ICIC Express Letters
Part B: Applications
Volume **7**, Number **9**, September **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **2045–2053**

UNSUPERVISED DEEP SPATIAL FEATURE ENCODING FOR AERIAL IMAGE CLASSIFICATION

Tao Shi, Chunlei Zhang, Hongge Ren, Fujin Li and Weimin Liu

College of Electrical Engineering
North China University of Science and Technology
No. 46, Xinhua Road, Tangshan 063009, P. R. China
randyray@126.com

Received March 2016; accepted June 2016

Abstract. *Due to the rapid development of satellite sensor technology, a rich aerial image data set can be easily acquired. How to efficiently classify and recognize the aerial image has become a critical task. In this paper, we propose an aerial image classification method based on local variance similarity sparse coding (LVSSC) and deep belief network (DBN). Low-level features are extracted by using scale-invariant feature transform (SIFT). These extracted features are encoded in terms of an improved sparse encoding mode by combining local variance similarity and sparse coding to generate new sparse representation. DBN is used to express the relationship between low-level features and high-level semantic representations and complete image classification. We apply our method to OT data set and UC Merced data set. Experimental results show that our method efficiently utilizes spatial information of images and improves the classification performance.*

Keywords: Aerial image classification, Local variance similarity, Sparse coding, Deep belief network, Feature extraction

1. **Introduction.** At present, the research on high resolution aerial image classification has been rapidly developed. Bruzzone and Carlin proposed an algorithm based on pixel level feature for scene classification [1]. However, the classification results are greatly influenced by the segmentation algorithm. Yang and Newsam have computed the co-occurrence of the visual words and combined this with the bag-of-visual-words (BoVW) method, and they reported higher classification accuracy than the traditional BoVW and the spatial pyramid matching kernel (SPMK) for their extended spatial co-occurrence kernel (SPCK++) method [2]. However, due to the aerial image character of high resolution and large data, we need to find a better classification method.

With the development of the sparse coding (SC) [3], the image representation method has changed greatly. Sparse coding simulates the activity of the neuron's sparse type. Lee et al. proposed an unsupervised learning model with 9 layers of sparse coding, which can effectively detect face from unlabeled images [4]. In this paper, we propose a novel sparse coding method by combining local variance similarity (LVS) and sparse coding to encode aerial image features [5].

In recent years, deep learning has been widely used in various fields of machine vision. Deep learning network has a hierarchical structure, which can effectively learn the features from a large number of input data. Lu et al. proposed a remote sensing image classification method based on DBN model [6] which can outperform support vector machine (SVM) and traditional neural network (NN) [7]. This paper takes DBN as an important tool for aerial image classification.

According to several characteristics of aerial images, this paper also proposes a classification method of aerial image based on sparse representation and deep belief network.

The algorithm flow chart is shown in Figure 1. First, we use SIFT features as aerial image feature descriptor for feature extraction and then the improved sparse representation of the extracted features is exploited to generate new sparse representation. Finally we put sparse feature into DBN which is used to express the relationship between low-level features and high-level semantic representations and complete image classification. We apply our model to OT data set and UC Merced data set and compare it with BoVW [2], SPMK [8] and SC+SVM [9]. Our method obtained results that were equal to or even better than the previous results with the OT data set and the UC Merced data set.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the SIFT feature extraction method and describe the unsupervised spatial feature encoding approach in detail. Section 3 describes the DBN model and the specific classification method on aerial images. The details of our experiments and the results are presented in Section 4. Finally, Section 5 concludes this paper and our ideas for future work.

