Journal of Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Multiscale deep features learning for land-use scene recognition

Baohua Yuan
Shijin Li
Ning Li

SPIE.

Baohua Yuan, Shijin Li, Ning Li, "Multiscale deep features learning for land-use scene recognition," *J. Appl. Remote Sens.* **12**(1), 015010 (2018), doi: 10.1117/1.JRS.12.015010.

Multiscale deep features learning for land-use scene recognition

Baohua Yuan,[a,b] Shijin Li,[a,*] and Ning Li[a]

[a]HoHai University, School of Computer and Information Engineering, Nanjing, China
[b]Taizhou Institute of Science and Technology of NanJing University of Science and Technology, Department of Computer Science and Technology, Taizhou, China

Abstract. The features extracted from deep convolutional neural networks (CNNs) have shown their promise as generic descriptors for land-use scene recognition. However, most of the work directly adopts the deep features for the classification of remote sensing images, and does not encode the deep features for improving their discriminative power, which can affect the performance of deep feature representations. To address this issue, we propose an effective framework, LASC-CNN, obtained by locality-constrained affine subspace coding (LASC) pooling of a CNN filter bank. LASC-CNN obtains more discriminative deep features than directly extracted from CNNs. Furthermore, LASC-CNN builds on the top convolutional layers of CNNs, which can incorporate multiscale information and regions of arbitrary resolution and sizes. Our experiments have been conducted using two widely used remote sensing image databases, and the results show that the proposed method significantly improves the performance when compared to other state-of-the-art methods. © *2018 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JRS.12.015010]

Keywords: convolutional neural network; convolutional features; locality-constrained affine subspace coding; multiscale ensemble; land-use scene recognition.

Paper 170935 received Oct. 28, 2017; accepted for publication Jan. 16, 2018; published online Feb. 9, 2018.

1 Introduction

In the past decade, with the ongoing development of various satellite sensors, a large volume of high-resolution remote sensing image data have become available. These high-resolution remote sensing images are generally rich in spatial arrangement information and textural structures, which are of great help in recognizing different land-use scene categories. Nevertheless, the high-resolution remote sensing images introduce new challenges in smart image interpretation.

In the past few years, there has been an intense research of remote sensing scene classification, with the focus on both the use of appropriate image descriptors and the appropriate classification task.[1–17] Bag-of-visual-words (BOVW) is one of the representative models in the field of image analysis and classification. The BOVW model represents an image as an orderless collection of local features (SIFT, HOG, etc.) extracted from a collection of images. The basic version of BOVW, however, neglects information on the spatial distribution of visual words. Hence, there have been several efforts in the literature to overcome the weakness. The BOVW model incorporating the spatial information of scene images has been successfully applied to remote sensing land-use scene classification and has exhibited good performance.[5] The spatial pyramid match kernel[18] is one approach to tackle the lack of spatial information. It consists of repeatedly subdividing the image and computing histograms of local features at increasingly fine resolutions. However, all the above-mentioned approaches are based on manually designed features, which heavily depend on the experience and domain knowledge of experts. Moreover, such features cannot adequately represent the complex image structures. This is mainly due to the lack of consideration for the details of remote sensing data.

*Address all correspondence to: Shijin Li, E-mail: lishijin@hhu.edu.cn

1931-3195/2018/$25.00 © 2018 SPIE

In 2006, a breakthrough in deep feature learning was made by Hinton and Salakhutdinov.[19] Since then, the aim of researchers has been to replace hand-engineered features with trainable multilayer networks and an amount of deep learning models have shown impressive feature representation capability for a wide range of applications including remote sensing image scene classification.[20] A number of recent works[6–10,21–24] show that convolutional neural networks (CNNs) pretrained on such large datasets have been shown to contain general-purpose feature extractors, transferrable to many other domains with a limited amount of training data. Employing the pretrained CNNs and fine-tuning them on the scene datasets, Penatti et al. experimentally evaluated ConvNets, showing impressive classification performance.[8] Marmanis et al. investigated the potential of using large pretrained neural networks for land-use classification and showed promising results on a public remotely sensed scene dataset.[7] Castelluccio et al. explored three design modalities of CNN for the semantic classification of remote sensing scenes and achieved a significant performance improvement.[10]

