Bidirectional Adaptive Feature Fusion for Remote Sensing Scene Classification

Weijun Ji[1,2], Xuelong Li[1], and Xiaoqiang Lu[1](✉)

[1] Center for OPTical IMagery Analysis and Learning (OPTIMAL),
State Key Laboratory of Transient Optics and Photonics,
Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences,
Xi'an 710119, Shaanxi, People's Republic of China
jiweijun2015@opt.cn, {xuelong_li,luxiaoqiang}@opt.ac.cn

[2] University of Chinese Academy of Sciences, 19A Yuquanlu, Beijing 100049,
People's Republic of China

Abstract. Convolutional neural networks (CNN) have been excellent for scene classification in nature scene. However, directly using the pretrained deep models on the aerial image is not proper, because of the spatial scale variability and rotation variability of the HSR remote sensing images. In this paper, a bidirectional adaptive feature fusion strategy is investigated to deal with the remote sensing scene classification. The deep learning feature and the SIFT feature are fused together to get a discriminative image presentation. The fused feature can not only describe the scenes effectively by employing deep learning feature but also overcome the scale and rotation variability with the usage of the SIFT feature. By fusing both SIFT feature and global CNN feature, our method achieves state-of-the-art scene classification performance on the UCM and the AID datasets.

Keywords: Feature fusion · Remote sensing scene classification

1 Introduction

High Spatial Resolution (HSR) remote sensing image scene classification aims to automatically label an aerial image with the specific category. It is the basis of land-use object detection and image understanding which are widely used in the field of military and civilian. Remote sensing image scene classification is a fundamental problem which has attracted much attention. The most vital and challenging task of the scene classification is to develop an effective holistic representation of the aerial image to directly model an image scene.

To develop the effective holistic representation, many methods have been proposed in recent years. Primitively, a bottom-up scheme was proposed to model

W. Ji—His research interests include remote sensing scene classification and computer vision.

J. Yang et al. (Eds.): CCCV 2017, Part II, CCIS 772, pp. 486–497, 2017.
https://doi.org/10.1007/978-981-10-7302-1_40

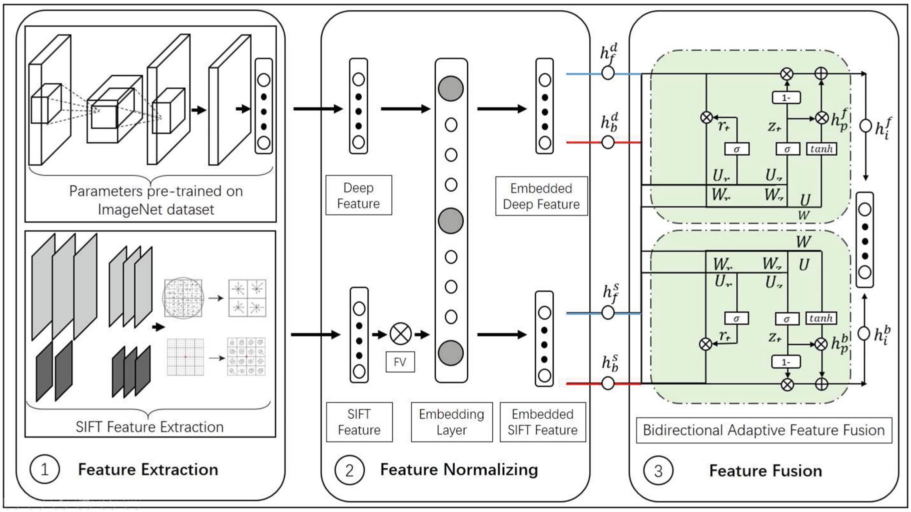

Fig. 1. The flowchart of the proposed method. Part 1 shows as the process of extracting features. Part 2 shows as the process of feature normalizing. Part 3 shows as the process of the bidirectional feature fusion.

a HSR remote sensing image scene by three “pixel-region-scene” steps. To further express the scene, many researchers tried to represent the scene directly without classifying pixels and regions. For example, Bag-of-the-Visual-Words (BoVW) [20] and some extensions of BoVW were proposed to improve the classification accuracy. Besides, the family of latent generative topic models [1] have also been applied in HSR image scene classification.