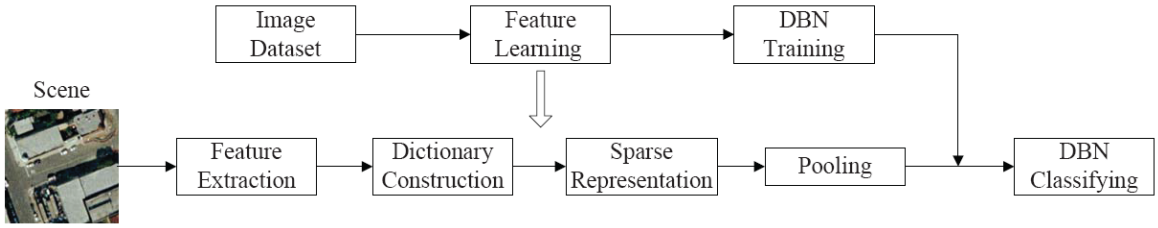

FIGURE 1. Framework of the proposed aerial image classification

2. Low-Level Feature Extraction.

2.1. SIFT feature extraction.

SIFT is the most widely used feature in the field of computer vision and also recognized as one of the best features. The feature extraction steps include feature detection and feature descriptor generation.

In the feature point detection: DOG equation is usually used to determine feature point detection, as shown in Formula (1):

$$D(x, y, \delta) = (G(x, y, k\delta) - G(x, y, \delta)) \cdot I(x, y) = L(x, y, k\delta) - L(x, y, \delta) \tag{1}$$

where k is a constant, $G(x, y, \delta) = \frac{1}{2\pi\delta^2} e^{-(x^2+y^2)/2\delta^2}$.

In the feature descriptor generation: first, determine the rectangular area of $R * R$ with each feature point as the center and calculate the gradient of each pixel. Then divide the region into $N * N$ sub regions and calculate the gradient of each feature point of the aerial image and form statistics of histogram. At this stage, the input image is represented as set of low-level feature vectors $X = [x_1, x_2, \ldots, x_n]$, where n is the number of samples. We set $n = 10000$ for all the experiments described in the later section.

2.2. Sparse representation.

Sparse representation is to represent images with a minimum of coefficients, which can provide a simple representation of redundant information and is propitious to extract the most essential feature of the image for DBN.

In this stage, we plan to find a set of basis functions and sparse weights that can be used to reproduce the original feature matrix X with least reconstruction error.

To construct a basis function (dictionary), first, we randomly sample low-level features from the entire data set to generate matrix $X = [x_1, x_2, \ldots, x_n]$. Next, given the feature matrix X, we learn the basis functions by finding best solution for a minimization problem which is similar to the sparse coding framework. The basis function D is learned using alternate minimization of Formula (2):

$$\min_{D, s_i} \sum_i \|Ds_i - x_i\|_2^2, \text{ subject to } \|D_j\|_2 = 1, \ \forall j \text{ and } \|s_i\|_0 \le k, \ \forall i \tag{2}$$

where $\|s_i\|_0$ is the number of nonzero elements in column vector s_i.

After learning the basis function D, we are going to represent the feature matrix X sparsely. According to sparse coding theory, the feature matrix X can be represented as a sparse linear combination of atoms in D. The process of feature encoding can be expressed as the following optimization problem:

$$f_i = \sum_i (x_i - Ds_i)^2 + \lambda \sum_i \theta(s_i) \tag{3}$$

where $\theta(s_i)$ is sparse penalty function. However, the formula uses error square sum as the standard for evaluating the similarity between sparse features and input features, which ignores the strong correlation between image features and the importance of structure and details for image. To get better sparse feature, we introduce the concept of LVS.

In recent years, people have a new breakthrough in the study of human eye vision system. The human eye is more sensitive to the high frequency part of the image which is related to the details of the image. The local variance of the image well reflects the spatial details of the image, so that the details of the image can be analyzed through the variation of the local variance of the image [10]. We can also consider that the distribution of the local variance of the image contains a large amount of structural information of the image [11]. The LVS between reconstruction feature Y and original feature X can be defined as Formula (4):

$$LVS = \frac{2\mu_x\mu_y}{\mu_x^2 + \mu_y^2} \cdot \frac{2\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} \tag{4}$$

where μ_x and μ_y are the mean value, σ_x and σ_y are the standard deviation.