However, general-purpose features extracted from CNNs pretrained on such a large dataset contain redundant information, which limits their performance for classification and robustness of highly variable land-use scenes. In recent years, most works mainly focus on the pooling scheme of deep learning. This MOP-CNN (multiscale orderless pooling)[22] extracts deep features from local patches at multiple scale, performs orderless VLAD encoding of these local patch activations at each level separately, and concatenates the result to form a new image representation. Cimpoi et al. proposed FV-CNN (Fisher vector pooling of a convolutional neural network),[23] which built on the fully connected layers of CNN form an orderless representation by fisher vector. Motivated by previous work on spatial and feature space pooling of local descriptors,[11,22,23,25] we develop a simple but effective framework for land-use scene recognition, which we refer to as locality-constrained affine subspace coding (LASC) pooling (LASC-CNN). This method builds universal image representations from CNN models with no training phase or use of labels. It first extracts the deep convolutional activations of an input multiscale image by an ImageNet pretrained networks. These deep activations are then encoded into a new high dimensional feature representation by overlaying a spatial pyramid partition. Then, a new feature representation is encoded via LASC forms the final image-level representation. The LASC, which can describe manifolds of high dimensional deep features by an ensemble of subspace attached to affine subspace, can obtain a more discriminative scene representation, and may exhibit better performance, as is commonly done in the BOVW approaches.[5]

The rest of this paper is organized as follows. Section 2 describes details of LASC-CNN methodology. The problem of multiscale deep feature extraction is then discussed in Sec. 2. Section 3 provides extensive experiments and analysis on the effectiveness of using our method. A conclusion is drawn in Sec. 4.

2 Description of the Proposed Method

The flowchart of the proposed method is shown in Fig. 1. The idea is to regard the convolutional layers of a CNN as a filter bank and build an orderless representation using LASC as a pooling mechanism, as is commonly done in the bag-of-words approaches. In the following sections, we respectively present the multiscale deep feature extraction method and the LASC-CNN algorithm in detail.

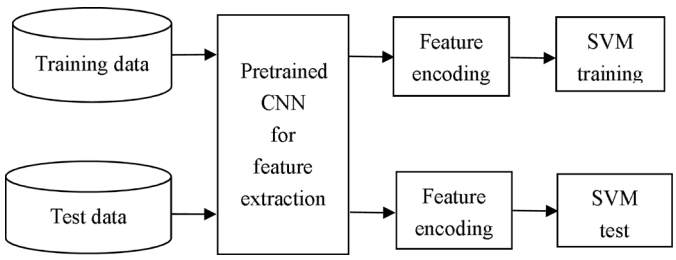

Fig. 1 The image classification framework.

2.1 *Multiscale Deep Feature Extraction*

The CNN is a trainable multilayer architecture composed of multiple feature-extraction stages, each comprising both linear and nonlinear operators, which are learnt jointly, in an end-to-end manner, to solve specific tasks.[6,23] Specifically, a typical CNN is commonly made up of four layers: (1) convolutional layer, (2) normalization layer, (3) pooling layer, and (4) fully connected layer. In actual CNNs, each layer comprises various sublayers of neurons operating in parallel on the previous layer, so as to extract a number of features at once, like a bank of filter does. Every filter is small spatially (along width and height), but extends through the full depth of the input image. A pretrained network can be used as a feature extractor for any image, since the generic features (learned in earlier layers) are less dependent on the final application and could be used in a myriad of tasks.

Recent advances in convolution layer features[11,24] are adapted to remote sensing data and shown to be as effective as in other domains. The adoption of convolution layer features to replace the full connection layer features has the following advantages. The first one is that features from convolutional layers are more generic than those from fully connected layers,[23] and thus these features may be more suitable for transfer learning. In addition, convolutional layer features contain more spatial information, which offers advantages for image classification, as compared to the activation of fully connected layers.[21] Furthermore, scale variation, which requires considering multiscale contextual and structural information in spatial domain, is quite common for objects detection in remote sensing images (e.g., roofs with different sizes). A second advantage is that the input to the CNN has to be of fixed size to be compatible with the fully connected layers, which requires an expensive resizing of the input image.

The feature map of top convolutional layers is known to contain mid- and high-level information, e.g., object parts or complete objects.[24] These deep descriptors contain more spatial information compared to the activation of the fully connected layers. The fully connected layers require a fixed image size (e.g., 224×224). On the contrary, convolutional layers accept input images of arbitrary resolution or aspect ratio. In addition, convolutional layer features contain more spatial information than fully connected layers. In this work, we take the output of the convolutional layers (before the fully connected layer) to represent the training and test images.

