INTERNATIONAL JOURNAL OF REMOTE SENSING, 2016
VOL. 37, NO. 17, 4119–4131
http://dx.doi.org/10.1080/01431161.2016.1207266

Taylor & Francis
Taylor & Francis Group

Scene classification using multi-scale deeply described visual words

Wenzhi Zhao and Shihong Du

Institute of Remote Sensing and GIS, Peking University, Beijing, China

ABSTRACT

This article presents a deep learning-based Multi-scale Bag-of-Visual Words (MBVW) representation for scene classification of high-resolution aerial imagery. Specifically, the convolutional neural network (CNN) is introduced to learn and characterize the complex local spatial patterns at different scales. Then, the learnt deep features are exploited in a novel way to generate visual words. Moreover, the MBVW representation is constructed using the statistics of the visual word co-occurrences at different scales, which are derived from a training data set. We apply our technique to the challenging aerial scene data set: the University of California (UC) Merced data set consisting of 21 different aerial scene categories with sub-metre resolution. The experimental results show that the statistics of deeply described visual words can characterize the scene well and improve classification accuracy. It demonstrates that the proposed method is highly effective in the scene classification of high-resolution remote-sensing imagery.

ARTICLE HISTORY
Received 2 March 2016
Accepted 23 June 2016

提出的方法是 MBVW
深度描述视觉单词的统计

1. Introduction

Nowadays, with the rapid progress in remote-sensing technologies, a large number of images of the Earth are available. However, it is difficult to extract valuable information from such an amount of images, especially when incorporating very-high-resolution (VHR) images. Recently, intensive studies have been conducted in remote-sensing image classification and interpretation (Bellens et al. 2008; Blaschke 2010; Bruzzone and Carlin 2006; Shackelford and Davis 2003). Most of the existing classification methods depend on the spectral and spatial properties of pixels or segmented patches. However, image classification can only assign labels to pixels that do not have access to the semantic content of image scenes. Therefore, to directly retrieve images of interest, the semantic content of the bulk of images should be automatically obtained.

从图片中提取大量有用的信息较困难，尤其是超高分辨率图片，虽有一些方法提出但是很难从一些场景中提取语义信息进行分类。

Scene classification (Boutell et al. 2004; Yang et al. 2016) is a bridge to link low-level image information with the semantic content of images. However, this task becomes more difficult as abstraction levels increase, i.e. going from pixels to scenes (Li et al. 2010). Therefore, scene classification is very challenging, because the scenes of the same class may present a great variability and even scenes of different classes may share

CONTACT S. Du dshgis@hotmail.com Institute of Remote Sensing and GIS, Peking University, Beijing 100871, China

similar appearances, especially when dealing with VHR images (Myint et al. 2011). To date, several scene classification algorithms, aiming at automatically labelling images, have been proposed and have shown remarkable success in image interpretation (Fernando, Fromont, and Tuytelaars 2014; Luo et al. 2005; Zhou, Zhou, and Hu 2013). One particular method, called the bag-of-visual-words (BoVW), has been proposed and successfully applied to scene classification (Yang et al. 2007). It can effectively represent an image with local feature clusters, termed 'visual words'. However, the BoVW neglects the information on the spatial distributions of visual words. Hence, several spatial extensions of BoVW have been proposed, such as spatial pyramid match kernel (SPMK) (Lazebnik, Schmid, and Ponce 2006) and spatial pyramid co-occurrence kernel (SPCK++) (Yi and Newsam 2011). Unfortunately, these methods rely heavily on low-level feature descriptors for generating visual words, such as Scale-Invariant Feature Transform (SIFT) (Lowe 2004) or histograms of oriented gradients (HOGs) (Dalal and Triggs 2005). However, due to their strong intra-class variability and scale variation, low-level features in VHR images are unreliable and fail to accurately characterize a complex scene. Therefore, more effective and representative features at higher levels are necessary for capturing the semantic information of scenes.