For the convenience of description, we divide the original image blocks into n dimensional column vector I. I_i $(i = 1, \ldots, N)$ represents each pixel point. ϕ_k is each of the N vectors. $\phi_{i,j}$ is the element in basis function matrix A. The reconstructed image block is represented by Y_i $(i = 1, \ldots, N)$.

Combining LVS and sparse coding, we can get the new optimization criterion:

$$z_i = \lambda_1 \sum_{i=1}^{N} (I_i - Y_i)^2 + \lambda_2(1 - LVS(I, Y)) + \lambda_3 \sum_{i=1}^{M} \theta(s_i) \tag{5}$$

where λ_1, λ_2 and λ_3 are weight coefficients.

Combining Formulas (4) and (5), we can get:

$$z_i = \lambda_1 \sum_{i=1}^{N} (I_i - Y_i)^2 + \lambda_2 \left(1 - \frac{2\mu_I\mu_Y}{\mu_I^2 + \mu_Y^2} \cdot \frac{2\sigma_I\sigma_Y}{\sigma_I^2 + \sigma_Y^2}\right) + \lambda_3 \sum_{i=1}^{M} \theta(s_i) \tag{6}$$

We use alternative optimization method to solve Formula (6), which is to solve one variable by fixing another variable. We define

$$Q_1 = \sum_{i=1}^{N} (I_i - Y_i)^2, \quad Q_{21} = 2\mu_I\mu_Y, \quad Q_{22} = 2\sigma_I\sigma_Y \tag{7}$$

$$Q_{23} = \mu_I^2 + \mu_Y^2, \quad Q_{24} = \sigma_I^2 + \sigma_Y^2, \quad Q_3 = \sum_{i=1}^{M} \theta(s_i) \tag{8}$$

Step 1. Fixing A, use conjugate gradient method to solve S.

$$\nabla_{\alpha_i} z_i = \lambda_1 \nabla_{\alpha_i} Q_1 - \lambda_2 \frac{Q_{21} \cdot Q_{22}}{Q_{23} \cdot Q_{24}} \cdot \left(\frac{\nabla_{\alpha_i} Q_{21}}{Q_{21}} + \frac{\nabla_{\alpha_i} Q_{22}}{Q_{22}} - \frac{\nabla_{\alpha_i} Q_{23}}{Q_{23}} - \frac{\nabla_{\alpha_i} Q_{24}}{Q_{24}}\right) + \lambda_3 \nabla_{\alpha_i} Q_3 \tag{9}$$

where

$$\nabla_{\alpha_i} Q_1 = -2 \sum_{k=1}^{N} (I_k - Y_k)\, \phi_{k,i},\ \nabla_{\alpha_i} Q_{21} = \tfrac{2}{N} \mu_I \sum_{k=1}^{N} \phi_{k,i},\ \nabla_{\alpha_i} Q_{22} = \tfrac{2}{N-1} \sum_{k=1}^{N} ((I_k - \mu_I)\phi_{k,i}) \tag{10}$$

$$\nabla_{\alpha_i} Q_{23} = \frac{2}{N}\mu_Y \sum_{k=1}^{N} \phi_{k,i},\ \nabla_{\alpha_i} Q_{24} = \frac{2}{N-1} \sum_{k=1}^{N} ((Y_k - \mu_Y)\phi_{k,i}) \tag{11}$$

Step 2. Fixing S, use simple gradient method to solve A.

