The Journal of Engineering

The 2nd 2018 Asian Conference on Artificial Intelligence Technology (ACAIT 2018)

IET Journals
The Institution of Engineering and Technology

Scene classification of remote sensing images based on hierarchical sparse coding

eISSN 2051-3305
Received on 17th July 2018
Accepted on 26th July 2018
E-First on 25th September 2018
doi: 10.1049/joe.2018.8268
www.ietdl.org

Xu Jiaqing[1] ✉, *Lv Qi*[2], *Liu Hongjun*[3], *He Jie*[1]

[1]*School of Computer, National University of Defense Technology, Changsha, People's Republic of China*
[2]*Unit 31104 of PLA, Nanjing, People's Republic of China*
[3]*College of Biomedical Engineering & Imaging Medicine, Army Medical University, Chongqing, People's Republic of China*
✉ *E-mail: xujiaqing@nudt.edu.cn*

Abstract: Remote sensing image scene classification is an important method for remote sensing image analysis and interpretation and plays an important role in civil and military fields. In this study, a scene classification method of remote sensing images based on hierarchical sparse coding is proposed. This method is essentially a kind of multi-layer, multi-scale, and multi-path sparse coding. It can extract features of optical remote sensing images more effectively, so that the features of the remote sensing images can be represented more sufficiently. The obtained codes are further used for spatial pyramid pooling (SPP) operation, and the corresponding SPP representation is obtained. SPP representations in different paths are combined and outputted to the support vector machine classifier, and the final classification results are obtained. Experiments on two data sets show that the proposed method can obtain better scene classification accuracy.

1 Introduction

With the development of remote sensing technology, the spatial resolution of spaceborne high-resolution remote sensing images (such as WorldView-3 satellite remote sensing images) can reach 0.31 m, and the resolution of airborne remote sensing images can be even higher. Remote sensing image scene classification is an important means for remote sensing image interpretation and analysis and plays an important role in monitoring of land use, urban planning, environmental monitoring, agriculture, forestry, and military reconnaissance. High-resolution remote sensing image scenes are often complex in content and have high background noise. It is a challenging problem in the field of remote sensing to efficiently represent and recognise features. It is also a hot research direction in remote sensing technology.

Scene classification is an important content of image comprehension, focusing on the overall perception and analysis of image scenes. In order to effectively model the remote sensing image scene to improve the classification accuracy, many related work with in-depth studies has been conducted in recent years. Generally, the description of the local structure and spatial attributes of scenes is the key to the accurate classification of high-resolution remote sensing image scenes. The direct modelling of scenes using low-level features is a common method. A typical example is the bag-of-visual-words (BOVW) model. BOVW originates from the bag-of-words (BOW) model in the field of text analysis. This model has been widely used in the field of computer vision. In recent years, it has gradually been introduced into image scene classification. The BOVW model can be roughly divided into two steps: feature learning and feature coding. In the feature learning phase, the low-level image features are clustered and the cluster centres form visual words. In the subsequent feature encoding phase, the image is mapped to the nearest visual word, and the new feature is constructed by calculating the histogram of the visual word. The SPMK (spatial pyramid match kernel) method divides a picture into multiple sub-regions, forms a word bag in each region, and then uses spatial pyramid matching to converge to generate a better scene representation. Considering the co-occurrence relationship of visual words, SPMK was transformed into SPCK (spatial pyramid co-occurrence kernel). SPCK+ and SPCK++ are obtained by combining SPCK with BOVW and SPMK, respectively, and the classification accuracy is improved [1]. Chen *et al.* [2] proposed a pyramid-of-spatial-relations (PSR) method to further mine the spatial relationship of local features and introduce them into the BOVW method.

BOVW generally uses only the low-level visual features of the image (mainly colours, textures, shapes etc., such as colour histogram, LBP, and SIFT), and there is a gap between high-level semantics. To overcome this deficiency, the probabilistic topic model (PTM) used in natural language processing is introduced into remote sensing image scene classification, such as latent Dirichlet allocation (LDA) and probability implicit Probabilistic latent semantic analysis (PLSA). LDA is a hierarchical probabilistic model that describes the distribution of visual words through unsupervised learning. Lienou *et al.* [3] first introduced LDA into remote sensing scene classification, while Cheng *et al.* [4] combined PLSA with BOVW to improve scene classification accuracy. Zhang *et al.* [5] combined LDA with support vector machine (SVM) and considered image saliency information to improve scene classification accuracy [6–9].

Sparse coding-based methods are also widely used in remote sensing scene classification tasks. Cheriyadat [10] proposed an airborne remote sensing image scene classification method based on unsupervised sparse coding. Zheng *et al.* [11] used spatial correlation constraints to propose a multifeature joint sparse coding method. Cheng *et al.* [12] proposed the concept of sparselet, and used an autoencoder with sparse constraints to extract the middle layer features. Qi *et al.* [13] proposed a sparse coding-based correlation model to discover the co-occurrence relationship of visual words to improve the accuracy of high-resolution remote sensing imagery scene classification.

Characterising the local structure and spatial attributes of scenes is the key to accurately classifying high-resolution remote sensing image scenes. Most of the existing BOVW models use only the low-level visual features of the image, and there is a gap between high-level semantics. Probabilistic topic models and existing sparse coding related methods have played a role in overcoming this deficiency. How to perform more effective feature representation on remote sensing image scenes to improve the classification accuracy is still a challenging problem in the current remote sensing field.

