use the fusion strategy of addition and gain a excellent accuracy of 91.87%. [10] uses deep ResNet to address the problem of training very deep convolutional networks, the classification accuracy achieves of almost 90% via training the top softmax layer of ResNet. Even better result 94.23% is obtained by fine-tuning of ResNet. In this work, we achieved the accuracy of 91.07% ± 0.33%, which is superior to the most state-of-the-art methods. Note that we don't adopt the fine-turn and multi-CNN combination approach in consideration of the computing cost. The high performance of the method benefits from the strong representation power of the fused hierarchical and local features. **Figure 5** shows the confusion matrix for the proposed method. We can observe that the classification accuracies achieve more than 0.8 for most scene class, even though the AID is a fairly challenging large-scale dataset. The scene class of centre, commercial, resort, school and square are easy to confuse with others due to the high intro-class variety and inter-class similarity of the AID dataset (**Table 1**).

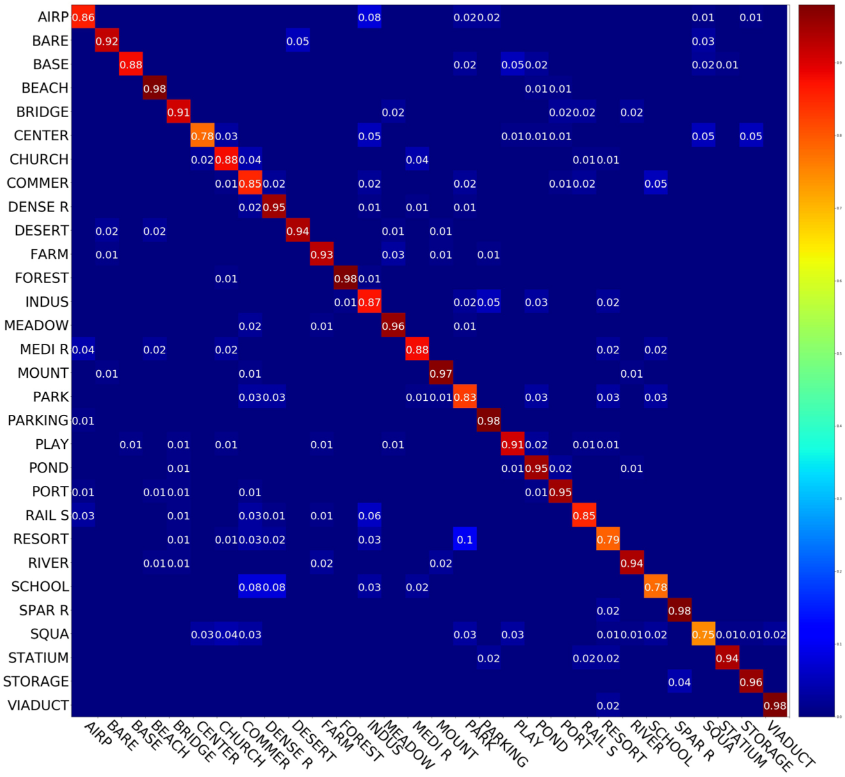

Figure 5. Effects of different parameter settings of scales and patch size on the classification results of the AID dataset.

Table 1. Performance comparison of the state-of-the-art methods on the AID data set.

Style name	Accuracy
DCA with concatenation [8]	89.71 ± 0.33
Fusion by addition [8]	91.87 ± 0.36
salM3LBP–CLM [9]	89.76 ± 0.45
CaffeNet [7]	89.53 ± 0.31
VGG-VD16 [7]	89.64 ± 0.36
GoogLeNet [7]	86.39 ± 0.55
ResNet softmax [10]	90.62 ± 0.56
ResNet fine-turn [10]	94.23 ± 0.34
Proposed method	91.07 ± 0.33

4. Conclusion

This paper presented a novel multi-feature fusion method for HRRS image scene classification. From the classification result we can conclude that: 1) the proposed method fully considers the hierarchical information hidden in the pre-trained CNN and the global and local information extracted from the patch based MS-CLBP method. The weighted concatenated features are more discriminating for classification. 2) The multi-SDAE classification method utilizes the deep feature learning abilities of each SDAE, and the experimental results indicate the effectiveness of our proposed methods. At present, there are there are still many technologies to be improved, such as a new feature coding method to encode these combined features into a more compact representation, or adopt more effective complementary low-level features.

Conflicts of Interest

No conflicts of interest regarding the publication of this paper.

References

[1] Ojala, T., Pietikainen, M. and Maenpaa, T. (2002) Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 971-987. http://dx.doi.org/10.1109/TPAMI.2002.1017623

[2] Perronnin, F., Sánchez, J. and Mensink, T. (2010) Improving the Fisher Kernel for Large-Scale Image Classification. *European Conference on Computer Vision*, Crete, 5-11 September 2010, 143-156. http://dx.doi.org/10.1007/978-3-642-15561-1_11

[3] Han, X.H., Chen, Y.W. and Xu, G. (2015) High-Order Statistics of Weber Local Descriptors for Image Representation. *IEEE Transactions on Cybernetics*, **45**, 1180-1193. http://dx.doi.org/10.1109/TCYB.2014.2346793

[4] Guo, Z., Zhang, L. and Zhang, D. (2010) A Completed Modeling of Local Binary Pattern Operator for Texture Classification. *IEEE Transactions on Image Processing*, **19**, 1657-1663. http://dx.doi.org/10.1109/TIP.2010.2044957

[5] Huang, L., Chen, C., Li, W. and Du, Q. (2016) Remote Sensing Image Scene Classi-

fication Using Multi-Scale Completed Local Binary Patterns and Fisher Vectors. *Remote Sensing*, **8**, 483. http://dx.doi.org/10.3390/rs8060483

[6] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. and Manzagol, P.A. (2010) Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, **11**, 3371-3408.

[7] Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y. and Lu, X. (2017) AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, **55**, 3965-3981. http://dx.doi.org/10.1109/TGRS.2017.2685945

[8] Chaib, S., Liu, H., Gu, Y. and Yao, H. (2017) Deep Feature Fusion for VHR Remote Sensing Scene Classification. *IEEE Transactions on Geoscience & Remote Sensing*, **55**, 4775-4784. http://dx.doi.org/10.1109/TGRS.2017.2700322

[9] Bian, X., Chen, C., Tian, L. and Du, Q. (2017) Fusing Local and Global Features for High-Resolution Scene Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **10**, 2889-2901. http://dx.doi.org/10.1109/JSTARS.2017.2683799

[10] Pilipović, R. and Risojević, V. (2017) Evaluation of Convnets for Large-Scale Scene Classification from High-Resolution Remote Sensing Images. *IEEE EUROCON* 2017-17*th International Conference*, Ohrid, 6-8 July 2017, 932-937. http://dx.doi.org/10.1109/EUROCON.2017.8011248