The objects of interest generally have different scales in different remotely sensed scenes, and even a single scene may contain objects with different sizes. Accordingly, a multiscale spatial feature extraction technique is proposed for improving classification accuracy. However, most of the CNNs require a fixed input image size. Therefore, it is difficult to extract multiscale deep features simultaneously from one network. The spatial pyramid pooling (SPP)-net method[21] adds a spatial pyramid pooling layer to deep nets, which allows us to feed images with varying sizes or scales during training. To explore multiscale deep features, we propose to adopt an SPP-net based framework to learn spatial features across different scales. Instead of relying on a fixed observation scale, a series of images at different observation scales are fed into the entire network for extracting multiscale features. Then, we use LASC to pool activations from a fully convolutional network.

We can form a spatial pyramid[26,27] by partitioning the cells of activations in the last convolutional layer into subregions and then pool deep descriptors in each region separately using LASC. Following Ref. 12, the spatial pyramid matching (SPM) structure as shown in Fig. 2 is employed for all the experiments.

The level 0 simply aggregates all cells using LASC. The level 1, however, splits the cells into five regions according to their spatial locations: the four quadrants and one centerpiece. Then,

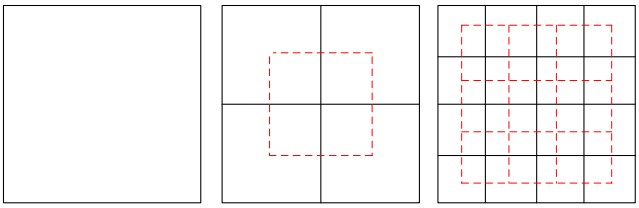

Fig. 2 Spatial pyramid matching.

five LASCs are generated from activations inside each spatial region. The level 2 splits the cells into 25 regions. Then, 25 LASCs are generated from activations inside each spatial region. The output of the spatial pyramid is realized by concatenating all 32 (25 + 6 + 1) LASCs from level 0 to level 2 to form the final image representation.

2.2 *LASC-CNN*

The deep CNN descriptors extracted from an image via a pretrained CNN model can be directly used to measure the scene similarity of a pair of images. However, as deep CNN descriptors are of high dimensionality and contain redundant information, we can encode the feature to obtain a more compact and discriminative representation, as pointed out in Refs. 22 and 23. The LASC-CNN descriptor is related to the FV-CNN proposed by Cimpoi et al. We need to pool the multiscale deep feature representation by LASC. Recently, Li et al. proposed a feature encoding method called LASC.[25] Taking into account the geometric subspace structure surrounding each visual word, the LASC algorithm characterizes the data manifold by an ensemble of subspaces attached to the representative points, resulting in a favorable classification performance in publicly available datasets.

Let y be an input feature to be encoded, μ_i indicates the i'th central representative point, A_i is an $n \times p$ matrix consisting of orthogonal basis of the linear subspace, and x_i are the linear approximate coefficients. The LASC is formulated as the following objective function:

$$\min_{\forall x_i} \sum_{S_i \in N_K^S(y)} \left\| (y - u_i) - A_i x_i \right\|_2^2 + \lambda \sum_i d(y, S_i) \|x_i\|_2^2. \tag{1}$$

Here, $\lambda > 0$ is a regularization parameter, $d(y, S_i)$ indicates the distance between y and its i'th subspace S_i determined by the proximity measure function, and $N_K^S(y)$ is the neighbor region of y defined by its K closest subspaces, where $\| \cdot \|_2$ indicates the Euclidean distance.

The method performs descriptor encoding only in a few most neighboring subspaces. We segment the deep feature space by the k-means algorithm to obtain mean vector (μ_i) of clusters. Then, we employ PCA to preserve the most significant principal directions (A_i) with larger variances. x_i can be regarded as the orthogonal projection of $y - \mu_i$ in the subspaces. Thus, the objective function Eq. (1) can be a simple form as

Algorithm 1 The LASC-CNN method.

Input: pyramid images $I = [I_1, \ldots, I_m]$; A pretrained CNN model;

Output: Accuracy;

1: **for all** $1 \le i \le m$ **do**

2: Extract deep descriptors X_i from I_i using the predefined model, $X_i = [x_1, \ldots, x_n]$;

3: Generate a spatial pyramid $\{X_i^1, \ldots, X_i^n\}$ for ;

4: **for all** $1 \le j \le n$ **do**

5: Encoding deep descriptors F_i^j (X_i^j) by LASC for X_i^j;

6: **end for**

7: Concatenate $F_i^j(X_i^j)$ to form the final spatial pyramid representation $F_i(X_i)$;

8: **end for**

9: Concatenate $F_i^{(X_i)}$ to $F(X)$ form the multiscale deep feature representation

10: Using an SVM as a classifier for land-use scene classification.

$$x_i = [1 + \lambda d(y, S_i)]^{-1} A_i^T (y - u_i). \quad (2)$$

Following Ref. 25, for multiscale deep features, we find its top-k nearest affine subspaces and perform linear decomposition in these subspaces weighted by the proximity measure. We produce the first-order (linear) and second-order LASC vector of the descriptor $x = [\ldots, x_i, x_i^2 \ldots]$.