如今一些针对场景分类的自动标识图像的分类方法提出来了并取得了较大进步，例如：BoVW

BoVW缺点忽略了视觉词汇的空间分布

Much research efforts have focused on the designing of more effective features for scenes recognition in the last few years and led to good results. For instance, the visual parts-based method (Felzenszwalb et al. 2010) was proposed to construct high-level feature representations for objects and pattern description. Moreover, filter banks were considered as predefined feature extractors in the work of Cheriyadat (Cheriyadat 2014). Instead of feature designing, deep learning can automatically learn the most effective feature representations from images and have shown great performances in several fields. Thus, even these sophisticated and successful designed descriptors are rapidly giving way to deep learning. Deep learning (Hinton, Osindero, and Teh 2006) takes inspiration from the hierarchical structure of the cognition functional zone of the human brain. Typically, deep learning frameworks comprise several layers of neurons feeding one another according to the hierarchical activation rules. Unlike the 'shallow' designed features mentioned above, deep learning utilizes 'deep' architecture to produce complex and abstract high-level deep features in a layer-wise manner. The learnt deep features are effective for describing complex objects and for further image understanding (Penatti, Nogueira, and dos Santos 2015). Thus far, deep learning has shown great potential in object recognition and image scene parsing in the field of computer vision (Jia et al. 2014; Lee et al. 2009). Therefore, the use of high-level deep features is advocated for complex remote-sensing scene classification. This article illustrates the effectiveness of deep learning in complex scene description. Some work has already been carried out in CNN-based remote-sensing scene classification (Jia, Liu, and Sun 2015). In these studies, CNNs were considered as the black box of image classifier (Hu, Xia, Wang, et al. 2015; Zhao et al. 2015; Hu, Xia, Hu, et al. 2015; Zhao and Du 2016a), which do not have access to the analysis of characteristics of scenes. Thus, to explore the basic contents of scenes, deeply described visual words at different scales (Zhao and Du 2016b) are introduced in this article. Specifically, we learn multi-scale visual words through deep convolutional neural networks (CNNs) and complex scenes can be described based on the learnt words. Experimental results on the University of California (UC) Merced data set indicate that the deeply described visual words using

深度学习的优势：能够自动且有效的学习这些特征

CNN很强大，但是被视为黑盒子，接着引出了“深度描述视觉词汇”

通过CNN学习视觉词，接着通过MBVW实现场景分类

CNN are much more effective and representative than the low-level feature descriptors. The experimental results indicate that Multi-scale Bag-of-Visual Words (MBVW) is a more efficient strategy for complex scene classification and it has achieved good performances in terms of classification accuracies.

2. The proposed method

In this section, the process of MBVW-based scene classification is briefly introduced. First, multi-scale training samples were randomly selected from the image pyramid, which later will be regarded as the input to the CNN framework. Then the CNN network can be trained with the feed of multi-scale training samples. In this step, the learnt deep features are flattened into vectors for further visual word generation. Third, the MBVW was applied to generate visual words from deep learning features at different scales. Finally, according to the trained visual word codebook, an unlabelled image can be classified according to the multi-scale visual word histogram. Figure 1 illustrates the detailed information about the classification of different scenes using deeply described visual words.

1）从图像金字塔中随机选择多尺度训练样本，后者将被视为CNN框架的输入
2）将学到的深层特征展平为矢量，以进一步生成视觉单词
3）应用MBVW从不同尺度的深度学习特征生成视觉词
4）根据训练的视觉词典码，可以根据多尺度视觉词直方图对未标记的图像进行分类

2.1. *Deep feature learning with CNN*

The first CNN, called NeoCognitron, was proposed by Fukushima (Fukushima 1980) and then refined by LeCun (LeCun et al. 1989). The CNN framework intuited by studies of the mammalian vision system, i.e. the visual cortex, is organized in layers and with lower layers to bright spots, corners, etc., which are then combined in higher layers to explore increasingly complex visual patterns (Hubel and Wiesel 1970). In CNN neurons have limited 'receptive fields', and all neurons in receptive fields share the same weights, as shown in Figure 2. Neurons of layer $(l+1)$ are connected to all neurons of the previous layer l. In this way, the image features extracted from the previous layer are passed on to

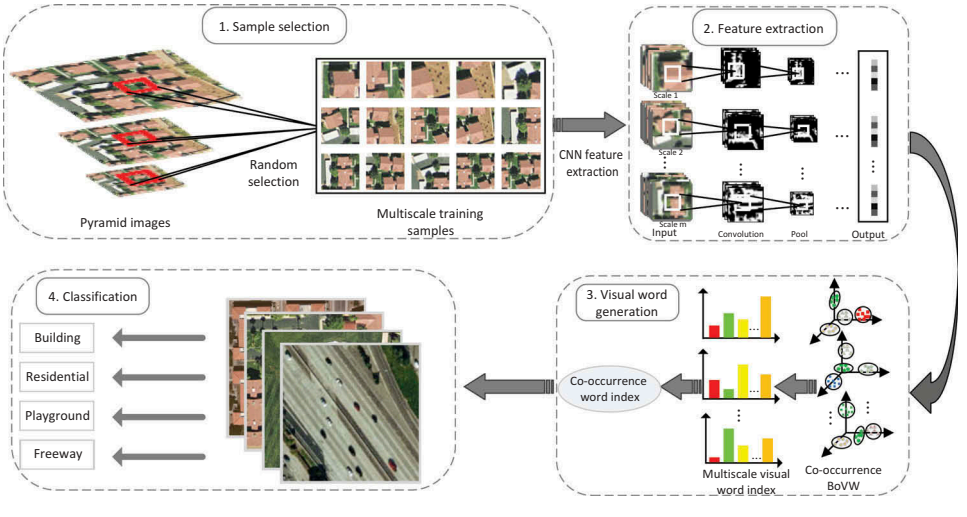

Figure 1. Flow chart of multi-scale deeply described visual words for scene classification.