All these methods have better performance, they cannot satisfy the demand of higher accuracy. One main reason is that these methods may lack the flexibility and adaptivity to different scenes [19]. Recently, it has been proven that deep learning methods can adaptively learn image features which are suitable for specific scene classification tasks and achieve far better classification performances. However, there are two major issues that seriously influence the use of the deep learning method. (1) The deep learning method needs large training data to train the model and is time consuming. (2) The pre-trained models take little consideration of the HSR remote sensing image characteristics [13]. The spatial scale variability and rotation variability of the HSR remote sensing images cannot be expressed precisely by pre-trained models.

In order to relieve the above issues, a feature fusion strategy is proposed in the paper. Many works have proved that SIFT feature has better performance on overcoming the scale variability and rotation variability [16,17]. The proposed method can be divided into three steps. First, the deep features and the SIFT features are extracted by CNN and SIFT filter respectively. Second, the two features are normalized by normalization layer to get the same dimension. Finally, the normalized features are adaptively assigned with optimal weights to get the fused feature. Specially, the fusion strategy is to assign average weights to confidence scores from the deep feature and the SIFT feature. The optimal weight of a feature is trained by the model and is optimized by Back Propagation Through Time (BPTT). With the help of above adaptive feature fusion strategy, the proposed method can get a discriminative representation for scene classification. The flowchart of the proposed method is shown as Fig. 1.

In general, the major contributions of this paper are as follows:

1. For HSR image scene classification, deep learning features are sensitive to scale and rotation variant. To address the aforementioned problem, deep learning features and SIFT features are together exploited remote sensing scene classification.
2. A new fusion architecture is investigated to take full advantage of the information of features.

The rest of the paper is organized as follows: In Sect. 2, the related works on HSR scene classification are reviewed. Section 3 gives the detailed description of our proposed method. The experiments of our method on two data sets are shown in Sect. 4. In the last Section, we conclude this paper.

2 Related Work

In this section we provide previous work on remote sensing scene classification and feature fusion.

Recently, several methods have been proposed for remote sensing scene classification. Many of these methods are based on deep learning [3,10,13–15,26]. The deep learning method becomes the mainstream approach, because it uses a multi-stage global feature learning strategy to adaptively learn image features and cast the aerial scene classification as an end-to-end problem. To further improve the classification accuracy, many researchers have proposed different methods to overcome the issues which pre-trained models taking little consideration of the HSR remote sensing image characteristics. Some of them have addressed the problem from datasets, they have enlarged the datasets to increase the scale and rotation invariant information. Y. Liu *et al.* [11] added random structure noise i.e. random-scale stretching to capture the essential feature robust to scale change. G. Cheng *et al.* [4,5] rotated the original data to enlarge the dataset and increased the rotation information. Others focused on the deep learning architecture. G. Cheng *et al.* [4] added two fully connected layers and changed the loss function to ensure all the rotation data has the same representation. Moreover, some researchers cared about the classifier. F. Zhang *et al.* [23] adapted the boosting method to merge some weak networks to get more accurate results.

There has been a long line of previous work incorporating the idea of fusion for scene classification. For example, [18] designed sparse coding based multiple feature fusion (SCMF) for HSR image scene classification. SCMF sets the fused result as the connection of the probability images obtained by the sparse codes of SIFT, the local ternary pattern histogram Fourier, and the color histogram features. [24] also used four features and concatenated the quantized vector by k-means clustering for each feature to form a vocabulary. All these fusion methods proposed for HSR image scene classification obtained good results to some degree. Nevertheless, all of these methods were limited because the fused features were not effective enough to represent the scene. In this paper, the strategy fuses deep learning features and SIFT features together to achieve comparable performance.