$$\nabla_{\phi_{i,j}} z_i = \lambda_1 \nabla_{\phi_{i,j}} Q_1 - \lambda_2 \frac{Q_{21} \cdot Q_{22}}{Q_{23} \cdot Q_{24}} \cdot \left(\frac{\nabla_{\phi_{i,j}} Q_{21}}{Q_{21}} + \frac{\nabla_{\phi_{i,j}} Q_{22}}{Q_{22}} - \frac{\nabla_{\phi_{i,j}} Q_{23}}{Q_{23}} - \frac{\nabla_{\phi_{i,j}} Q_{24}}{Q_{24}} \right) \tag{12}$$

where

$$\nabla_{\phi_{i,j}} Q_1 = -2\,(I_i - Y_i)\,\alpha_j,\ \nabla_{\phi_{i,j}} Q_{21} = \frac{2}{N}\mu_I \alpha_j \tag{13}$$

$$\nabla_{\phi_{i,j}} Q_{22} = \frac{2}{N-1}\,(I_k - \mu_I)\,\alpha_j,\ \nabla_{\phi_{i,j}} Q_{23} = \frac{2}{N}\mu_Y \alpha_j,\ \nabla_{\phi_{i,j}} Q_{24} = \frac{2}{N-1}(Y_i - \mu_Y)\alpha_j \tag{14}$$

After sparse representation, the new feature representation for an image scene will usually have a very high dimensionality. For computational efficiency and storage volume, it is a standard practice to use a pooling strategy to reduce the dimensionality of the image representation. With the sparse features z_i computed for an image patch, we can estimate the final feature representation as follows:

$$m = \frac{1}{N} \sum_{i=1}^{N} z_i \tag{15}$$

At this point, we get the sparse feature representation of aerial image m. m is used as the input vector of DBN to complete the high-level image classification.

3. **High-Level Image Classification.** Deep learning network has a hierarchical structure, which can automatically learn high-level features from low-level features. This paper uses deep belief networks to represent the relationship between low-level features and high-level semantic representations of aerial image and then complete image classification.

In 2006, Hinton et al proposed a model of DBN and successfully applied it to the recognition of handwritten font [12]. The basic structure of DBN is restricted boltzmann machine (RBM), as shown in Figure 2.

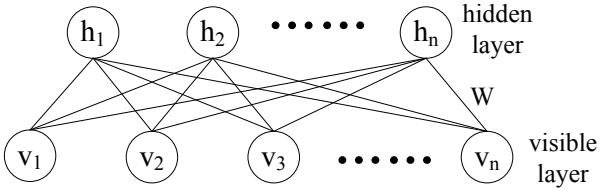

FIGURE 2. RBM model

RBM is a type of two layer neural networks comprised of a visible layer that represents the observed data and a hidden layer that represents the hidden variables. Connections only exist between the visible layer and the hidden layer.

RBM is an energy based model, and its energy function is defined in Formula (16):

$$E(v,h) = -\sum_{i=1}^{I}\sum_{j=1}^{J} v_i h_j w_{ij} - \sum_{i=1}^{I} a_i v_i - \sum_{j=1}^{J} b_j h_j \tag{16}$$

where w is weight matrix, b are visible unit biases and a are hidden unit biases.

Based on the energy function, the definition of the joint distribution is in Formula (17):

$$P(v,h) = \frac{\exp(-E(v,h))}{Z} \tag{17}$$

where Z is called the partition function and $Z = \sum_v \sum_h \exp(-E(v,h))$.

Conditional probability distributions are as follows:

$$p(h_j = 1 \,|v) = \delta\left(b_j + \sum_{i=1}^{I} v_i w_{ji}\right), \quad p(v_i = 1 \,|h) = \delta\left(a_i + \sum_{j=1}^{J} h_j w_{ji}\right) \tag{18}$$

where $\delta(x)$ is sigmoid function.

Each two hidden layers constitute an RBM network and the top layer is back propagation (BP) neural network, as shown in Figure 3.