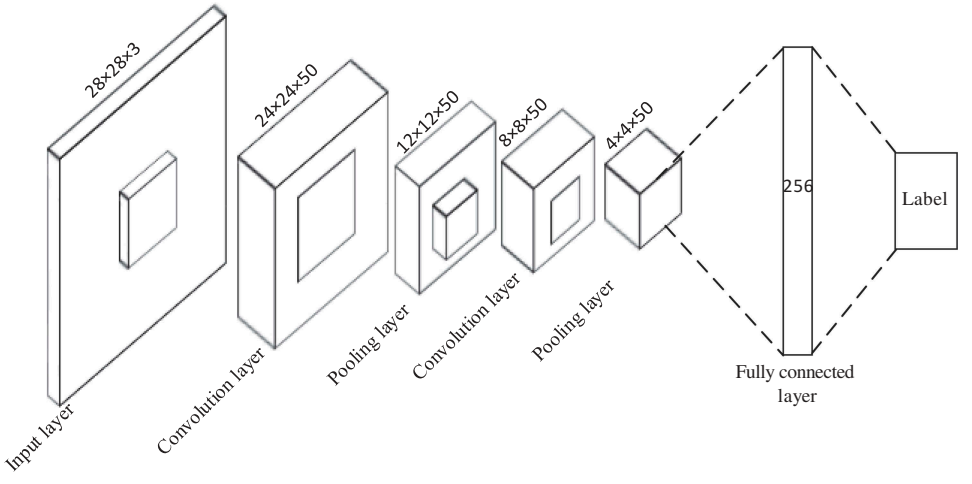

Figure 2. CNN framework with three kinds of layers: convolution layer, pooling layer, and fully connected layer.

the next layer for further processing. A typical CNN consists of three kinds of layers: convolutional layers, pooling layers, and fully connected layers.

(1) Convolutional layers. They work like a filter bank, which includes learnable weights of filters to produce convolutional information of the input image. The first convolutional layer works on only a small image window, thus only locally low-level features can be learnt. With going deeper, the convolutional layer views become broader, and thus they are able to learn more robust and complex features from low-level ones. Several parameters should be determined before training the whole CNN: the number of filters, the size of filters, etc.
(2) Pooling layers. They reduce the spatial size of the input layer through locally non-linear operations. With pooling layers, the size of an image is greatly reduced and becomes more general and translation invariance. The number of undetermined parameters is also reduced according to the spatial pooling strategy. The max-pooling operator is widely chosen for many applications.
(3) Fully connected layers. They typically appear in the last few layers of the network. A fully connected layer connects all neurons in the previous layer (pooling, or convolutional layers) to every single neuron of itself. With the fully connected layer, all of the features can be better summarized and proper classification results can be obtained.

The CNN can be described as a sequence of convolutional layers, interspersed with pooling layers. The maps in the pyramid of image $\mathbf{I}$, $\boldsymbol{P}_s$ are computed using a scaling function g_s, i.e. $\boldsymbol{P}_s = g_s(\mathbf{I})$, for all scales $s \in \{1, \ldots, N\}$. For each scales, a CNN f_s with parameters θ_s and L layers can be represented as

$$f_s(\boldsymbol{P}_s; \theta_s) = \mathbf{W}_L \boldsymbol{H}_{L-1}, \tag{1}$$

where the vector of hidden units at layer $l \in \{1, \ldots, L-1\}$ is

$$\boldsymbol{H}_l = \text{pool}(\text{sigmoid}(\mathbf{W}_l \boldsymbol{H}_{l-1} + b_l)), \tag{2}$$

where $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$. b_l represents a vector of bias parameters. The matrices $\mathbf{W}_l$ are learnable filter weights; thus, each hidden unit vector $\boldsymbol{H}_l$ can be expressed as a convolution operation between kernels from $\mathbf{W}_l$ and the previous hidden unit vector $\boldsymbol{H}_{l-1}$, squashed through a sigmoid function, and finally pooled spatially (max-pooling strategy was applied through this study). The filter parameters $\mathbf{W}_l$ and the bias parameter b_l constitute the trainable parameters θ_s of the CNN framework. Finally, the outputs of the N networks are concatenated so as to produce the output feature vector across all scales $\mathbf{O} = [f_s]_{s=1}^{N}$, where f_s refers to the output feature map of network$f$at scale$s$.