3 Proposed Method

The proposed method involves three parts: (1) feature extraction. (2) feature normalizing. (3) feature fusion. The proposed method fuses the two features to get the final classification result.

3.1 Feature Extraction

This Section of feature extraction is divided into two parts, the first part is the deep feature extraction and the second part is the SIFT feature extraction. Deep features are extremely effective features for scene classification, and they are the guarantee of classification accuracy. Considering the limitation of the training data in the satellite image and the high computation complexity of tuning convolutional neural networks, we use the pre-trained convolutional neural networks. [19] compares three representative high-level deep learned scene classification methods. The comparison shows that VGG-VD-16 gets the best result.

VGG gives a thorough evaluation of networks by increasing depth using an architecture with very small (33) convolution filters, which shows a significant improvement on the accuracies, and can be generalised well to wide range of tasks and datasets. In our work, we use one of its best-performance models named VGG-VD-16, because of its simpler architecture and slightly better results. It is composed of 13 convolutional layers and followed by 3 fully connected layers, thus results in 16 layers. The parameters of the model are pre-trained on approximately 1.2 million RGB images in the ImageNet ILSVRC 2012. We carry on the experiment to verify that the first fully connected layer perform best in the experiment. So the first fully connected layer is chosn as the feature vectors of the images [19].

The SIFT feature [12] describes a patch by the histograms of gradients computed over a (4×4) spatial grid. The gradients are then quantized into eight bins, so the dimension of the final feature vector is 128 $(4 \times 4 \times 8)$.

The SIFT feature searches over all scales and image locations. It is implemented effectively by using a difference-of Gaussian function to identify potential interest points that are invariant to scale and orientation. The initial image is incrementally convolved with Gaussians to produce images separated by a constant factor k in scale space. Adjacent images scales are subtracted to produce the difference-of Gaussian images. This step detect the scale invariance feature from different scale space. And the date will experience key point localization, orientation assignment and key point descriptor. The features with scale invariant and rotation invariant are extracted.

3.2 Feature Normalizing

After feature extraction, the dimension of the deep feature is 4096 and the dimension of the SIFT feature is $128 \times N$ (where N indicates the number of the points of the interest). The dimensions of the two features are different so we need to utilize the normalizing method to reshape the two features in the same dimension.

We firstly use the FV [16] to encode the SIFT feature. In essence, the SIFT features generated from FV encoding method is a gradient vector of the log-likelihood. By computing and concatenating the partial derivatives, the mean and variance of Gaussian functions, the dimension of the final feature vector is $2 \times K \times F$ (where F indicates the dimension of the local feature descriptors and K denotes the size of the dictionary). And then, we resize the size of the dictionary to control the dimension of the output feature. Finally, we use the encoding method to express SIFT feature more effectively.

Then the deep feature and the SIFT feature are together to get the same dimension by a normalizing layer. The normalizing layer is composed of two fully connected layers. The dimensions of the fully connected layers and the number of the connection layers are verified by the experiments. The number of the fully connection layer is set to 2 and the dimensions of the fully connection layers are 4096 dimensions and 2048 dimensions. The parameters of the fully connection layer are trained together with the below proposed fusion strategy.

3.3 Feature Fusion

We propose a novel bidirectional adaptive feature fusion method inspired by recurrent neural networks [7,8]. The process of the feature fusion can be divided into three procedures. The input features pass through the primary fusion node to get a primary fusion feature. Then the primary feature and the first input node feature are fused with different weights to get the intermediate feature. After getting the two intermediate features from two directions, the two intermediate features are summed together with different weights and with the bias to get the final fusion feature. The reset ratio and the update ratio should be calculated firstly before the feature fusion. The details of the feature fusion is described as below.

