



Image classification method rationally utilizing spatial information of the image

Chao Wu¹ · Yaqian Li²  · Zhibiao Zhao¹ · Bin Liu¹

Received: 19 June 2018 / Revised: 2 December 2018 / Accepted: 22 January 2019 /

Published online: 8 February 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

In order to improve the accuracy of image classification problem, this paper proposes a new classification method based on image feature extraction and neural network. The method consists of two stages: image feature extraction and neural network classification. At the stage of feature extraction, spatial pyramid matching (SPM) feature, local position feature and global contour feature are extracted. The utilization of spatial information in SPM feature is effectively improved by combining the above three features. At the stage of neural network classification, a multi-hidden layer feedforward Histogram Intersection Kernel Weighted Learning Network (HWLN) is proposed to take advantage of three features to improve the classification accuracy. In the structure of network, the hidden layer output features are used as the bias of the input features to weight the coefficients of the features. The input layer and the hidden layers are directly connected with the output layer to realize the combination of linear mapping and nonlinear mapping. And the Histogram Intersection Kernel is used instead of random initialization of the input weight matrix. Taking combined features as input information, HWLN can realize the mapping relationship between input information and target category, so as to complete the image classification task. Extensive experiments are performed on Caltech 101, Caltech 256 and MSRC databases respectively. The experimental results show that the proposed method further utilizes the spatial information, and thus improves the accuracy of image classification.

Keywords Spatial pyramid matching · Local position feature · Global contour feature · Neural network · Histogram intersection kernel weighted learning network

✉ Yaqian Li
yaqianli@126.com

¹ Institute of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China

² Key Lab of Industrial Computer Control Engineering of Hebei Province, Yanshan University, Qinhuangdao 066004, China

1 Introduction

Image classification has always been a challenging problem in the field of computer vision because objects in images are affected by many factors, such as changes in lighting and viewing angles, large differences between objects, object deformation, occlusion, and background noisy, etc. With the development of the image classification field, many scholars have developed many effective classification methods [4, 7, 14, 20, 35, 36].

Classification methods usually consist of two essential stages: feature extraction and classification. Features should be relevant and distinguishable. It is quite reasonable that both feature and its spatial distribution are important for classification. Therefore, researchers attempted to fuse different features to improve the utilization of spatial information [23, 27, 42]. For example, in our recent work [27], local position features and global contour features are fused with Spatial Pyramid Matching (SPM) [24] features by multiplying respective weights to take advantage of the spatial distribution of the features. However, the weighting coefficients are manually set by experience. Therefore, a method that can automatically learn the weights is needed.

In classification stage, Support Vector Machines (SVM) is a popular classifier [7, 24, 26, 27, 37]. Limited by the structure of the SVM, Coefficient determined by experiment in a database could not be used directly in another database. Therefore, a large number of experiments are required on each database to verify the appropriate weight coefficients, which takes a lot of time to experiment. To solve this problem, we need a classifier that can perform automatic weighted fusion and classification on different features. Inspired by Squeeze-and-Excitation Network (SENet) [15], we use shallow neural network as classifier to achieve automatic weighting of the three features. In order to avoid iterative calculation of weights, we use Extreme Learning Machine (ELM) [16, 17] that trains the network quickly and efficiently without iteration. The fusion of the three features and classification are achieved by improving its structure.

In this paper, a complete image classification method, which includes two stages of feature extraction and classification, is proposed. At the stage of feature extraction, Local position features and global contour features are extracted to improve spatial information utilization of SPM features. At the stage of classification, we propose Histogram Intersection Kernel Weighted Learning Network (HWLN) to achieve feature combination and classification, which can be used to automatically and effectively fuse the above three features on different databases. In the structure of HWLN, the hidden layer output features are added to the input features to weight the coefficients of the input features. The input layer and the hidden layers are directly connected with the output layer to realize the combination of linear mapping and nonlinear mapping. And the Histogram Intersection Kernel (HIK) [13] is introduced between input layer and hidden layer instead of random initialization. Our method is tested on the three typical classification databases Caltech 101, Caltech 256 and MSRC. Compared with state-of-the-art methods, the proposed method obtains the higher accuracy for above classification databases.