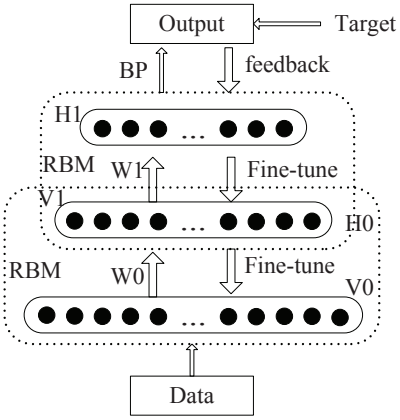

FIGURE 3. DBN structure

The lower layer of RBM extracts and abstracts the input data and puts it as the high layer input. The DBN is trained by the combination of pre-training and fine tuning. First, the unsupervised training of each layer of RBM is carried out in a bottom-up way and then the supervised BP neural network is used to fine tuning the whole model in a top-down way.

After training DBN, we put the sparse feature of the image m as input and the category of the image as output to train a new network structure, which is suitable to classify aerial images.

4. **Experiment.**

4.1. **Experimental setup.** This paper uses dense-SIFT algorithm [13]. First, divide the image into image blocks with the size of 16×16 pixels and interval of 8 pixels. Then divide the image block into 4×4 sub regions and calculate the gradient histogram of 8 directions in each sub region as the seed point. Finally, the seed points are connected to the 128 dimensional feature vector. We set the size of the basis functions to 1024.

We validate the aerial scene classification on two data sets, OT data set and UC Merced data set. With OT data set and UC Merced data set, we randomly select 80 samples from each class to initialize the training set and the remaining 20 samples as the testing set. We repeated 10 classification experiments on each data set and the classification accuracy of the 10 experiments was averaged as the final classification accuracy.

4.2. **Experiment on OT data set.** OT data set contains 8 aerial scene categories: (1) Forest, (2) Mountain, (3) Open Country, (4) Coast, (5) Highway, (6) City, (7) Tall Building, and (8) Street. We select 100 images per class. Figure 4 shows some of the images in OT data set.

In order to study the sensitivity of the sparsity parameter, we varied their values over a wide range. Figure 5 shows the classification performance with different sparsity parameter values. The results showed that there was a wide range of sparseness values for which the classification performance was consistent, and the best classification performance was

FIGURE 4. Few example images from the OT data set

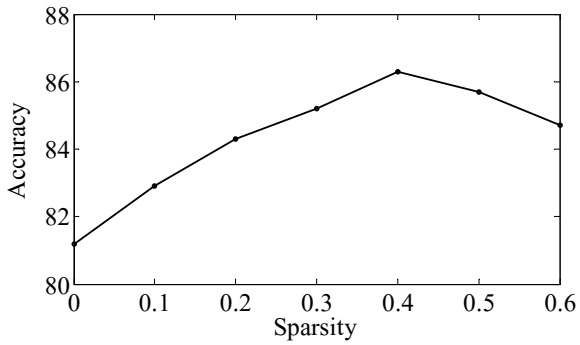

FIGURE 5. Effect of the sparsity parameter value on the classification accuracy

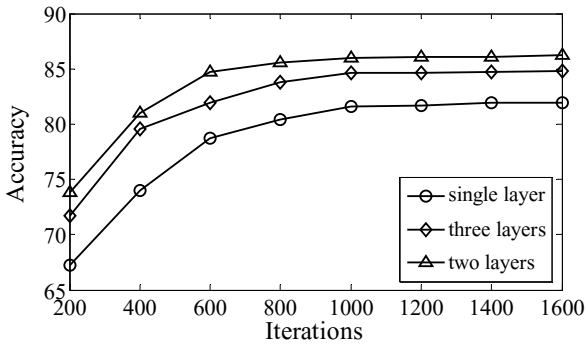

FIGURE 6. Effect of the layer of DBN and iterations of RBM on the classification accuracy

obtained at a sparsity value close to 0.4. Based on this analysis, we set the sparsity value as equal to 0.4 to generate sparse features.

To test the effects of different layers of DBN and different iterations of RBM on the classification performance, we verify the classification accuracy of three different layers of DBN as shown in Figure 6.