In the first step, we feed multiscale images into a pretrained CNN model to extract deep activations. Then, a visual dictionary is trained on the deep descriptors from training images. The third step overlays a spatial pyramid partition to the deep activations of an image, pools deep descriptors in each region separately using LASC, and then concatenates these regions feature to form the multiscale deep feature representation. Finally, using a support vector machine (SVM) as a classifier for land-use scene classification. The LASC-CNN method is summarized in Algorithm 1.

3 Experiments

Using two remote-sensing datasets, we carried out a number of experiments to assess the performance of the proposed approach in comparison with state-of-the-art results. The well-known UC Merced Land Use dataset[5] (UCMerced for short[28]), includes aerial optical images, with low-level characteristics similar to those of the Imagenet. In recent years, many researchers have used this dataset, allowing for an extensive comparison of results with the literature. All deep learning models are required to be trained on large training datasets with abundant and diverse images to avoid overfitting. The NWPU-RESISC45[29] is created by Northwestern Polytechnical University (NWPU[30] is a large-scale dataset with big image variations and diversity). Since it has been published very recently, limited results are available, including results with CNNs. In the next two sections, we discuss results separately for the two datasets. All experiments have been carried out on a notebook equipped with an NVIDIA GeForce GT 750M 2048 MB GPU. In feature vector modality, only the last fully connected layer is trained. In the LASC method, as this was empirically shown, the number of nearest subspaces $k = 3$, subspace dimension $M = 64$, regularization parameter $\lambda = 1$, we normalize the first-order and second-order subvectors separately per subspace by $l2$ norm. Following Ref. 25, too small parameters (number of nearest subspaces, subspace dimension) are insufficient to describe the structure of the subspace, while much larger ones give little benefit.

3.1 *UCMerced Land Use Dataset*

3.1.1 *Dataset description and experimental setup*

This dataset contains land-use aerial orthoimagery from 21 classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golfcourse, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparseresidential, storage tanks, and tennis courts. Each class contains 100 images, which are cropped to 256×256 pixels. This is a challenging dataset due to a variety of spatial patterns in those 21 classes. The dataset represents highly overlapping classes such as the dense residential, medium residential, and sparse residential, which mainly differs in the density of structures. Sample images of each land-use class are shown in Fig. 3.

In each experiment, besides the original scale, the images are warped into three different scales, including 128×128, 204×204, and 256×256 pixels. We choose 256×256 pixels to be consistent with the input scale of the pretrained CaffeNet. The VGGNet-16 model, which was pretrained on ImageNet dataset, is available in a Github repository: https://github.com/BVLC/caffe/wiki/Model-Zoo for deep CNN feature extraction. The dataset is randomly divided into two sets: the training set and the testing set. For the pixels in each convolutional layer, K-means clustering is employed to form the visual codebook with 300 code words. The encoded deep features by LASC are then fed into SVM classifiers with histogram intersection kernels, which are implemented using the LIBSVM package,[31] and one-against-one strategy is adopted to address the multiclass issue. The testing set is used to evaluate the performance of classifiers. In

Fig. 3 Example images from the UCMerced dataset.

order to reduce the effect of random selection, we repeat each algorithm execution on 10 different training/testing splits of the dataset and report means and standard deviations of the obtained accuracies.

We randomly select samples of each class for training the SVM classifier and the rest for testing, following the same sampling setting as Ref. 5 for the datasets: 80 training samples per class for the UCMerced dataset.

3.1.2 *Different CNN feature coding*

We focus on testing different CNN feature coding discriminative powers. To make a sufficient comparison, LASC-CNN is compared with two other descriptors: (1) MOP-CNN and (2) FV-CNN. To generate codewords, we use the standard K-means clustering algorithm for all coding methods. Except for Fisher coding wherein the GMM is applied. We choose 256×256 pixels to be consistent with the input scale of the pretraining CaffeNet for all coding methods. Experimental results are shown in Table 1. From Table 1, we can see that LASC-CNN achieves good performance for remote sensing scene classification.