2.2. *MBVW with deeply described visual words*

2.2.1. *Traditional BoVW representation*

Visual words constructed from training data sets are the key to the BoVW method. Visual vocabulary is a novel approach to link the feature representation of undetermined images with the corresponding semantic labels. To explore visual words, the *k*-means clustering algorithm is commonly applied. For a specific scene, an image is first randomly split into several image patches based on the grid rule, and then the local features of each image patch are represented by low-level feature descriptors. Accordingly, each scene is described by a set of patch descriptors, and training data sets are constructed by different patch descriptors for all scenes. Finally, M clusters are generated from the local feature representations, and their cluster centroids constitute a vocabulary of visual words. This visual dictionary can be used to quantize the extracted features by simply assigning the label of the closest visual word in the feature space. Mathematically, the final representation of an image is the frequency or histogram of the visual words in the vocabulary:

这一段在介绍BoVW算法

$$\text{BoVW} = [v_1, v_2, \ldots, v_M], \tag{3}$$

where v_m represents the number of occurrences of visual word m in the image and M is the size of visual vocabulary. To measure the similarity of two images with BoVW representations, the intersection kernel is used:

为了测量两张图片的相似度采用以下的公式Kbovw来表示

$$K_{\text{BoVW}}\left(\text{BoVW}^1, \text{BoVW}^2\right) = \sum_{m=1}^{M} \min\left(\text{BoVW}^1(m), \text{BoVW}^2(m)\right), \tag{4}$$

where BoVW^1 and BoVW^2 correspond to the BoVW representations of two different images. The intersection kernel can be used in kernel-based learning algorithms for classification, such as non-linear support vector machines.

2.2.2. *The MBVW representation*

Unlike the traditional BoVW, the MBVW explores the visual word vocabulary with multi-scale deep feature representations. First, multi-scale deep features are extracted by the CNN framework, and then multi-scale visual words are constructed by clustering the

extracted features. Finally, the co-occurrence word indices at different scales are calculated and used for scene classification.

More specifically, if an image is partitioned into a sequence of square image patches at the scale 0,..., N, there are $T = 4^s$ image patches at the scale s. Let P_1^s and P_2^s be the visual word histograms of two images at the scale of s, then the number of matches at scale s is computed as the histogram intersection $T(P_1^s, P_2^s)$: 根据（4）类比 此时的T 就相当于(4)中的K

$$T(P_1^s, P_2^s) = \sum_{m=1}^{M} \min(P_1^s(m), P_2^s(m)), \tag{5}$$

where $P_1^s(m)$ and $P_2^s(m)$ represent the counts of visual word m. The weight associated with level s is set to $\frac{1}{2^{N-s}}$ for penalizing the matches found in the larger scale. Finally, the co-occurrence index K_s for two images across all scales is given by S级上权重 设为1/2^n-s 以此来对惩罚较大规模的匹配

$$K_s = \sum_{s=0}^{N} \frac{1}{2^{N-s}} T(P_1^s, P_2^s). \tag{6}$$

3. Experimental results

A series of experiments have been conducted to access the performance of the proposed deep feature-based BoVW, and compared with other state-of-the-art results. In the experiment, we refer to the well-known UC Merced Land Use data set consisting of high-resolution aerial images with different scenes. Recently, numerous studies have used this data set; thus, an extensive comparison of the classification results between different algorithms is already available.

3.1. *Data set*

The UC Merced data set was released in 2010 (Yang and Newsam 2010), and consists of 21 classes of images selected from aerial orthoimagery with a pixel resolution of 0.3 m. For each class, 100 images with the size of 256 × 256 pixels are selected and manually labelled. Generally, the classes in the UC Merced data set include agriculture area, airport, baseball diamond, beach, buildings, chaparral, dense residential areas, forest, freeway, golf course, harbour, intersection, medium residential areas, mobile home park, overpass, parking lot, river, runway, sparse residential areas, storage tanks, and tennis court. The representatives of each class are shown in Figure 3. Owing to the complex patterns in the land-use scenes, low-level features are inefficient for visual word description. However, the CNN can effectively find complex and representative features to describe complex visual words in different land-use scenes.

3.2. *Experiments and comparison*

During experiments, 80 images for each class were randomly selected and marked as labelled images, whereas the remaining images were marked as the unlabelled ones. To test the robustness of the deeply described visual words, instead of using uniform grid