This paper proposes a remote sensing image scene classification framework based on hierarchical sparse coding. The proposed method utilises multi-layer multi-scale multi-path sparse coding to extract the features of remote sensing images effectively, so that the features of the images can be fully expressed. Experiments on

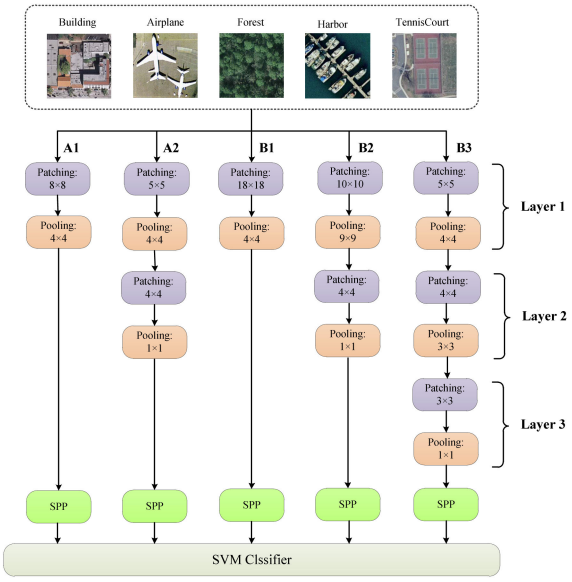

Fig. 1 *Schematic diagram of the classification framework of M³SC*

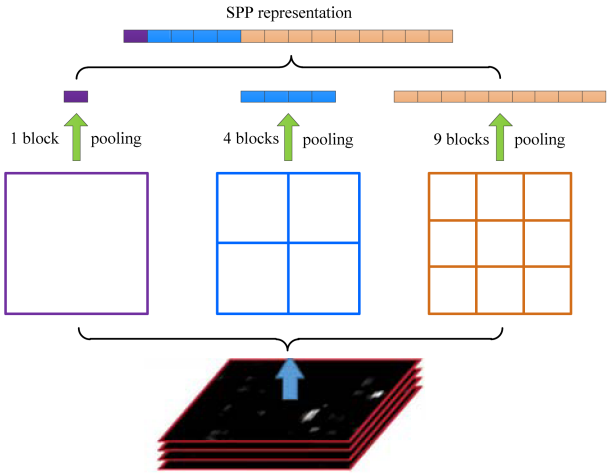

Fig. 2 *Schematic diagram of SPP*

two high-resolution optical remote sensing image datasets show that the proposed method can obtain better scene classification accuracy.

This rest of this paper is organised as follows. Section 2 describes the proposed method for optical remote sensing image scene classification based on hierarchical sparse coding, and the experimental results are given in Section 3. Section 4 summarises this paper.

2 Multi-layer multi-scale multi-path sparse coding

To better perform feature extraction, we propose a remote sensing scene classification framework based on hierarchical sparse coding. The process is shown in Fig. 1. This method is called multi-layer multi-scale multi-path sparse coding and is abbreviated as M³SC. The proposed framework is inspired by the hierarchical matching pursuit model proposed by Bo *et al.* [14] and combines information from multiple sparse coding paths and applies it to remote sensing image scene classification. The sparse encoding part of the framework mainly includes two operations: patching and pooling. Each pooling operation follows the patching operation. The code obtained by the last pooling is further used as the input of Spatial Pyramid Pooling (SPP) operation to obtain the corresponding SPP representation. SPPs of different paths are sent together to the SVM classifier to get the final classification result. The number of layers is used to represent the number of 'patching + pooling' structures on a path. It can be seen from Fig. 1 that the number of layers of paths A1 and B1 is 1, the number of layers of paths A2 and B2 is 2, and the number of layers of path B3 is 3.

2.1 Patching operation

Specifically, the patching operation includes three sub-processes: dictionary samples generation, dictionary learning, and sparse coefficients solving.

2.1.1 Dictionary samples generation: The image is first preprocessed, including normalisation and subtraction of the mean. Subsequently, a sliding window operation is performed on the entire image using a square window with a step size of one to obtain patches. Then the patches obtained by the sliding window are randomly selected as the samples to the train the dictionary. The number of the selected patches is equal to the pre-given dictionary size.

2.1.2 Dictionary learning: The dictionary learning is performed after sliding window and randomly selecting samples. The K-SVD algorithm [15] is used as the learning algorithm in this classification framework. The K-SVD algorithm is actually a combination of K-means clustering and singular value decomposition. The goal is to solve the optimisation problem as follows:

$$\hat{D} = \arg\min_{D} \| Y - DX \|_F^2 = \arg\min_{D} \| Y - \sum_{j=1}^{m} d_j x_j^T \|_F^2$$

$$= \arg\min_{D} \| Y - \sum_{j \neq l} d j x_j^T - d l x_l^T \|_F^2 = \arg\min_{D} \| El - d l x_l^T \|_F^2$$

where $Y \in \mathbb{R}^{n \times N}$ is the data of samples (n denotes the dimension of data and N denotes the number of samples), $D \in \mathbb{R}^{n \times m}$ is the dictionary (m denotes the size of codewords), and $X \in \mathbb{R}^{m \times N}$ is the sparse coefficients matrix.