The major contributions of this paper are: 1) HWLN is proposed based on concept of weighted fusion of features. The parameters of the network could be derived by performing the HIK kernel mapping and a solution matched with the HWLN structure, thus effectively solves the instability caused by the random initialization of parameters in ELM. 2) Feature extraction and HWLN are used to implement an image classification method, in which three features are fused and classified by benefiting from HWLN structure.

The main content of this paper is organized as follows. Related work is given in Section 2. The steps of feature extraction, the training process of HWLN and the steps of complete method are introduced in Section 3. Section 4 presents our experimental results on Caltech 101, Caltech 256 and MSRC databases. The paper concludes in section 5.

2 Related work

As a classic problem in computer vision, researchers have extensively studied image classification using various methods [5, 7, 10, 11, 18, 21, 33, 34]. By introducing frequency statistics to solve classification problem, the Bag of Keypoints (BoK) has become the most popular and effective image classification method. BoK proposed by Csurka et al. [7] uses the frequency statistics method to represent each image as an unordered collection of visual words. In order to solve the problem that BoK does not consider spatial information, researchers have proposed many methods [2, 3, 20, 22, 24, 41], such as Vector of Locally Aggregated Descriptors (VLAD) [20], SPM [24], BossaNova [3], Soft Pairwise Spatial Angle-Distance Histograms (SPSad) [22], etc. In order to solve the problem of losing information about the relative position, Zafar et al. [40] calculated the orthogonal vector of each point in the triplets of identical visual words, and the histogram is calculated based on the magnitude of these orthogonal vectors. The above methods all extract a single kind of feature on the image for classification.

The BoK framework is designed to extract a single kind of feature. The extracted features are limited by the single content that can be described, and cannot effectively solve the image classification problem. Many research efforts are focused on combining other types of features with BoK-based features to improve classification accuracy. Koniusz et al. [23] proposed the Higher-order Occurrence Pooling method, which obtains distinguishable features by aggregating co-occurrence relationship coefficients of codebook features. It solves the fusion problem of multiple local features, but it does not effectively utilize the spatial information contained in the image. Zou et al. [42] used multiscale completed local binary patterns (MS-CLBP) to extract the global features. BoK features and SPM features are extracted as local features. Global features and local features are encoded by kernel collaborative representation-based classification (KCRC) and the sum of weighted residuals is used to classify. However, the weighting coefficients of feature fusion need to be manually adjusted according to different databases. Xiong et al. [39] proposed a new pooling method to pool each sub-region into features. Then, features are weighted and concatenated to be classified using SVM. However, the weights are set empirically. Nilsback et al. [32] use color features, HOG features, SIFT features, and multi-kernel SVM for flower classification. The weighting coefficients of multi-kernel are obtained by optimization method. Ahmed et al. [1] use texture analysis with a uniform local binary pattern to extract local features, and use intensity-based detected local interest points and optimized sliding window to extract global features. These two features are reduced by the principal component analysis (PCA) before concatenation. Combined features are classified by SVM. Mansourian et al. [30] proposed Salient Dominant Color Descriptor and Pyramidal Histogram of Visual Words fusion model (SDCD and PHOW fusion model). In the model, color-based feature and PHOW-based feature are combined into a vector, which is classified by SVM. In the papers [1, 30], the features are not weighted when they are concatenated. The used SVM could not automatically train weighting coefficients based on different databases.

After features are extracted, neural network is utilized to perform classification. ELM is an extremely fast single hidden layer feed-forward neural network. Its main principle is that the output weights can be calculated by hidden layer weights and thresholds, which allows ELM to train the network quickly and efficiently without iteration. Li et al. [28] used ELM as a classifier, which gets better classification accuracy than SVM. Li et al. [25] improved network performance by combining linear and nonlinear mapping from input to output, this method takes full advantage of the features, but the randomly initialized hidden layer weights increase the instability of the network. Huang et al. [19] used the Karush-Kuhn-Tucker (KKT) principle to train ELM and proposed an extreme learning machine with kernel (KELM). Deng et al. [9] obtained the hidden layer output features by inputting the sample features and its subsets into the kernel function, which avoids random initialization and saves the calculation cost of the training process. Both [9, 19] avoid the instability caused by random initialization, but they cannot fully combine and utilize features.