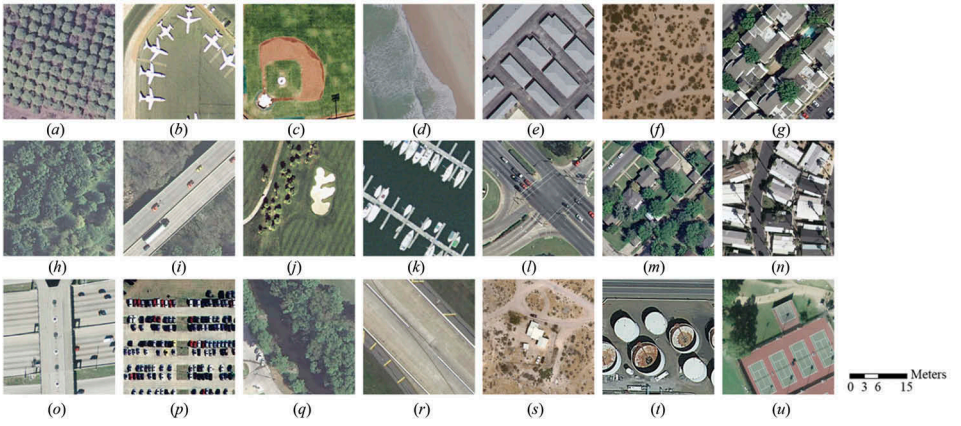

Figure 3. Sample images in the UC Merced data set. (a) agricultural area; (b) airport; (c) baseball diamond; (d) beach; (e) buildings; (f) chaparral; (g) dense residential area; (h) forest; (i) freeway; (j) golf course; (k) harbour; (l) intersection; (m) medium residential area; (n) mobile home park; (o) overpass; (p) parking lot; (q) river; (r) runway; (s) sparse residential area; (t) storage tanks; and (u) tennis court.

sampling, for each labelled image, 50 training samples (image patches) with the fixed size of 28×28 pixels were randomly chosen from different image scales. Training samples were taken as the input data for deep feature learning and visual word assignment. Additionally, overall accuracy (OA) was used to evaluate the classification performances. Fivefold stratified cross-validation was used for all experiments, where four folds are used for training and model selection and the remaining unseen fold is classified to measure accuracy. 所有实验均使用五重分层交叉验证，其中四个折叠用于训练和模型选择，剩余的未见折叠被分类为测量准确度。

3.2.1. *Parameter selection in CNN*

The parameter of the CNN architecture given in this section was heuristically selected based on experimental investigation. Before deep feature extraction, image patches at different scales are transformed into the size of 28×28 pixels for fitting the input size of our CNN framework. At the first convolutional layer, the input is converted to $24 \times 24 \times 50$ pixels (units are the same hereinafter) with 50 filters of 5×5, before being subsampled with a 2×2 max-pool layer to obtain a $12 \times 12 \times 50$ output. Then, the second convolutional layer includes 50 filters with size of 5×5 to generate an $8 \times 8 \times 50$ output, which then subsampled to $4 \times 4 \times 50$ with the max-pool operator. Finally, a fully connected dense layer with 256 hidden units is used before resolving to 21 different labels in a soft-max output layer. In the training stage, the learning rate was set to 0.005 with L2 normalization, and the mini-batch size was set to 150. To avoid over-fitting, dropout strategy was also implemented in the CNN framework. During the prediction stage of CNN, for each image patch, a 256-dimension feature vector was extracted by the fully connected layer and then used for visual word assignment.

3.2.2. *Analysis of MBVW*

3.2.2.1. *Number of scales.* Objects in the VHR images show great variations in sizes and shapes. Therefore, observation scale is critical to effectively represent complex scenes. In this study, multi-scale image patches are used in the MBVW framework in order to capture the spatial features at different scales. To quantitatively evaluate the effects of the chosen scales on the MBVW-based classification accuracy, five scales (images were resized to 28×28 at scale five) were used. In these experiments, the number of training sets varies from 10 to 80 and the size of vocabulary was 500. Figure 4 (*a*) reports the results of the MBVW considering the different scales of training samples. Generally, higher classification accuracies can be achieved when more training samples and visual word scales are used. However, classification accuracies decrease dramatically if more than three scales are considered. The reason is that the pyramid images with more than three scales are too blurred to differentiate different scenes. Accordingly, in this study, pyramid images with three scales were adopted.

本文采用三个尺度的金字塔图像。原因是 太深容易产生模糊，影响分类效果

3.2.2.2. *Visual word vocabulary construction.* For each training sample, a feature vector with the size of $4 \times 4 \times 50$ can be obtained with the help of CNN. In addition, there are 80×50 training samples for each class. Accordingly, during the training stage, we can obtain 4000 feature representations for each image class. Then, the *k*-means clustering is used to construct the visual word vocabulary. Each clustering centre represents a visual word. Accordingly, the number of clusters potentially impacts the expressive power of MBVW. To measure the performance of MBVW with different numbers of clusters (visual words), the parameter of *k*-means was set to 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, and 600, respectively. To illustrate the effectiveness of multi-scale deep features in visual word construction, the traditional features for visual word construction such as SIFT and image colour information (denoted as 'raw' hereinafter) are also included for comparison. With different kinds of image feature representations, scene classification results using different strategies are reported in Figure 4(*b*). Generally, the classification accuracy increases as the size of visual dictionary becomes bigger; however, it decreases greatly when it reaches the highest point at 500 for both