In forward direction, the framework of the feature fusion is shown as Fig. 2. There are two input nodes in the framework, the first input feature is the deep learning feature marked as h_f^d and the second input feature is the SIFT feature marked as h_f^s. The reset ratio and the update ratio are calculated respectively by Eqs. 1 and 2. The primary fusion feature marked as h_f^p is calculated by Eq. 3. The primary fusion feature and the deep learning feature are as the two input features to calculate the intermediate fusion feature (marked as h_f^i) Eq. 4.

$$z_f = sigmoid(W_z h_f^s + U_z h_f^d) \tag{1}$$

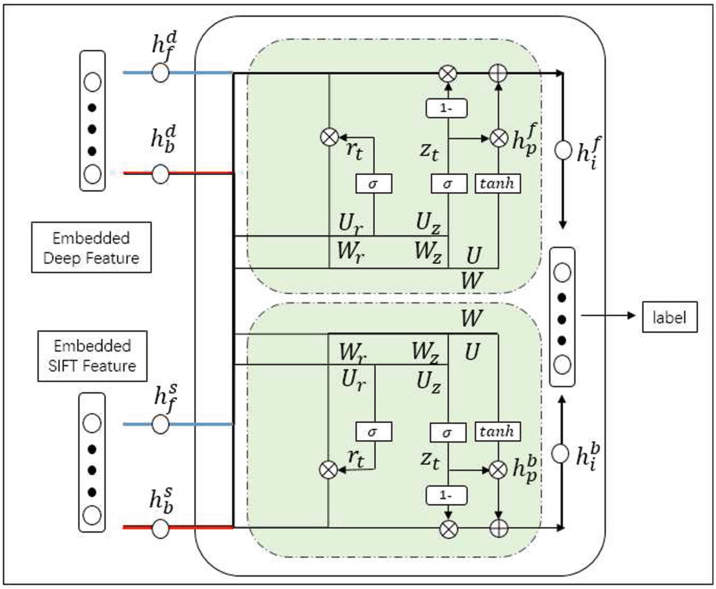

Fig. 2. The details of the feature fusion strategy, the input features are the embedded deep feature and embedded SIFT feature. The output is the label of the image. The two nodes are described as the bidirectional feature fusion strategy.

$$r_f = sigmoid(W_r h_f^s + U_r h_f^d) \tag{2}$$

$$h_f^p = tanh(W h_f^s + r_f \circ U h_f^d) \tag{3}$$

$$h_f^i = z_f * h_f^d + (1 - z_f) * h_f^p \tag{4}$$

In backward direction, the first input feature is the SIFT feature marked as h_b^s and the second input feature is the deep learning feature marked as h_b^d. The reset ratio and the update ratio are also calculated respectively by Eqs. 5 and 6. The primary fusion feature marked as h_b^p is calculated by Eq. 7. The primary fusion feature and the SIFT feature are as the two input features to calculate the intermediate fusion feature (marked as h_b^i) by Eq. 8.

$$z_b = sigmoid(W_z h_b^d + U_z h_b^s) \tag{5}$$

$$r_b = sigmoid(W_r h_b^d + U_r h_b^s) \tag{6}$$

$$h_b^p = tanh(W h_b^d + r_b \circ U h_b^s) \tag{7}$$

$$h_b^i = z_b * h_b^s + (1 - z_b) * h_b^p \tag{8}$$

In bidirectional feature fusion, h_b^i and h_f^i are calculated together to get the final fusion feature y_t. The formula is show as (9).

$$y = W_f h_f^i + W_b h_b^i + b_y \tag{9}$$

In all these formula Eqs. 1 to 9. $W_z, W_r, U_z, U_r, W, U, W_f, W_b, b_y$ are all the weight parameters needed to be learned during the training the model. After getting the final fusion feature, the fusion feature passes through the softmax layer to classify the HSR images into different categories. The model is trained by BPTT method. The proposed fusion method takes full advantage of the deep learning feature and the SIFT feature to achieve comparable performance in the HSR image scene classification.

4 Experimental Results and Analysis

To evaluate the effectiveness of the proposed method for scene classification, we perform experiments on two datasets: the UC Merced Land Use dataset [20] and the AID dataset [19]. At the same time, in order to evaluate the fusion strategy, we perform experiments on both single features (deep feature and SIFT feature) and the fused feature. The details of the experiments and the results are described in the following sections.