From Figure 6, in the initial stage, the classification accuracy is significantly improved with the increase of the number of iterations. When the number of iterations is greater than 1000, the classification accuracy is almost unchanged. So we set the number of iterations to 1000. Besides, the 2-layers DBN outperforms the single-layer DBN and 3-layers DBN. So we select 2-layers DBN.

To compare the scene classification performance of the spatial extension of BoVW reported in [2], the proposed method with SPMK [8] and the SC+SVM method described in [9], we measured the classification performance with the OT data set. Of the four strategies that we tested, our method produced better performance, as shown in Table 1. We compared the classification performances with and without the sparse representation.

TABLE 1. Comparison with the previous reported accuracies on OT data set

Methods	*BoVW*	*SPMK*	*SC + SVM*	*Non Sparsity*	*With Sparsity*
Accuracy	76.87%	79.12%	85.62%	84.73%	86.23%

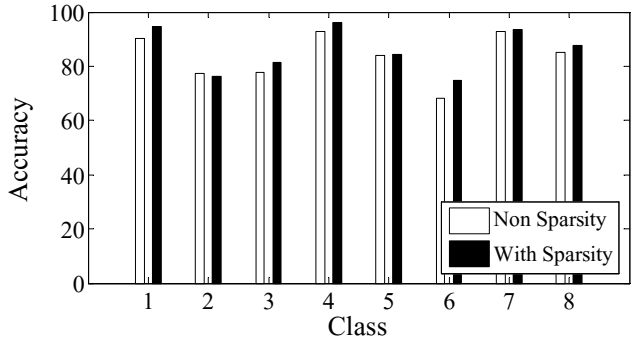

FIGURE 7. The overall accuracies with the OT data set for the proposed method

FIGURE 8. Few example images from the UC Merced data set

The results illustrate that using sparse representation is an efficient way to increase the scene classification accuracy. The reason is that sparse representation provides a simple representation of redundant information and represent features in a more concise and effective way. The overall accuracies of OT data set are reported in Figure 7.

4.3. **Experiment on UC Merced data set.** UC Merced data set consists of 256×256 color images from 21 aerial scene categories: (1) Agricultural, (2) Airplane, (3) Baseball diamond, (4) Beach, (5) Buildings, (6) Chaparral, (7) Dense residential, (8) Forest, (9) Freeway, (10) Golf course, (11) Harbor, (12) Intersection, (13) Medium residential, (14) Mobile home park, (15) Overpasses, (16) Parking lot, (17) River, (18) Runway, (19) Sparse residential, (20) Storage tanks, and (21) Tennis court. The data set contains highly overlapping classes and has 100 images per class. Figure 8 shows some of the images in UC Merced data set.

To test the classification performance of our proposed method in a larger data set, we measure the classification performance with the UC Merced data set. Compared with BoVW, SPMK and SC+SVM, the results are shown in Table 2. Our proposed method also has a better performance on a larger data set. We also compare the classification performance with and without the sparse representation to validate that sparse representation is a required step to characterize the scene effectively.

According to Table 1 and Table 2, we obtain that our method does not have obvious advantages compared with SC+SVM. The reason is the limited training data. There is not enough data in each category to train a DBN network which can well express the relationship between low-level features and high-level semantic representations of aerial

TABLE 2. Comparison with the previous reported accuracies on UC Merced data set

Methods	*BoVW*	*SPMK*	*SC + SVM*	*Non Sparsity*	*With Sparsity*
Accuracy	71.86%	74%	81.67%	81.15%	82.07%

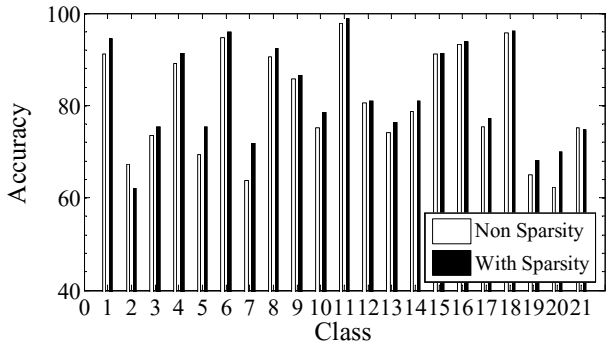

FIGURE 9. The overall accuracies with the UC Merced data set for the proposed method

image. With the increase of training data, deep network will have more obvious advantages than shallow network.