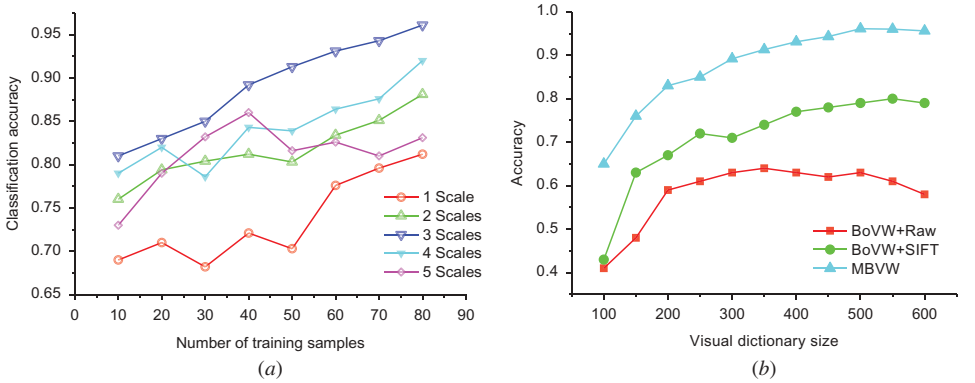

Figure 4. (a) Classification accuracy of MBVW with different numbers of training data sets at five scales. (b) Experimental results of using three different scene classification strategies with various sizes of visual dictionary: BoVW + Raw; BoVW + SIFT; and MBVW.

methods. This indicates that the MBVW-based classification method obtains 65% OA with just 100 visual words, which is much larger than the traditional ones (around 42%). Therefore, this further illustrates the effectiveness of deep feature representations. Moreover, MBVW outperforms the SIFT and Raw feature-based scene classification strategies (BoVW+SIFT and BoVW+Raw) in terms of classification accuracies. According to the analysis of visual word vocabulary, we set the size of vocabulary to 500 during our experiments.

3.2.2.3. *Comparison with existing methods.* To illustrate the effectiveness of the MBVW, three-misclassified image examples are shown in Figure 5. The appearances of these misclassified images are extremely confusing with a similar appearance of different semantic classes. We can conclude that it is difficult to accurately interpret the misclassified images even with the help of experienced users. Figure 6 provides the confusion matrix for MBVW-based scene classification, where the entry in the *i*th row and the *j*th column denotes the rate of test images of the *i*th class that is wrongly

Figure 5. Illustration of mis-classification areas. (a) Baseball diamond classified into golf course. (b) Dense residential classified into medium residential. (c) Sparse residential classified into forest.

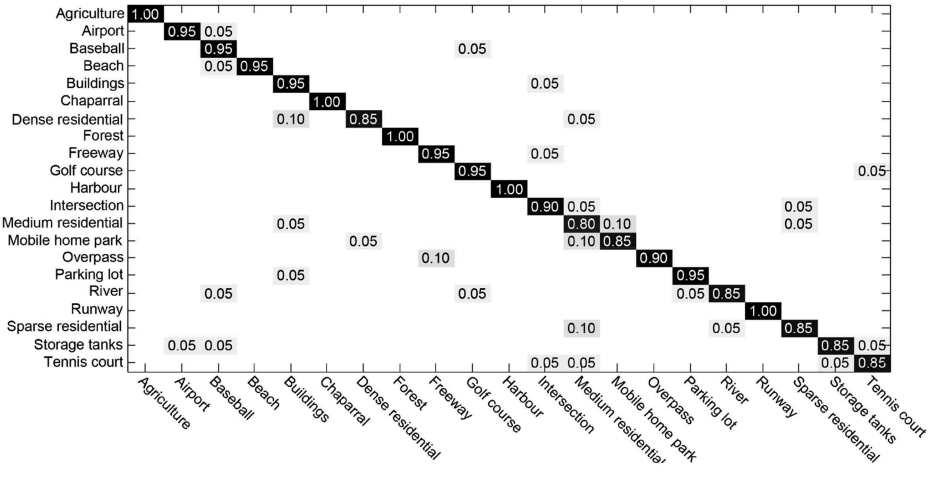

Figure 6. Confusion matrix showing classification performance on the UC Merced data set for the proposed method. The rows and columns of the matrix denote actual and predicted classes, respectively.

Table 1. Comparison between the accuracy of the proposed method and the previously used methods for the UC Merced data set.

Method	BoVW	SPMK*	SPCK++*	UFL*	UFL-SC*	MBVW	GoogLeNet*	OverFeat*
Accuracy (%)	71.86	75.29	77.38	81.67	90.26	96.14	97.10	99.43

*Results are directly taken from Penatti, Nogueira, and Dos Santos (2015).

classified as the *j*th class. In this figure, 12 among the 21 UC Merced land-cover classes have classification accuracies larger than 90%. Especially, for the agricultural, chaparral, forest, harbour, and runway classes, their classification accuracies are 98%, which are much larger than the traditional ones.