4.1 Dataset and Experiment Set Up

UC Merced Land Use dataset: It contains 21 scene categories. Each category contains 100 images. Each scene image consists of 256 × 256 pixels, with a spatial resolution of one foot per pixel. The example images are shown as Fig. 3. In the experiment, we randomly choose 80 images from each category as the training set, the rest images are chosen as the testing set. Some of the scenes are similar to each other causing the scene classification to be challenging.

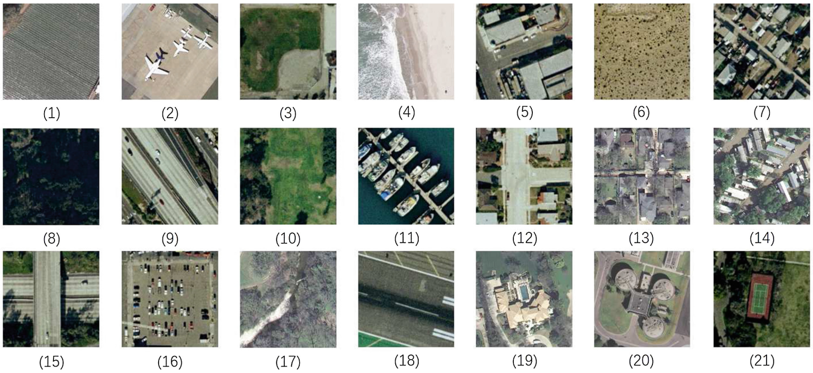

Fig. 3. Example images associated with 21 land-use categories in the UC-Merced data set: (1) agricultural, (2) airplane, (3) baseball diamond, (4) beach, (5) buildings, (6) chaparral, (7) dense residential, (8) forest, (9) freeway, (10) golf course, (11) harbor, (12) intersection, (13) medium residential, (14) mobile home park, (15) overpass, (16) parking lot, (17) river, (18) runway, (19) sparse residential, (20) storage tanks and (21) tennis court

AID: It is a new dataset containing 10000 images which make it the biggest dataset of high spatial resolution remote sensing images. Thirty scene categories are included in this dataset. The number of different scenes is different from each other changing from 220 to 420. The shape of the scene image consists of 600 × 600 pixels. We follow the parameters setting in the [19]. Fifty images of each category are chosen as the training set and the rest of the images are set as the testing set. The dataset is famous for its images with high intra-class diversity and low inter-class dissimilarity. The example images are shown as Fig. 4.

Table 1. Comparison with the previous reported accuracies with the UC merced data set

Method	OA(%)	Method	OA(%)
BOVW	72.05	S-UFL	82.30
pLSA	80.71	SAL-PTM	88.33
SPCK++	76.05	GBRCN	94.53
LDA	81.92	SRSCNN	95.10
SPM+SIFT	82.30	CNN	93.10
SIFT+FC	81.67	PROPOSED	95.48

4.2 Experiment on UC Merced Dataset

To evaluate the proposed method, we compare our method with the state-of-the-art methods in scene classification. The comparative scene classification methods include BOVW [20], pLSA [2], LDA [1], SPM [9], SPCK [21], SIFT+SC [6], S-UFL [22], GBRCN [23], SAL-PTM [25] and SRSCNN [11]. The results are showed in Table 1. From Table 1, it can be concluded that the proposed method performs better than other methods. Compared with GBRCN which combines many CNN for scene classification, our method is 0.91% better than it. Our method is 0.38% better than SRSCNN which randomly selects patches from image and stretch to the specific scale as input to train CNN. The experiment results indicate the effectiveness of the proposed method.

4.3 Experiment on AID Dataset

To evaluate the proposed method on AID, the compared methods are chosen to code SIFT feature with models such as BOVW, pLSA, LDA, SPM, VLAD, FV and LLC. The deep models are chosen as VGG-VD-16 and the proposed method. Table 2 gives the comparison, which also shows the effectiveness of the proposed method.