3 Methodology

In order to utilize the spatial information in the image, an image classification method is proposed in this paper. This method can be divided into two stages: feature extraction and feature classification. The feature extraction stage is used to extract the SPM feature, local position feature and global contour feature. The feature classification stage uses HWLN to perform a series of weighted combinations of the three features and classify combined feature. The overall structure of the method is shown in Fig. 1. In the HWLN structure, the red solid lines between A layer and C layer represent the input features A are directly input into C layer. In the C layer, BW_b is added to A to approximate the effect of weighting. The red solid lines between B layer and D layer represent the same meaning. The blue dotted lines represent the output features of each layer are multiplied by the weight matrix and then input to the T layer. In the T layer, the multiplied output features are added together.

3.1 Feature extraction

In the feature extraction stage, SPM features, local position features and global contour features are extracted. We briefly introduce these three features. For detailed extraction steps, please refer to the paper [27].

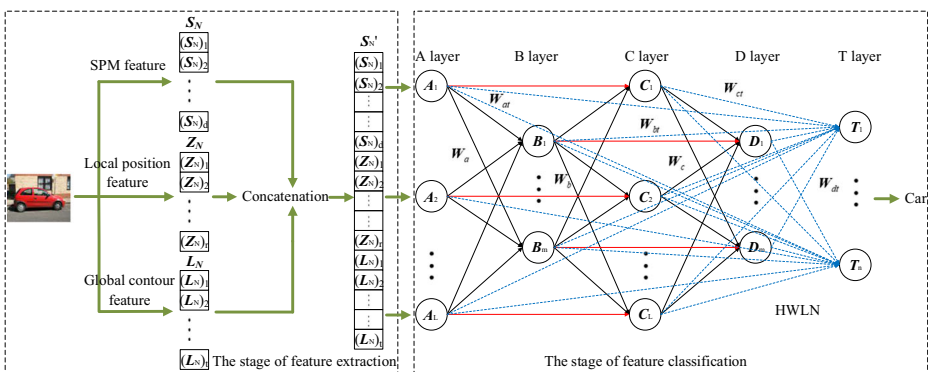


Fig. 1 The structure of object image classification method

The SPM method is described as below: (1) Dense-SIFT of each image are extracted. Dense-SIFT of training images are clustered into codebook. Localized Soft-assignment Coding (LSAC) [29] and codebook are used to encode local features of each image as visual words. (2) All the images are meshed according to different resolution levels, and the visual words in each mesh are pooled to generate a visual histogram. These histograms are multiplied by the weight coefficients of their respective resolutions and concatenated to generate SPM features.

As shown in Fig. 2, the coordinates of each Dense-SIFT feature in the image are extracted. The coordinates $w_j(j = 1, 2, \dots, m)$ of all Dense-SIFT features encoded by the dictionary vector $d_i(i = 1, 2, \dots, n)$ in one image are used to calculate the average coordinate $p_i(i = 1, 2, \dots, n)$. $w_j(j = 1, 2, \dots, m)$ and $p_i(i = 1, 2, \dots, n)$ are used to calculate the relative position distribution $z_i(i = 1, 2, \dots, n)$ of coordinates of Dense-SIFT feature around each codebook vector in the image. Local position feature is generated by concatenating $z_i(i = 1, 2, \dots, n)$.

The process of extracting the global contour feature is shown in Fig. 3. The Nonsubsampled Contourlet Transform (NSCT) [8] is used to extract the contour images, which are transformed into features $N_i^1(i = 1, 2, \dots, 7)$ by using Linear Discriminant Analysis (LDA) [38]. A series of feature encodings are performed on the low-dimensional features $N_i^1(i = 1, 2, \dots, 7)$ to generate the global contour feature of the image. Finally, the global contour feature is combined with the SPM feature and the local position feature to generate a new combined feature.

3.2 Histogram intersection kernel weighted learning network

In order to make rational use of the combined features of three features, we design a Histogram Intersection Kernel Weighted Learning Network (HWLN) based on ELM. For clarity, we first present the Weighted Learning Network (WLN) structure. The WLN structure is described from two aspects of Fig. 4a and b. In Fig. 4, the meaning of the red solid lines and the blue dotted lines is consistent with Fig. 1.