First, the overall representation residual $E_l = Y - \sum_{j \neq l} d_j x_j^T$ is computed, and then d_l and x_l are updated. In order to maintain the sparsity of x_l^T in this step, only the non-zero elements of x_l^T should be preserved and only the non-zero items of E_l should be reserved, i.e. E_l^P, from $d_l x_l^T$. Then, SVD decomposes E_l^P into $E_l^P = U \Delta V$ and then updates dictionary d_l.

2.1.3 Sparse coefficients solving: The sparse coefficients are solved using the orthogonal matching pursuit (OMP) algorithm [16]. OMP selects the codeword best correlated with the current residual at each iteration, which is the reconstruction error remaining after the codewords chosen thus far are subtracted. At the first iteration, this residual is exactly the observation. Once a new codeword is selected, the observation is orthogonally projected onto the span of all the previously selected codewords and the residual is recomputed. The procedure is repeated until the desired sparsity level is reached.

2.2 Pooling operation

Each pooling operation following after a patching operation is a type of max pooling. The maximum value of the sparse coefficients of the patching operation is used as the result of the pooling. If the number of layers is greater than one, the result of the pooling operation is passed as input to the subsequent patching operation.

After the last pooling operation, the encoded feature map is segmented by 1×1, 2×2, 3×3 to obtain 1, 4, and 9 blocks, respectively. The max pooling operation is performed on each block to obtain the spatial pyramidal pooling (SPP) representation, as shown in Fig. 2.

In general, the M³SC classification framework proposed in this paper has several advantages: (i) Some paths has multi-layered structure, and the deeper and more abstract features of the image can be extracted through the hierarchical feature representation. (ii) Feature extraction can be achieved at different scales. This different scale is reflected in different patching scales, different pooling sizes, and the SPP operation; (iii) The number of layers on different paths, the patching size, and the size of the pooling may

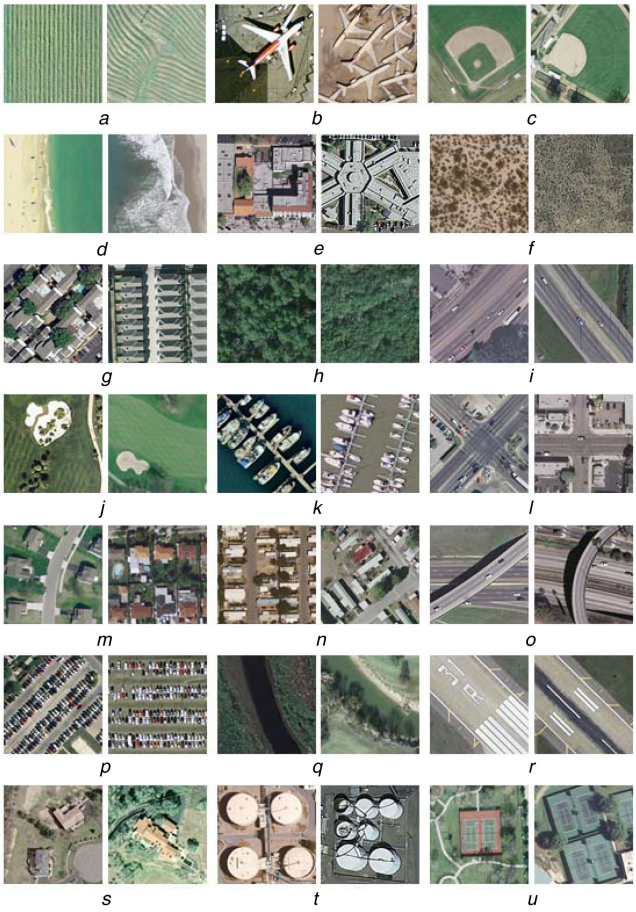

Fig. 3 *Sample images of the UCM dataset*

(a) Agricultural, ***(b)*** Aeroplane, ***(c)*** Baseball-diamond, ***(d)*** Beach, ***(e)*** Buildings, ***(f)*** Chaparral, ***(g)*** Dense-residential,***(h)*** Forest, ***(i)*** Freeway, ***(j)*** Golf-course, ***(k)*** Harbour, ***(l)*** Intersection, ***(m)*** Medium-residential, ***(n)*** Mobile-homepark, ***(o)*** Overpass, ***(p)*** Parking lot, ***(q)*** River, ***(r)*** Runway, ***(s)*** Sparse-residential, ***(t)*** Storage-tanks, ***(u)*** Tennis-court

be different, and each path may serve to compensate for each other. Due to these advantages, M^{3}SC hopes to achieve higher performance in remote sensing image scene classification.

3 Experimental results

To verify the performance of M^{3}SC for the task of remote sensing image scene classification, we conducted experiments on the UCM and WHU-RS data sets.

3.1 Experiments on UCM dataset

The UCM (University of California Merced) dataset has 21 land-use categories, each containing 100 airborne remote sensing images. The size of the image is 256 × 256 and the resolution is 1 foot (0.3 m). In Fig. 3, two sample images are given for each category. For each category, 80 pictures were randomly selected as training samples, and the remaining 20 pictures were used as test samples. A total of five runs were performed, and the average of five results was used as the classification accuracy.

The parameters of the M^{3}SC method are set as follows: The dictionary size of the A1 and B1 paths in layer 1 is 500; the dictionary sizes of the first layer of the A2 and B2 paths in layer 2 are 75 and 200, respectively. The dictionary size in the second layer is 500; the three layers dictionary size of the path B3 in layer 3 is 75, 300, and 500; the size of the patching and pooling in each path is shown in Fig. 1.