3.1.3 *Comparison with state-of-the-art methods*

Several approaches have been proposed recently for remote sensing scene classification, and most of them have been tested on the UCMerced dataset, following the same experimental protocol, with fivefold cross-validation. Therefore, there is plenty of data available for a solid comparison with the state-of-the-art. In Table 2, we report the overall accuracies for all these comparable methods, as they appear in the original papers, together with the accuracy of our best CNN solution.

An overview of the performance of multiscale LASC-CNN is shown in the confusion matrix in Fig. 4. There is some confusion between dense residential and medium residential. This can be

Table 1 Comparison of classification accuracy with different CNN feature coding on the UCMerced dataset.

Different encoding method	UCMerced (%)
MOP-CNN	94.1
FV-CNN	95.2
LASC-CNN	97.14

Table 2 Comparison of classification accuracy with the state-of-the-art methods on the UCMerced dataset.

Method	Accuracy (%)
BOVW[5]	76.8
SPM[5]	75.3
BOVW + spatial co-occurrence kernel[5]	77.7
Concentric circle-structure BOVW[13]	86.6 (±0.8)
Wavelet BOVW[14]	87.4 (±1.3)
Pyramid-of-spatial-relations[16]	89.1
CLBP[15]	85.5 (±1.9)
MS-CLBP1[15]	90.6 (±1.4)
cCENTRIST (HSV)[12]	75.2 (±2.4)
Sparse correlaton[4]	84.31 (±0.51)
meCENTRIST[17]	91.24 (±0.78)
GoogLeNet[6]	93.0
VGG16[6]	92.8 (±0.61)
CNN with overfeat feature[7]	92.4
Caffe[8]	93.42 (±1.0)
OverFeat[8]	90.91 (±1.19)
Our VGG16 (single-scale)	97.14
Our VGG16 (multiscale)	98.10

explained by the fact that the pairs of classes have similar spectral or structural features, such as both dense residential and medium residential. Therefore, more work needs to be done with regard to the use of the structural feature in the future.

3.2 *NWPU-RESISC45 Dataset*

3.2.1 *Dataset description and experimental setup*

The NWPU-RESISC45 dataset consists of 31,500 remote sensing images divided into 45 scene classes. Each class includes 700 images with a size of 256×256 pixels in the red green blue (RGB) color space. The spatial resolution varies from about 30 to 0.2 m per pixel for most of the scene classes except for the classes of island, lake, mountain, and snowberg that have lower spatial resolutions. These 45 scene classes are as follows: airplane, airport, baseball diamond, basketball, court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, dense residential, desert, forest, freeway, golfcourse, ground track field, harbor, industrial area, intersection, island, lake, meadow, medium residential, mobile home park, mountain, overpass, palace, parking lot, railway, railway station, rectangular farmland, river, roundabout, runway, sea ice, ship, snowberg, sparse residential, stadium, storage tank, tennis court, terrace, thermal power station, and wetland. Sample images of each land-use class are shown in Fig. 5.

3.2.2 *Different CNN feature coding*

To make a comprehensive evaluation, two training-test ratios are considered. (i) 10% to 90%: the dataset was randomly split into 10% for training and 90% for testing (70 training samples and