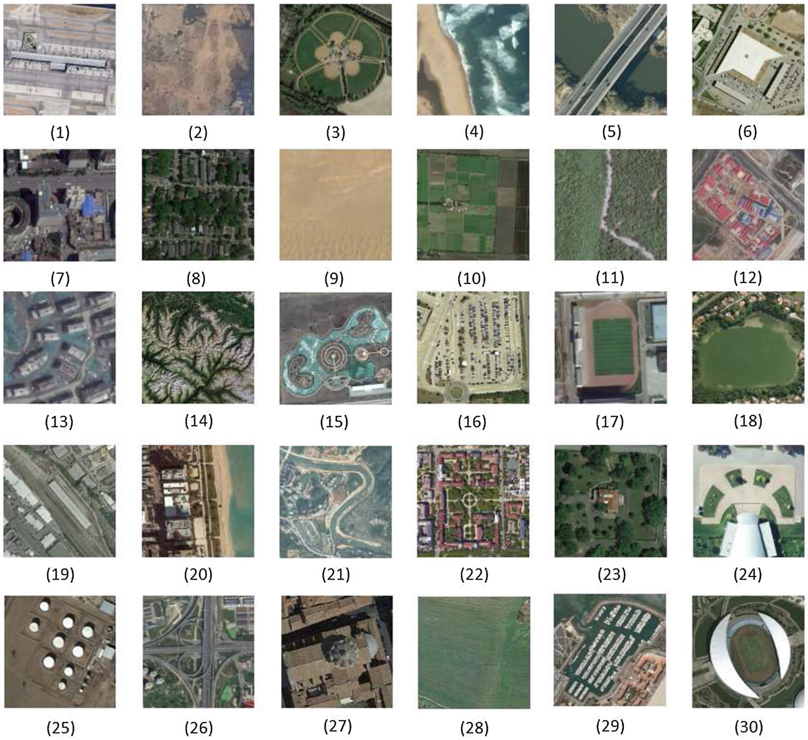

Fig. 4. Example images associated with 30 land-use categories in the UC-Merced data set: (1) airport, (2) bare land, (3) baseball field, (4) beach, (5) bridge, (6) center, (7) church, (8) commercial, (9) dense residential, (10) desert, (11) farmland, (12) forest, (13) industrial, (14) meadow, (15) medium residential, (16) mountain, (17) park, (18) parking, (19) play ground, (20) pond, (21) port, (22) railway station, (23) resort, (24) river, (25) school, (26) sparse residential, (27) square, (28) stadium, (28) storage tanks, (29) viaduct and (30) viaduct

Table 2. Comparison with the feature fusion with the aid data set

Method	BOVW	FV	LLC	pLSA
OA(%)	68.37	78.99	63.24	63.07
Method	SPM	VGG-VD-16	VLAD	PROPOSED
OA(%)	45.52	89.64	68.96	93.56

4.4 Experiment on Feature Fusion

To evaluate the proposed fusion strategy, we compare the classification accuracy in experiments with single deep feature and experiments with single SIFT feature encoded with fisher vector. Moreover, we compare the fusion strategy with sum of the features and the average of the features. The result shows that the single SIFT feature encoded with fisher vector performs worst in the experiment. At the same time, the deep feature and the fusion feature all get the classification accuracy exceeding 90%. The proposed fusion strategy obtains the best performance. The sum of feature gets the accuracy of 94.01% and the average of the features gets the accuracy of 93.05%. The classification accuracy of the proposed fusion strategy is shown as the best result compared with other fusion strategys. The final result is shown in Table 3. Our experiment is conducted on the UC Merced data set.