Figs. 4 and 5 show the initialised DCT dictionary and the dictionary obtained through K-SVD learning, respectively. It can be seen that compared to the DCT dictionary, the K-SVD learned dictionary contains more abundant image feature information, not only contains multiple line segments with directionality but also contains information related to colours.

Next, we discuss the effect of different path combinations on the classification accuracy, as shown in Table 1. Since the number of different types of test samples is the same, the average accuracy (AA) and overall accuracy (OA) are also the same, so the AA is not listed in this table. From Table 1 we can see that in a single path, the accuracy of A2 is the highest, with the OA of 0.8548 and the Kappa coefficient of 0.8475. In the multi-path combination, the accuracy of the A1 + A2 path is relatively high, with the OA of 0.9095 and the Kappa coefficient of 0.9050. Overall, the classification performance of the method considering all five routes is the highest, with the OA of 0.9214 and the Kappa coefficient of 0.9175. Unless otherwise specified, the M^{3}SC method mentioned in this paper refers to as the method that considers all paths.

Furthermore, we examine the effect of different paths on class-specific accuracy, as shown in Fig. 6. We can observe that on the UCM data set, the method of considering all paths gets much higher classification accuracies than single paths for baseball-diamond, buildings, freeway, golf-course, medium-residential, mobile-homepark, overpass, sparse-residential, and storage-tanks.

Fig. 7 shows the confusion matrix for the M^{3}SC classification method on UCM data. From this figure, it can be seen that the confusion between medium-residential, dense-residential, buildings, dense-residential is relatively large. Fig. 8 gives examples of several misclassified images on the UCM data set. In Fig. 8*a*, the true label is dense-residential while the predicted label is medium-residential; in Fig. 8*c*, the medium-density residential area is misclassified as high density residential area; in Fig. 8*e*, storage-tanks are mistakenly classified as sparse-residential in Fig. 8*f*, tennis-courts are misclassified ad dense-residential (high density residential area). There are two main reasons for the misclassification: (i) The category attributes are relatively similar. For example, even for human eyes, it is difficult to distinguish whether it is a high density residential area or a medium-density residential area; (ii) The pictures contained various objects. For

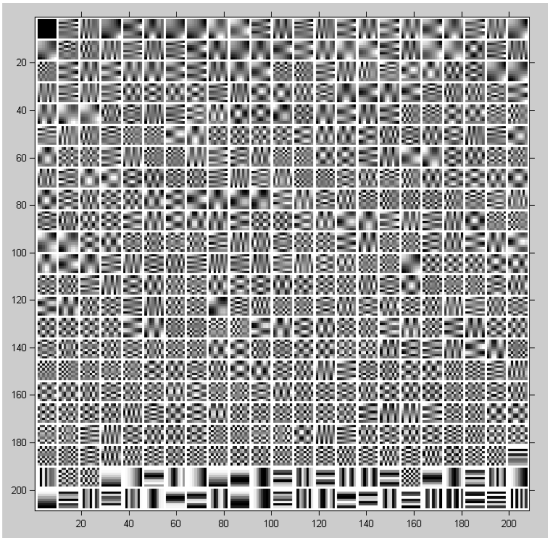

Fig. 4 *Initialised DCT dictionary*

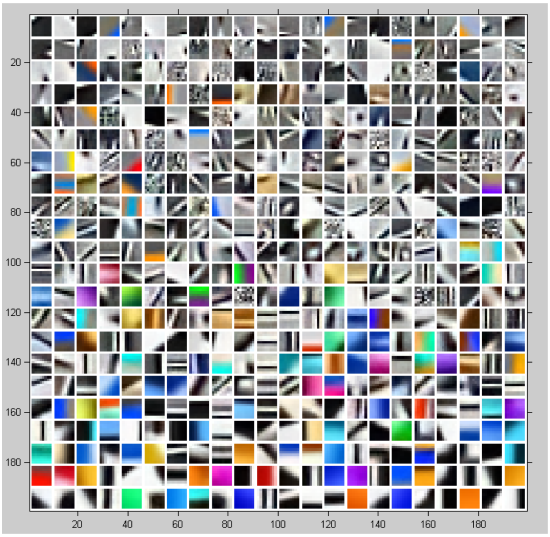

Fig. 5 *Dictionary obtained by K-SVD*

Table 1 Accuracy comparison of different path combinations on the UCM dataset

Path	OA	Kappa
A1	0.7857	0.7750
A2	0.8548	0.8475
B1	0.8048	0.7950
B2	0.8262	0.8175
B3	0.8262	0.8175
A1 + A2	0.9095	0.9050
B1 + B2	0.8762	0.8700
B1 + B3	0.8857	0.8800
B2 + B3	0.8810	0.8750
B1 + B2 + B3	0.8929	0.8875
ALL	0.9214	0.9175

example, the buildings in Figs. 8*e* and *f* interfere with the classification of the scene. The misclassification of Fig. 8 is basically consistent with the confusion between the classes of the confusion matrix shown in Fig. 7.