In Fig. 4a, the input feature matrix is training combined samples $A = [a_1, a_2, \dots, a_N]^T \in \mathbb{R}^{N \times L}$, where N is the number of samples. The i -th sample feature is an L -dimensional vector $a_i = [a_{i1}, a_{i2}, \dots, a_{iL}]^T \in \mathbb{R}^L$, $t_i = [t_{i1}, t_{i2}, \dots, t_{in}]^T \in \mathbb{R}^n$ is the n -dimensional output vector. The input weight matrix $W_a = [w_{a1}, w_{a2}, \dots, w_{am}] \in \mathbb{R}^{L \times m}$ links the input layer A to the hidden layer B, and the output feature matrix of hidden layer B is $B = [b_1, b_2, \dots, b_N]^T \in \mathbb{R}^{N \times m}$.

$$B = \tanh(AW_a) \tag{1}$$

SENet uses the full-connected layer to calculate the channel weights of each convolutional layer, and to weight the features of each layer by channel-wise multiplication. In the process of training WLN, the method of channel-wise multiplication is not conducive to the derivation of

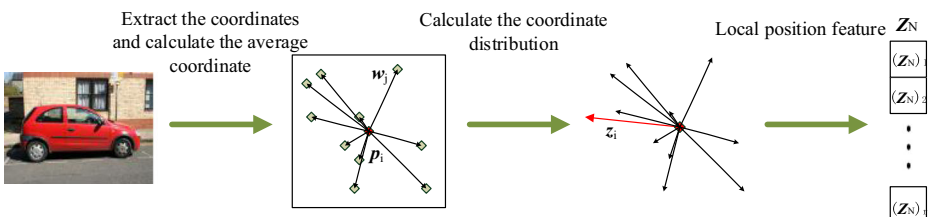


Fig. 2 The process of extracting the local position feature

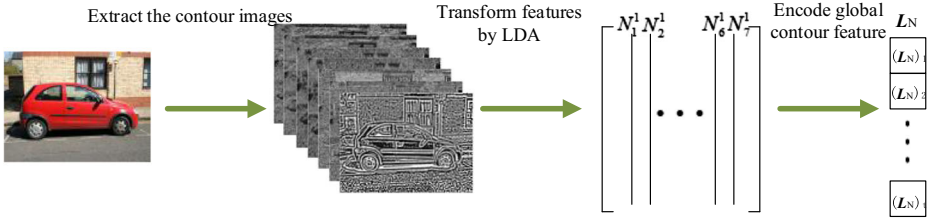


Fig. 3 The process of extracting the global contour feature

parameters of each layer. We are inspired by the weighted matrix qA equivalent to $(q - 1)A + A$, where q is weight coefficient. $(q - 1)A$ is seen as the function $f(A)$ of A . According to the equivalent thought of $qA = f(A) + A$, the hidden layer output feature matrix is added to the input feature matrix to achieve the effect of weighting.

The input layer A is connected to the hidden layer C by red solid lines. BW_b is added to the input feature matrix $A = [a_1, a_2, \dots, a_N]^T \in \mathbb{R}^{N \times L}$, and the biased matrix is mapped to the weighted matrix C using the activation function $\tanh(x)$.

$$C = \tanh(BW_b + A) \tag{2}$$

Where $W_b = [w_{b1}, w_{b2}, \dots, w_{bL}] \in \mathbb{R}^{m \times L}$ is the weight matrix between hidden layer B and hidden layer C , the output matrix of hidden layer C is $C = [c_1, c_2, \dots, c_N]^T \in \mathbb{R}^{N \times L}$. Similarly, the matrix B is biased by the coefficient CW_c , and the biased matrix is mapped to the weighted matrix D .

$$D = \tanh(CW_c + B) \tag{3}$$

Where $W_c = [w_{c1}, w_{c2}, \dots, w_{cm}] \in \mathbb{R}^{L \times m}$ is the weight matrix between hidden layer C and hidden layer D , the output matrix of hidden layer D is $D = [d_1, d_2, \dots, d_N]^T \in \mathbb{R}^{N \times m}$.