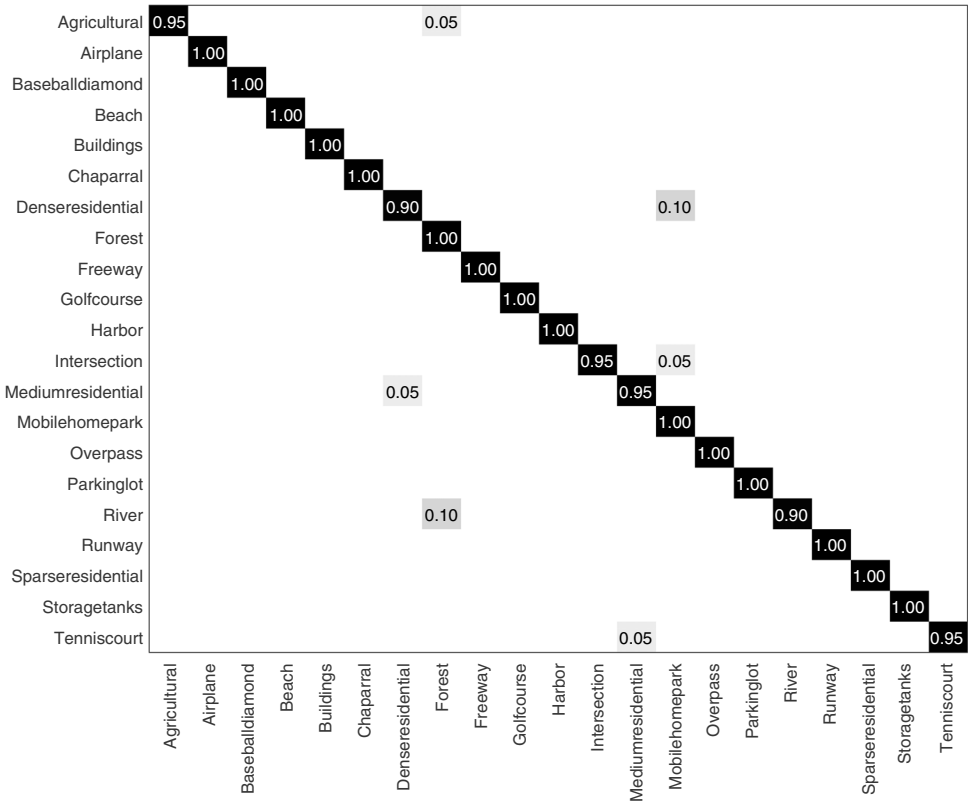

Fig. 4 Confusion matrix for the UCMerced dataset using the proposed LASC-CNN.

Fig. 5 Example images from the NWPU-RESISC45 dataset.

630 testing samples per class). (ii) 20% to 80%: the dataset was randomly divided into 20% for training and 80% for testing (140 training samples and 560 testing samples per class). Experimental results are shown in Table 3. Table 3 demonstrates the classification accuracies of different feature coding from each class and 256×256 input pixels. Same as the results on the

Table 3 Comparison of classification accuracy with different CNN feature coding on the NWPU-RESISC45 dataset.

	NWPU-RESISC45 Training ratios (%)	
Different encoding method	10%	20%
MOP-CNN	77.93	80.7
FV-CNN	78.32	81.6
LASC-CNN	80.68	84.21

UCMerced Land Use dataset, LASC-CNN achieves superior performance compared to the other feature coding method.

3.2.3 *Comparison with state-of-the-art methods*

In order to comprehensively analyze the superiority of the proposed method, we compare it with the three state-of-the-art approaches ever tested on this dataset, including LBP,[29] BoVW + SPM,[29] and CNN.[29] Table 4 reports the classification accuracies achieved by different methods with five training samples from each class. All of the LASC-CNN models are better than the best

Table 4 Comparison of classification accuracy (mean SD) with the state-of-the-art methods on the NWPU-RESISC45 dataset.

	NWPU-RESISC45 Training ratios (%)	
Different encoding method	10%	20%
LBP[29]	19.20 (±0.41)	21.74 (±0.18)
BoVW + SPM[29]	27.83 (±0.61)	32.96 (±0.47)
CNN[29]	76.47 (±0.18)	79.79 (±0.15)
LASC-CNN (single-scale)	80.69	83.64
LASC-CNN (multiscale)	81.37	84.30

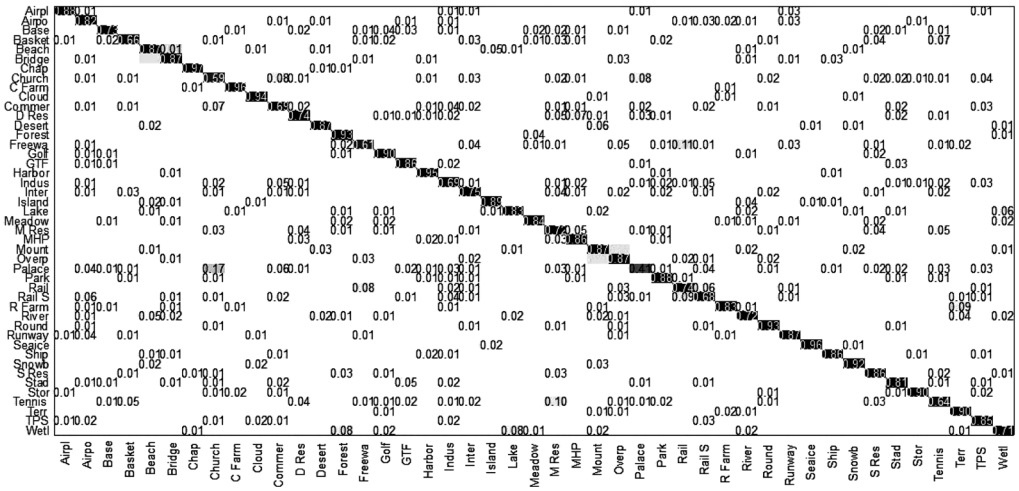

Fig. 6 Confusion matrix for the NWPU-RESISC45 dataset under the training ratio of 10% by using the proposed multiscale LASC-CNN.