Table 2 compares the classification accuracy of the proposed M^{3}SC method with other over 30 methods on the UCM dataset. In each method, the training sample and the test sample are divided by 80 and 20%. Compared with the earlier SPMK [17] and BOVW [17] methods, the M^{3}SC method has an ∼15% improvement in OA. In recent years, deep learning or other hierarchical methods have been used more and more in UCM remote sensing image scene classification. Among them, the classification accuracy of EPLS + CNN [18] is 84.53%, 89.1% for MFL [19], 89.90% for LPCNN [20], 90.5% for HCV [21], and 91.12% for TLFR [22]. HCV-FV [21] combines the HCV method with the Fisher vector method, achieving a classification accuracy of 91.8%. The proposed M^{3}SC method has further improved the classification performance, achieving an OA of 92.14%.

3.2 Experiments on WHU-RS dataset

The WHU-RS data set was provided by the Signal Processing Laboratory of Wuhan University. The data set includes 19 images of the ground scene category. The image size is 600 × 600. Fig. 9 gives an exemplary image of this data set. Each category randomly selected 30 pictures as training samples and the remaining pictures as test samples. The number of test samples of the corresponding category in Figs. 9*a–s* is 25, 20, 22, 26, 20, 20, 20, 23, 23, 31, 20, 20, 20, 24, 23, 20, 24, 26, and 28, respectively. A total of five runs were performed and the average of five results was used as the classification accuracy.

The effect of different path combinations on the classification accuracy is shown in Table 3, which gives the OA, AA, and Kappa coefficient under different path combinations. As can be seen from the table, the accuracy of B3 in the single path is the highest, with the OA of 0.7655, the AA of 0.7710 and the Kappa coefficient of 0.7523. In addition, the multi-path combination has significantly improved the classification performance over a single path. In the

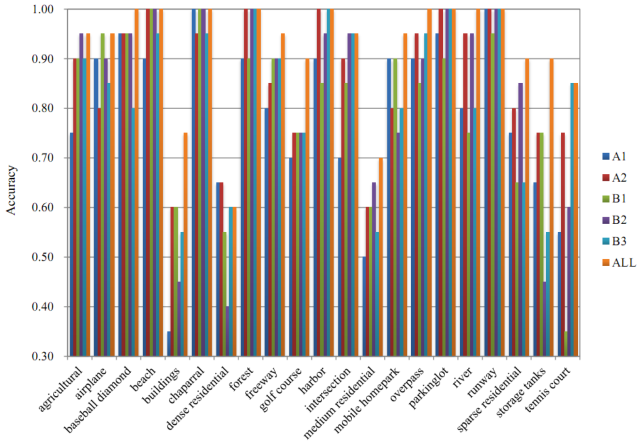

Fig. 6 *Class-specific accuracy of different paths on the UCM dataset*

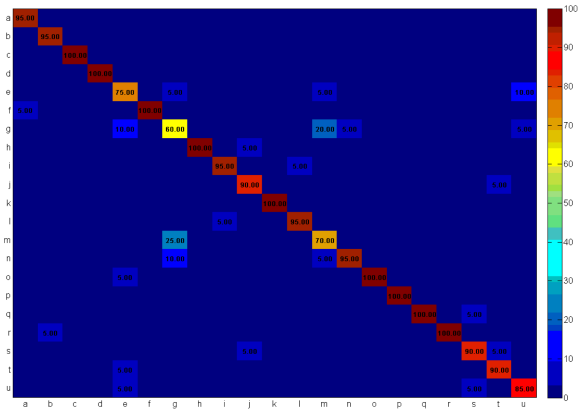

Fig. 7 *Confusion matrix (in per cent) of the proposed method on the UCM dataset (values <5% are not displayed)*

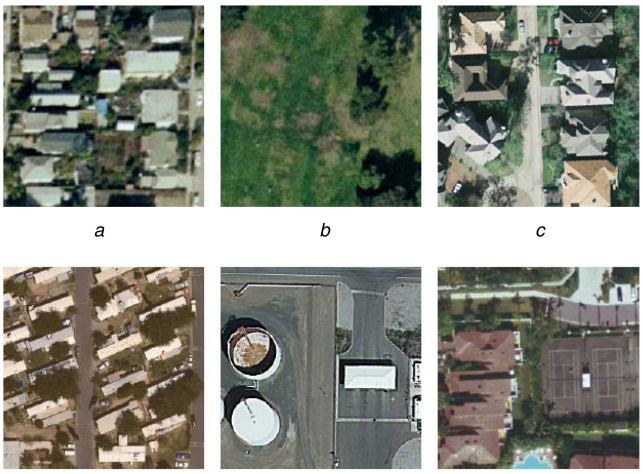

Fig. 8 *Samples of misclassification on the UCM dataset*
(a) Dense-residential→medium-residential, ***(b)*** Golf-course→forest, ***(c)*** Medium-residential→dense-residential ***(d)*** Mobile-homepark→dense-residential, ***(e)*** Storage-tanks→sparse-residential, ***(f)*** Tennis-court→dense-residential

multi-path combination, the accuracy of the B1 + B2 + B3 path is relatively high, with an OA of 0.8460, an AA of 0.8505, and a Kappa coefficient of 0.8373. Overall, the classification performance of the method considering all five paths is the highest, with the OA of 0.8483, the AA of 0.8530, and the Kappa coefficient of 0.8397.

From the classification accuracy of various categories, we can see from Fig. 10 that the method of considering all the paths in the WHU-RS data set has a greater degree of improvement than the single path for airport, forest, meadow, and railway stations. Fig. 11 shows the confusion matrix for the M^{3}SC classification method on the WHU-RS data.