Table 3. Comparison with the feature fusion with the UC merced data set

Method	FV	VGG-VD-16	SUM	AVG	Proposed
OA(%)	85.11	93.09	94.01	93.05	95.48

5 Conclusion

The proposed method aims to solve the scale variance in high spatial resolution remote sensing images scene classification. Fusion of deep feature and SIFT feature is evaluated to get the state-of-the-art performance. The experiments on UCM and AID demonstrate the effectiveness of the proposed method.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006). https://doi.org/10.1007/11744085_40
3. Castelluccio, M., Poggi, G., Sansone, C., Verdoliva, L.: Land use classification in remote sensing images by convolutional neural networks. J. Mol. Struct. Theochem **537**(1), 163–172 (2015)
4. Cheng, G., Ma, C., Zhou, P., Yao, X., Han, J.: Scene classification of high resolution remote sensing images using convolutional neural networks. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, pp. 767–770 (2016)
5. Cheng, G., Zhou, P., Han, J.: Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Trans. Geosci. Remote Sens. **54**(12), 7405–7415 (2016)

6. Cheriyadat, A.M.: Unsupervised feature learning for aerial scene classification. IEEE Trans. Geosci. Remote Sens. **52**(1), 439–451 (2014)
7. Chung, J., Gulcehre, C., Cho, K.H., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. Eprint Arxiv (2014)
8. Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649 (2013)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178 (2006)
10. Li, X., Lu, Q., Dong, Y., Tao, D.: SCE: a manifold regularized set-covering method for data partitioning. IEEE Trans. Neural Netw. Learn. Syst. (2017)
11. Liu, Y., Zhong, Y., Fei, F., Zhang, L.: Scene semantic classification based on random-scale stretched convolutional neural network for high-spatial resolution remote sensing imagery. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, pp. 763–766 (2016)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
13. Lu, X., Zheng, X., Yuan, Y.: Remote sensing scene classification by unsupervised representation learning. IEEE Trans. Geosci. Remote Sens. **55**, 5185–5197 (2017)
14. Luus, F.P.S., Salmon, B.P., Bergh, F.V.D., Maharaj, B.T.J.: Multiview deep learning for land-use classification. IEEE Geosci. Remote Sens. Lett. **12**(12), 1–5 (2015)
15. Nogueira, K., Penatti, O.A.B., Santos, J.A.D.: Towards better exploiting convolutional neural networks for remote sensing scene classification. Patt. Recogn. **61**, 539–556 (2017)
16. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
17. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_11
18. Sheng, G., Yang, W., Xu, T., Sun, H.: High-resolution satellite scene classification using a sparse coding based multiple feature combination. Int. J. Remote Sens. **33**(8), 2395–2412 (2012)
19. Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X.: AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens. **55**, 3965–3981 (2017)
20. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of Sigspatial International Conference on Advances in Geographic Information Systems, pp. 270–279 (2010)
21. Yang, Y., Newsam, S.: Spatial pyramid co-occurrence for image classification. In: Proceedings of International Conference on Computer Vision, pp. 1465–1472 (2011)
22. Zhang, F., Du, B., Zhang, L.: Saliency-guided unsupervised feature learning for scene classification. IEEE Trans. Geosci. Remote Sens. **53**(4), 2175–2184 (2015)
23. Zhang, F., Du, B., Zhang, L.: Scene classification via a gradient boosting random convolutional network framework. IEEE Trans. Geosci. Remote Sens. **54**(3), 1793–1802 (2016)
24. Zheng, X., Sun, X., Fu, K., Wang, H.: Automatic annotation of satellite images via multifeature joint sparse coding with spatial relation constraint. IEEE Geosci. Remote Sens. Lett. **10**(4), 652–656 (2012)

25. Zhong, Y., Zhu, Q., Zhang, L.: Scene classification based on multifeature probabilistic latent semantic analysis for high spatial resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. **53**(11), 6207–6222 (2015)
26. Zou, Q., Ni, L., Zhang, T., Wang, Q.: Deep learning based feature selection for remote sensing scene classification. IEEE Geosci. Remote Sens. Lett. **12**(11), 2321–2325 (2015)