Table 1 reports the average classification accuracy of all of the 21 classes of our MBVW and some state-of-the-art approaches such as GoogLeNet and Overfeat-based deep networks. Although the MBVW performs worse than the well-designed deep learning frameworks in terms of classification accuracy, it provides a new path to link the mid-level deeply described visual elements and the content of scenes with relatively low computational complexity. Therefore, to illustrate the effectiveness of our MBVW, we mainly focus on the comparison of the classification performance of MBVW with the SIFT-based BoVW and its extension forms such as SPMK (Lazebnik, Schmid, and Ponce 2006) and SPCK++ (Yi and Newsam 2011). Furthermore, unsupervised feature learning methods (UFLs) (Cheriyadat 2014) and their spectral clustering version UFL-SC (Hu, Xia, Wang, et al. 2015) were also involved in the comparison. The comparison shows that our method can improve the accuracy with UFL-SC by 5.88%. The comparison results indicate that the MBVW-based scene classification method using deeply described visual words has great potential for image content retrieval. However, there is a significant accuracy gap between the MBVW and the well-designed deep learning methods, such as GoogLeNet and Overfeat. The reason for this phenomenon probably is that the MBVW explores deep feature representations on local image patches, which overlooks contextual information of the entire image scene. However, the MBVW uses multi-scale strategy to cover different spatial layouts on image scenes and for efficient classification with much fewer computational costs compared with the previous methods.

尽管跟CNN框架比，不如其精度但是MBVW却连接了中级视觉层和较低级视觉层。表现仍然很好

4. Conclusions

In contrast to the previous works focusing on exploring low-level features for scene classification, we introduced the deep learning framework to automatically extract robust features at high levels for high-resolution scenes description. The proposed method involves deep feature extraction, visual words construction, and scene classification. Rather than using the low-level feature for visual words construction, we derived deep features through a well-designed CNN framework. The visual word vocabulary was constructed by applying the *k*-mean clustering strategy on deep features. The experimental results show that deep feature-based MBVW outperforms the existing methods in classification accuracy.

As future extension, we plan to improve this method in terms of spatial extensions. The current MBVW-based scene classification method overlooks the spatial arrangements of visual words. However, it will helpful to consider the spatial distribution of the visual words, especially for high-resolution scenes.

当前基于MBVW的场景分类方法忽略了视觉词的空间排列。

但是考虑到空间分布对于视觉词很有帮助，尤其是对高分辨率场景

Acknowledgements

The work presented in this article was supported by the Weng Hongwu Scientific Research Foundation of Peking University, China (No. WHW201505).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Weng Hongwu Scientific Research Foundation of Peking University, China [No. WHW201505].

References

Bellens, R., L. Gautama Sidharta, M.-F. W. Philips, J. C. Chan, and F. Canters. 2008. "Improved Classification of VHR Images of Urban Areas Using Directional Morphological Profiles." *IEEE Transactions on Geoscience and Remote Sensing* 46 (10): 2803–2813. doi:10.1109/TGRS.2008.2000628.

Blaschke, T. 2010. "Object Based Image Analysis for Remote Sensing." *ISPRS Journal of Photogrammetry and Remote Sensing* 65 (1): 2–16. doi:10.1016/j.isprsjprs.2009.06.004.

Boutell, M. R., J. Luo, X. Shen, and C. M. Brown. 2004. "Learning Multi-Label Scene Classification." *Pattern Recognition* 37 (9): 1757–1771. doi:10.1016/j.patcog.2004.03.009.

Bruzzone, L., and L. Carlin. 2006. "A Multilevel Context-Based System for Classification of Very High Spatial Resolution Images." *IEEE Transactions on Geoscience and Remote Sensing* 44 (9): 2587–2600. doi:10.1109/TGRS.2006.875360.

Cheriyadat, A. M. 2014. "Unsupervised Feature Learning for Aerial Scene Classification." *IEEE Transactions on Geoscience and Remote Sensing* 52 (1): 439–451. doi:10.1109/TGRS.2013.2241444.

Dalal, N., and B. Triggs. 2005. "Histograms of Oriented Gradients for Human Detection". Paper presented at the Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, San Diego, CA, June 20–June 25.

Felzenszwalb, P. F., R. B. Girshick, D. McAllester, and D. Ramanan. 2010. "Object Detection with Discriminatively Trained Part-Based Models." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (9): 1627–1645. doi:10.1109/TPAMI.2009.167.

Fernando, B., E. Fromont, and T. Tuytelaars. 2014. "Mining Mid-level Features for Image Classification." *International Journal of Computer Vision* 108 (3): 186–203. doi:10.1007/s11263-014-0700-1.

Fukushima, K. 1980. "Neocognitron: A Self-Organizing Neural Network Model for A Mechanism of Pattern Recognition Unaffected by Shift in Position." *Biological Cybernetics* 36 (4): 193–202. doi:10.1007/BF00344251.