Table 4 compares the accuracy of our proposed M^{3}SC method with other methods on the WHU-RS dataset. For each method, 30 samples are selected as training samples and the remaining samples as test samples. The bag of colours model and the tree of c-shapes model [35] achieve classification accuracy of 70.63 and 80.42%, respectively. The classification accuracy of LTP-HF [36] and SIFT [36] are 77.6 and 82.8%, respectively. Among several methods, the M^{3}SC method achieves the highest classification accuracy of 84.83%, again confirming the effectiveness of our proposed method.

Table 2 Accuracy comparison of different methods on the UCM dataset

Method	Accuracy, %	Year	Method description	References
BOVW	76.81	2010	bag-of-visual-words	[17]
SPMK	75.29	2010	spatial pyramid match kernel	[17]
SCK	72.52	2010	spatial co-occurrence kernel	[17]
BOVW + SCK	77.71	2010	combination of BOVW and SCK	[17]
SPCK	73.14	2011	spatial pyramid co-occurrence kernel	[1]
SPCK +	76.05	2011	combination of SPCK and BOVW	[1]
SPCK + +	77.38	2011	combination of SPCK and SPMK	[1]
UFL	81.67	2014	unsupervised feature learning by sparse coding	[10]
wavelet-BOVW	87.38	2014	wavelet based bag-of-visual-words model	[23]
mCENTRIST	89.9	2014	multi-channel census transform histogram	[24]
COPD	91.33	2014	collection of part detectors	[25]
SVM-LDA	80.33	2015	combination of SVM and LDA	[5]
saliency-UFL	82.72	2015	saliency-guided unsupervised feature learning	[5]
MCBGP	86.52	2015	multi-channel binary Gabor patterns descriptor	[26]
MCBGP	86.52	2015	multi-channel binary Gabor patterns descriptor	[26]
BOW- RD	87.67	2015	bag-of-words using random dictionary	[27]
SAL-PTM	88.33	2015	semantic allocation level probabilistic topic model	[8]
PSR p	89.1	2015	pyramid of spatial relatons model	[2]
UFL-SC	90.26	2015	unsupervised feature learning via spectral clustering	[28]
partlets	91.33	2015	partlets-based method	[12]
ASP	80.7	2016	adaptive spatial pooling	[29]
SCM	84.31	2016	sparse correlaton model	[13]
EPLS + CNN	84.53	2016	sparse unsupervised deep convolutional networks	[18]
CLBP	85.5	2016	completed local binary patterns	[30]
FBC	85.53	2016	fast binary coding	[31]
CKC	86.28	2016	supervised collaborative kernel coding	[32]
ERT	86.69	2016	extremely randomised trees	[33]
MFL	89.1	2016	multi-layer feature learning via convolution and pooling operations	[19]
MS-BOV	89.10	2016	multi-scale bag-of-visual-words representation	[34]
LPCNN	89.90	2016	large patch convolutional neural networks	[20]
HCV	90.5	2016	hierarchical coding vectors	[21]
MS-CLBP	90.6	2016	multi-scale completed local binary patterns	[30]
TLFR	91.12	2016	two-level feature representation	[22]
HCV-FV	91.8	2016	hierarchical coding vectors with Fisher vectors	[21]
ours	92.14	2018	M^3SC	—

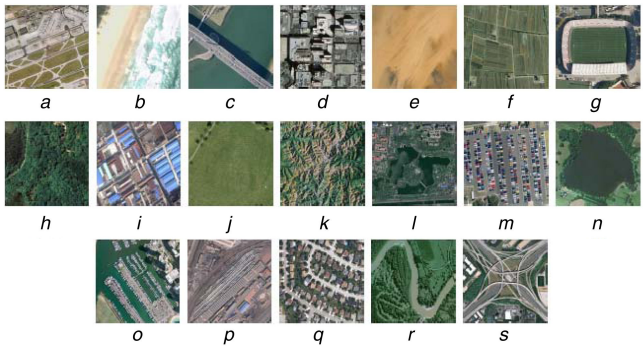

Fig. 9 *Sample images of the WHU-RS dataset*
(a) Airport, ***(b)*** Beach, ***(c)*** Bridge, ***(d)*** Commercial, ***(e)*** Desert, ***(f)*** Farmland, ***(g)*** Football field, ***(h)*** Forest, ***(i)*** Industrial, ***(j)*** Meadow, ***(k)*** Mountain, ***(l)*** Park, ***(m)*** Parking, ***(n)*** Pond, ***(o)*** Port, ***(p)*** Railway station, ***(q)*** Residential, ***(r)*** River, ***(s)*** Viaduct

4 Conclusion

In this paper, a framework of remote sensing image scene classification based on hierarchical sparse coding is proposed. The method uses a M^3SC to extract the features of remote sensing images. Experiments on two high-resolution optical remote sensing images (UCM and WHU-RS dataset) show that the proposed method can obtain better scene classification accuracy.

5 Acknowledgments

This work is mainly supported by the National Key Research and Development Program of China (2016YFB0200203), the National Natural Science Foundation of China (61572509).