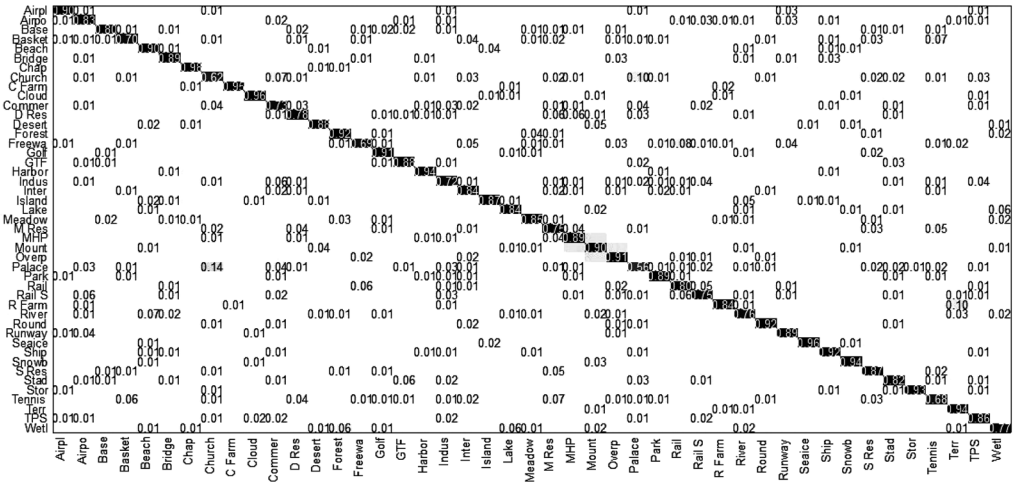

Fig. 7 Confusion matrix for the NWPU-RESISC45 dataset under the training ratio of 20% by using the proposed multiscale LASC-CNN.

state-of-the-art methods CNN, and the multiscale LASC-CNN can further increase the accuracy as compared to the single-scale models. In this paper, we mainly discuss the different encoding methods for deep features extracted from pretrained CNN. Certainly, by fine-tuning an off-the-shelf CNN model, the accuracy was further boosted by at least 6% points.[29] However, in comparison, fine-tuning the off-the-shelf CNN model is more time consuming and also requires a certain amount of data to train the deep network.

Figures 6 and 7 show the confusion matrices of multiscale LASC-CNN under the training ratios of 10% and 20%, respectively. For multiscale LASC-CNN-based CNN features, the relatively big confusions happen between church and palace and dense residential and medium residential because of their similar spectral or structural features. As expected, a larger training ratio can induce an increased recognition rate because of the availability of more spatial information. This suggests that a potential way to classify more challenging image scenes may be deep learning-based methods in combination of remote sensing data and spatial technology.

4 Conclusion

This paper has presented a multiscale orderless pooling scheme that is built on top of convolutional layers. The pooling scheme is to regard the convolutional layers of a CNN as a filter bank and build an orderless representation using LASC as a pooling mechanism, as is commonly done in the bag-of-words approaches. On two very challenging datasets, we have achieved a substantial improvement over global CNN activations, in some cases outperforming the state-of-the-art. The experimental results indicate that effectively encoding convolutional layer features can generate a more powerful representation, and that multiscale input images can provide much more discriminative information as compared to single-scale ones. In the future, it is interesting to integrate into LASC-CNN by the computational visual attention model[32] for remote sensing image retrieval.

Acknowledgments

This work was partially funded by the National Natural Science Foundation of China (Grant Nos. 61170200 and 61370091), the Key R&D Program of Jiangsu Province under Grant No. BE2015707, Non-profit science and technology research Program of MWR under Grant No. 201501022, and Water resources applications of the National High Resolution Earth Observation System (08-Y30B07-9001-13/15-01), Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (No. MJUKF201740).