Hinton, G. E., S. Osindero, and Y.-W. Teh. 2006. "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation* 18 (7): 1527–1554. doi:10.1162/neco.2006.18.7.1527.

Hu, F., G.-S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun. 2015. "Unsupervised Feature Learning Via Spectral Clustering of Multidimensional Patches for Remotely Sensed Scene Classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8 (5): 2015–2030. doi:10.1109/JSTARS.2015.2444405.

Hu, F., G.-S. Xia, J. Hu, and L. Zhang. 2015. "Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery." *Remote Sensing* 7 (11): 14680–14707. doi:10.3390/rs71114680.

Hubel, D. H., and T. N. Wiesel. 1970. "The Period of Susceptibility to the Physiological Effects of Unilateral Eye Closure in Kittens." *The Journal of Physiology* 206 (2): 419–436. doi:10.1113/jphysiol.1970.sp009022.

Jia, S., H. Liu, and F. Sun. 2015. "Aerial Scene Classification with Convolutional Neural Networks." In *Advances in Neural Networks – ISNN 2015*, edited by H. Xiaolin, Y. Xia, Y. Zhang, and D. Zhao, 258–265. Cham: Springer International Publishing.

Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. "Caffe: Convolutional Architecture for Fast Feature Embedding." In *Proceedings of the ACM International Conference on Multimedia*, November 03–07, 675–678. Orlando, FL: ACM.

Lazebnik, S., C. Schmid, and J. Ponce. 2006. "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories". Paper presented at the Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, New York, June 17–June 22.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation* 1 (4): 541–551. doi:10.1162/neco.1989.1.4.541.

Lee, H., R. Grosse, R. Ranganath, and A. Y. Ng. 2009. "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations." In *Proceedings of the 26th Annual International Conference on Machine Learning*, June 14–18, 609–616. Montreal, QC: ACM.

Li, L-J, H. Su, L. Fei-Fei, and E. P. Xing. 2010. "Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification". Paper presented at the Advances in neural information processing systems Vancouver, BC, December 6–9.

Lowe, D. G. 2004. "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision* 60 (2): 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

Luo, J., M. Boutell, R. T. Gray, and C. Brown. 2005. "Image Transform Bootstrapping and its Applications to Semantic Scene Classification." *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)* 35 (3): 563–570. doi:10.1109/TSMCB.2005.846677.

Myint, S. W., P. Gober, A. Brazel, S. Grossman-Clarke, and Q. Weng. 2011. "Per-Pixel Vs. Object-Based Classification of Urban Land Cover Extraction Using High Spatial Resolution Imagery." *Remote Sensing of Environment* 115 (5): 1145–1161. doi:10.1016/j.rse.2010.12.017.

Penatti, O. A. B., K. Nogueira, and J. A. dos Santos. 2015. "Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains?" Paper presented at the Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on, Boston, MA, June 7–12

Shackelford, A. K., and C. H. Davis. 2003. "A Combined Fuzzy Pixel-Based and Object-Based Approach for Classification of High-Resolution Multispectral Data over Urban Areas." *IEEE Transactions on Geoscience and Remote Sensing* 41 (10): 2354–2364. doi:10.1109/TGRS.2003.815972.

Yang, C., H. Liu, S. Wang, and S. Liao. 2016. "Scene-Level Geographic Image Classification Based on a Covariance Descriptor Using Supervised Collaborative Kernel Coding." *Sensors* 16 (3): 392. doi:10.3390/s16030392.

Yang, J., Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. 2007. "Evaluating Bag-Of-Visual-Words Representations in Scene Classification." In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, September 23–28, 197–206. Augsburg: ACM.

Yang, Y., and S. Newsam. 2010. "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification". Paper presented at the Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, November 3–5.

Yi, Y., and S. Newsam. 2011. "Spatial Pyramid Co-Occurrence for Image Classification". Paper presented at the Computer Vision (ICCV), 2011 IEEE International Conference on, Barcelona, November 6–13.

Zhao, W., and S. Du. 2016a. "Spectral–Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach." *IEEE Transactions on Geoscience and Remote Sensing* 54 (8): 4544–4554. doi:10.1109/TGRS.2016.2543748.

Zhao, W., and S. Du. 2016b. "Learning Multiscale and Deep Representations for Classifying Remotely Sensed Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 113: 155–165. doi:10.1016/j.isprsjprs.2016.01.004.

Zhao, W., Z. Guo, J. Yue, X. Zhang, and L. Luo. 2015. "On Combining Multiscale Deep Learning Features for the Classification of Hyperspectral Remote Sensing Imagery." *International Journal of Remote Sensing* 36 (13): 3368–3379. doi:10.1080/2150704X.2015.1062157.

Zhou, L., Z. Zhou, and D. Hu. 2013. "Scene Classification Using a Multi-Resolution Bag-Of-Features Model." *Pattern Recognition* 46 (1): 424–433. doi:10.1016/j.patcog.2012.07.017.