Table 3 Accuracy comparison of different path combinations on the WHU-RS dataset

Path	OA	AA	Kappa
A1	0.7218	0.7296	0.7064
A2	0.7586	0.7643	0.7451
B1	0.6575	0.6645	0.6383
B2	0.7517	0.7596	0.7378
B3	0.7655	0.7710	0.7523
A1 + A2	0.8414	0.8457	0.8325
B1 + B2	0.8253	0.8302	0.8155
B1 + B3	0.8437	0.8479	0.8349
B2 + B3	0.7931	0.7999	0.7815
B1 + B2 + B3	0.8460	0.8505	0.8373
all	0.8483	0.8530	0.8397

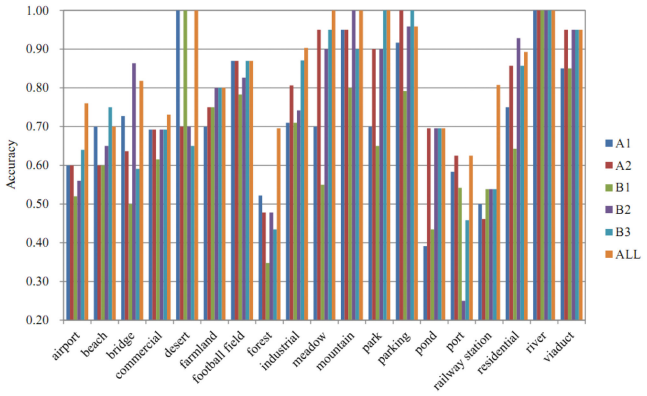

Fig. 10 *Class-specific accuracy of different paths on the WHU-RS dataset*

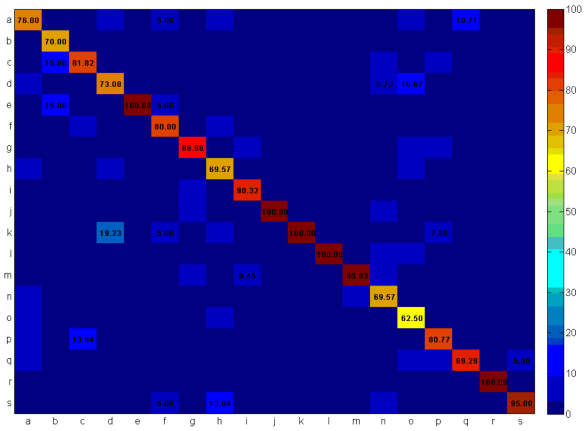

Fig. 11 *Confusion matrix (in per cent) of the proposed method on the WHU-RS dataset (values <5% are not displayed)*

Table 4 Accuracy comparison of different methods on the WHU-RS dataset

Method	Accuracy, %	Method description
bag of colours [35]	70.63	bag of features model using colour descriptors
tree of c-shapes [35]	80.42	tree of coloured shapes
LTP-HF [36]	77.6	local ternary pattern histogram Fourier features
SIFT [36]	82.8	scale invariant feature transform features
ours	84.83	multi-layer multi-scale multi-path sparse coding

6 References

[1] Yang, Y., Newsam, S.: 'Spatial pyramid co-occurrence for image classification'. 2011 Int. Conf. on Computer Vision, Barcelona, Spain, November 2011, pp. 1465–1472

[2] Chen, S., Tian, Y.: 'Pyramid of spatial relatons for scene-level land use classification', *IEEE Trans. Geosci. Remote Sens.*, 2015, **53**, (4), pp. 1947–1957

[3] Lienou, M., Maitre, H., Datcu, M.: 'Semantic annotation of satellite images using latent Dirichlet allocation', *IEEE Geosci. Remote Sens. Lett.*, 2010, **7**, (1), pp. 28–32

[4] Cheng, G., Guo, L., Zhao, T., *et al.*: 'Automatic landslide detection from remote-sensing imagery using a scene classification method based on BOVW and PLSA', *Int. J. Remote Sens.*, 2013, **34**, (1), pp. 45–59

[5] Zhang, F., Du, B., Zhang, L.: 'Saliency-guided unsupervised feature learning for scene classification', *IEEE Trans. Geosci. Remote Sens.*, 2015, **53**, (4), pp. 2175–2184

[6] Zhao, B., Zhong, Y., Xia, G.S., *et al.*: 'Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery', *IEEE Trans. Geosci. Remote Sens.*, 2016, **54**, (4), pp. 2108–2122

[7] Zhao, B., Zhong, Y., Zhang, L.: 'Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery', *Remote Sens. Lett.*, 2013, **4**, (12), pp. 1204–1213