References

1. S. Liu, Z. Zhang, and X. Mei, "Ground-based cloud classification using weighted local binary patterns," *J. Appl. Remote Sens.* **9**(1), 095062 (2015).
2. K. Chehdi, M. Soltani, and C. Cariou, "Pixel classification of large-size hyperspectral images by affinity propagation," *J. Appl. Remote Sens.* **8**(1), 083567 (2014).
3. B. Liao, W. Liu, and J. Shen, "Multispectral image fusion based on joint sparse subspace recovery," *J. Appl. Remote Sens.* **9**(1), 095068 (2015).
4. Q. Kunlun et al., "Sparse coding-based correlaton model for land-use scene classification in high-resolution remote-sensing images," *J. Appl. Remote Sens.* **10**(4), 042005 (2016).
5. Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. of the 18th SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems*, pp. 270–279, ACM (2010).
6. K. Nogueira, O. Penatti, and J. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.* **61**, 539–556 (2017).
7. D. Marmanis et al., "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geosci. Remote Sens. Lett.* **13**(1), 105–109 (2016).
8. O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 44–51, IEEE (2015).
9. G. J. Scott et al., "Training deep convolutional neural networks for land-cover classification of high-resolution imagery," *IEEE Geosci. Remote Sens. Lett.* **14**(4), 549–553 (2017).
10. M. Castelluccio et al., "Land use classification in remote sensing images by convolutional neural networks," arXiv preprint arXiv:1508.00092, https://arxiv.org/abs/1508.00092 (2015).
11. Q. Liu et al., "Adaptive deep pyramid matching for remote sensing scene classification," arXiv preprint arXiv:1611.03589, https://arxiv.org/abs/1611.03589 (2016).
12. Y. Xiao, J. Wu, and J. Yuan, "mCENTRIST: a multi-channel feature generation mechanism for scene categorization," *IEEE Trans. Image Process.* **23**(2), 823–836 (2014).
13. L.-J. Zhao, P. Tang, and L.-Z. Huo, "Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7**(12), 4620–4631 (2014).
14. L. Zhao, P. Tang, and L. Huo, "A 2-d wavelet decomposition-based bag-of-visual-words model for land-use scene classification," *Int. J. Remote Sens.* **35**(6), 2296–2310 (2014).
15. C. Chen et al., "Land-use scene classification using multi-scale completed local binary patterns," *Signal Image Video Process.* **10**(4), 745–752 (2016).
16. S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.* **53**(4), 1947–1957 (2015).
17. B. Yuan and S. Li, "Extended census transform histogram for land-use scene classification," *J. Appl. Remote Sens.* **11**, 025003 (2017).
18. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, Vol. **2**, pp. 2169–2178, IEEE (2006).
19. G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science* **313**(5786), 504–507 (2006).
20. L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: a technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.* **4**(2), 22–40 (2016).
21. K. He et al., "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1904–1916 (2015).
22. Y. Gong et al., "Multi-scale orderless pooling of deep convolutional activation features," in *European Conf. on Computer Vision*, pp. 392–407 (2014).
23. M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3828–3836, IEEE (2015).
24. B.-B. Gao et al., "Deep spatial pyramid: the devil is once again in the details," arXiv preprint arXiv:1504.05277, https://arxiv.org/abs/1504.05277 (2015).

25. P. Li, X. Lu, and Q. Wang, "From dictionary of visual words to subspaces: locality-constrained affine subspace coding," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2348–2357, IEEE (2015).
26. Y. Deng et al., "Differences help recognition: a probabilistic interpretation," *PLoS ONE* **8**(6), e63385 (2013).
27. J. Yang et al., "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1794–1801 (2009).
28. S. D. Newsam, "UC merced land use dataset," http://weegee.vision.ucmerced.edu/datasets/landuse.html.
29. G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: benchmark and state of the art," *Proc. IEEE* **105**(10), 1865–1883 (2017).
30. J. Han, "NWPU-RESISC45 dataset-Junwei Han," http://www.escience.cn/people/JunweiHan/NWPU-RESISC45.html.
31. C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intell. Syst. Technol.* **2**(3), 1–27 (2011).
32. G. Liu, J. Yang, and Z. Li, "Content-based image retrieval using computational visual attention model," *Pattern Recognit.* **48**(8), 2554–2566 (2015).

Baohua Yuan received his MS degree in computer science and technology from NanJing University of Science and Technology, Nanjing, China, in 2005. Currently, he is pursuing his PhD in computer science and technology at HoHai University, Nanjing, China. He is the author of more than 20 journal papers and has written five book chapters. His current research interests include high-resolution remote sensing image understanding and image retrieval.

Shijin Li received his MS and PhD degrees in computer science and technology from NanJing University of Science and Technology, Nanjing, China. Currently, he is a professor with the Department of Computer Science and Technology, HoHai University. He is the author of more than 50 journal papers. His current research interests include high-resolution remote sensing image understanding and computer vision.

Ning Li is a PhD in computer science and technology at Hohai University and an associate professor of software engineering at the University of Changzhou. His research interests include deep learning, image segmentation, object recognition, image annotation, and so on.