The overall accuracies are reported in Figure 9. The proposed method also shows the highest accuracy for the classification of the agricultural, chaparral, harbor, and runway scenes, which have a regular textural and spatial structure.

5. **Conclusions.** In this paper, we propose a classification method for high resolution aerial images based on sparse coding theory and deep belief network. First, we use dense-SIFT to extract the features of aerial image and then we use an improved sparse coding to represent the extracted features, which can represent the complex characteristics of aerial image more concise and effective. Finally, the relationship between low-level features and high-level semantic representations is represented by DBN. Experimental results on OT data set and UC Merced data set indicate that our method efficiently utilizes spatial information of images and obtains results that are equal to or even better than the previous results and deep learning can be effectively applied to the field of aerial image. Under the premise of meeting the amount of data, deep learning will have a greater advantage. As future extensions, we plan to apply this method to different scale data sets and further optimize feature extraction algorithm and sparse encoding mode.

Acknowledgment. This work is supported by National Natural Science Foundation (NNSF) of China under Grant 61203343.

REFERENCES

[1] L. Bruzzone and L. Carlin, A multilevel context-based system for classification of very high spatial resolution images, *IEEE Trans. Geoscience and Remote Sensing*, vol.44, no.9, pp.2587-2600, 2006.

[2] Y. Yang and S. Newsam, Spatial pyramid co-occurrence for image classification, *Proc. of IEEE International Conference on Computer Vision*, pp.1465-1472, 2011.

[3] B. A. Olshausen and D. J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature*, vol.381, no.6583, pp.607-609, 1996.

[4] Q. V. Lee, Building high-level features using large scale unsupervised learning, *Proc. of IEEE International Conference on Acoustics, Speech and Signal*, pp.8595-8598, 2013.

[5] Y. Wang, W. Liu and Y. Wang, Image quality assessment based on local variance and structure similarity, *Journal of Optoelectronics and Laser*, vol.19, no.11, pp.1546-1553, 2008.

[6] T. Kuremoto, S. Kimura and K. Kobayashi, Time series forecasting using a deep belief network with restricted boltzmann machines, *Neurocomputing*, vol.137, no.15, pp.47-56, 2014.

[7] Q. Lu, Y. Dou and X. Niu, Remote sensing image classification based on DBN model, *Journal of Computer Research and Development*, vol.51, no.9, pp.1911-1918, 2014.

[8] S. Lazebnik, C. Schmid and J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *IEEE Trans. Computer Vision and Pattern Recognition*, no.2, pp.2169-2178, 2006.

[9] A. M. Cheriyadat, Unsupervised feature learning for aerial scene classification, *IEEE Trans. Geoscience and Remote Sensing*, vol.52, no.1, pp.439-451, 2014.

[10] Y. Shi and Y. Ding, On the role of local variance in image fidelity assessment, *Proc. of IEEE International Conference on Signal Processing Systems*, pp.403-406, 2010.

[11] L. Dragut, C. Eisank and T. Strasser, Local variance for multi-scale analysis in geomorphometry, *Geomorphology*, vol.130, nos.3-4, pp.162-172, 2011.

[12] G. E. Hinton, S. Osindero and Y. W. Teh, A fast learning algorithm for deep belief nets, *Neural Computation*, vol.18, no.7, 2006.

[13] A. Bosch, A. Zisserman and X. Muoz, Image classification using random forests and ferns, *Proc. of IEEE International Conference on Computer Vision*, pp.1-8, 2007.