[8] Zhong, Y., Zhu, Q., Zhang, L.: 'Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery', *IEEE Trans. Geosci. Remote Sens.*, 2015, **53**, (11), pp. 6207–6222
[9] Vaduva, C., Gavat, I., Datcu, M.: 'Latent Dirichlet allocation for spatial analysis of satellite images', *IEEE Trans. Geosci. Remote Sens.*, 2013, **51**, (2013–05), pp. 2770–2786
[10] Cheriyadat, A.M.: 'Unsupervised feature learning for aerial scene classification', *IEEE Trans. Geosci. Remote Sens.*, 2014, **52**, (1), pp. 439–451
[11] Zheng, X., Sun, X., Fu, K., *et al.*: 'Automatic annotation of satellite images via multifeature joint sparse coding with spatial relation constraint', *IEEE Geosci. Remote Sens. Lett.*, 2013, **10**, (4), pp. 652–656
[12] Cheng, G., Han, J., Guo, L., *et al.*: 'Effective and efficient midlevel visual elementsoriented land-use classification using VHR remote sensing images', *IEEE Trans. Geosci. Remote Sens.*, 2015, **53**, (8), pp. 4238–4249
[13] Qi, K., Zhang, X., Wu, B., *et al.*: 'Sparse coding-based correlaton model for land-use scene classification in high-resolution remote-sensing images', *J. Appl. Remote Sens.*, 2016, **10**, (4), p. Art. ID 042005, doi: 10.1117/1.JRS.10.042005
[14] Bo, L., Ren, X., Fox, D.: 'Hierarchical matching pursuit for image classification: architecture and fast algorithms'. NIPS'11 Proceedings of the 24th Int. Conf. on Neural Inf. Process. Syst., Granada, Spain, December 2011, pp. 2115–2123
[15] Aharon, M., Elad, M., Bruckstein, A.: 'K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation', *IEEE Trans. Signal Process.*, 2006, **54**, (11), pp. 4311–4322
[16] Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.S.: 'Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition'. Proc. of 27th Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, USA, November 1993, pp. 40–44
[17] Yang, Y., Newsam, S.: 'Bag-of-visual-words and spatial extensions for land-use classification'. Proc. ACM GIS, 2010, pp. 270–279
[18] Romero, A., Gatta, C., Camps-Valls, G.: 'Unsupervised deep feature extraction for remote sensing image classification', *IEEE Trans. Geosci. Remote Sens.*, 2016, **54**, (3), pp. 1349–1362
[19] Li, Y., Tao, C., Tan, Y., *et al.*: 'Unsupervised multilayer feature learning for satellite image scene classification', *IEEE Geosci. Remote Sens. Lett.*, 2016, **13**, (2), pp. 157–161
[20] Zhong, Y., Fei, F., Zhang, L.: 'Large patch convolutional neural networks for the scene classification of high spatial resolution imagery', *J. Appl. Remote Sens.*, 2016, **10**, (2), p. Art. ID 025006, doi: 10.1117/1.JRS.10.025006
[21] Wu, H., Liu, B., Su, W., *et al.*: 'Hierarchical coding vectors for scene level land-use classification', *Remote Sens.*, 2016, **8**, (5), p. Art. ID 436, doi: 10.3390/rs8050436
[22] Gan, J., Li, Q., Zhang, Z., *et al.*: 'Two-level feature representation for aerial scene classification', *IEEE Geosci. Remote Sens. Lett.*, 2016, **13**, (11), pp. 1626–1630
[23] Zhao, L., Tang, P., Huo, L.: 'A 2-D wavelet decomposition-based bag-of-visual-words model for land-use scene classification', *Int. J. Remote Sens.*, 2014, **35**, (6), pp. 2296–2310
[24] Xiao, Y., Wu, J., Yuan, J.: 'mCENTRIST: a multi-channel feature generation mechanism for scene categorization', *IEEE Trans. Image Process.*, 2014, **23**, (2), pp. 823–836
[25] Cheng, G., Han, J., Zhou, P., *et al.*: 'Multi-class geospatial object detection and geographic image classification based on collection of part detectors', *ISPRS J. Photogramm. Remote Sens.*, 2014, **98**, pp. 119–132
[26] Cvetkovic, S., Stojanovic, M.B., Nikolic, S.V.: 'Multi-channel descriptors and ensemble of extreme learning machines for classification of remote sensing images', *Signal Process., Image Commun.*, 2015, **39**, (11), pp. 111–120
[27] Cui, S., Schwarz, G., Datcu, M.: 'Remote sensing image classification: no features, no clustering', *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2015, **8**, (11), pp. 5158–5170
[28] Hu, F., Xia, G.S., Wang, Z., *et al.*: 'Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification', *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2015, **8**, (5), pp. 2015–2030
[29] Liu, Y., Zhang, Y., Zhang, X., *et al.*: 'Adaptive spatial pooling for image classification', *Pattern Recognit.*, 2016, **55**, pp. 58–67
[30] Chen, C., Zhang, B., Su, H., *et al.*: 'Land-use scene classification using multi-scale completed local binary patterns', *Signal. Image. Video. Process.*, 2016, **10**, (4), pp. 745–752
[31] Hu, F., Xia, G., Hu, J., *et al.*: 'Fast binary coding for the scene classification of high-resolution remote sensing imagery', *Remote Sens.*, 2016, **8**, p. Art. ID 555, doi: 10.3390/rs8070555
[32] Yang, C., Liu, H., Wang, S., *et al.*: 'Scene-level geographic image classification based on a covariance descriptor using supervised collaborative kernel coding', *Sensors*, 2016, **16**, p. Art. ID 392, doi: 10.3390/s16030392
[33] Maree, R., Geurts, P., Wehenkel, L.: 'Towards generic image classification using treebased learning: an extensive empirical study', *Pattern Recognit. Lett.*, 2016, **74**, pp. 17–23
[34] Zhang, J., Li, T., Lu, X., *et al.*: 'Semantic classification of high-resolution remote-sensing images based on mid-level features', *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2016, **9**, (6), pp. 2343–2353
[35] Shao, W., Yang, W., Xia, G., *et al.*: 'A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization'. Proc. Int. Conf. on Computer Vision Systems, St. Petersburg, Russia, July 2013, pp. 324–333
[36] Sheng, G., Yang, W., Xu, T., *et al.*: 'High-resolution satellite scene classification using a sparse coding based multiple feature combination', *Int. J. Remote Sens.*, 2012, **33**, (8), pp. 2395–2412