In Fig. 4b, the input layer and all hidden layers are connected to the output layer by blue dotted lines. The mapping relationship from the input layer A to the output

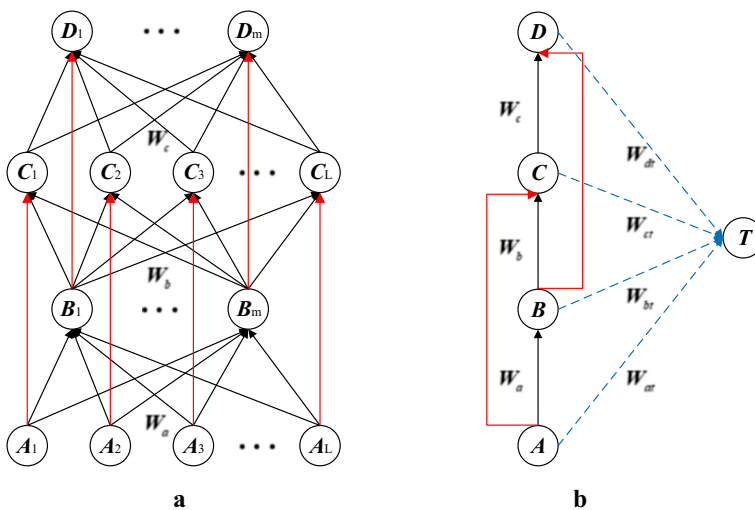


Fig. 4 Weighted Learning Network Structure. **a** represents the connection relationship between the layers in the WLN, **b** represents the connection relationship between each layer (A, B, C and D) and the T layer

layer T is linear, which can solve linear problems with higher efficiency. The relationship from each hidden layer to the output layer T is non-linear, which can very well realize the capability of nonlinear approximation from the input matrix to the output matrix. These matrices AW_{at} , BW_{bt} , CW_{ct} and DW_{dt} are added together to generate prediction matrix $T = [t_1, t_2, \dots, t_N]^T \in \mathbb{R}^{N \times n}$, which effectively achieves a combination of linear and non-linear. Another function of this structure is to combine weighted features with unweighted features.

The WLN, which is designed by combining the above two aspects, can perform a series of coefficient weightings on the input combination features to improve the distinguishability of features, and improve the classification accuracy of the network through the reasonable combination of linear and nonlinear.

In Fig. 4b, the output function of WLN is

$$f(A) = AW_{at} + BW_{bt} + CW_{ct} + DW_{dt} \tag{4}$$

The problem of solving the weight matrix of each layer of the network can be transformed into an optimization problem that satisfies the equality constraint:

$$\begin{aligned} \text{Minimize : } L_{PWLN} &= \frac{1}{2} \|W_{at}\|^2 + \frac{1}{2} \|W_{bt}\|^2 + \frac{1}{2} \|W_{ct}\|^2 + \frac{1}{2} \|W_{dt}\|^2 + \frac{1}{2} K \sum_{i=1}^N \|\xi_i\|^2 \tag{5} \\ \text{s.t. } a_i^T W_{at} + b_i^T W_{bt} + c_i^T W_{ct} + d_i^T W_{dt} &= t_i^T - \xi_i^T, \quad i = 1, \dots, N. \end{aligned}$$

Where K is penalty coefficient. Based on the KKT theorem, the problem of solving the weight matrix of each layer of WLN is equivalent to solving the following dual optimization problem:

$$\begin{aligned} L_{DWLN} &= \frac{1}{2} \|W_{at}\|^2 + \frac{1}{2} \|W_{bt}\|^2 + \frac{1}{2} \|W_{ct}\|^2 + \frac{1}{2} \|W_{dt}\|^2 + \frac{1}{2} K \sum_{i=1}^N \|\xi_i\|^2 \tag{6} \\ &- \sum_{i=1}^N \sum_{j=1}^n \alpha_{i,j} \left(a_i^T (W_{at})_j + b_i^T (W_{bt})_j + c_i^T (W_{ct})_j + d_i^T (W_{dt})_j - t_{i,j} + \xi_{i,j} \right) \end{aligned}$$

From the KKT optimality conditions of (6), we have

$$W_{at} = A^T \left(\frac{I}{K} + AA^T + BB^T + CC^T + DD^T \right)^{-1} T \tag{7}$$

$$W_{bt} = B^T \left(\frac{I}{K} + AA^T + BB^T + CC^T + DD^T \right)^{-1} T \tag{8}$$

$$W_{ct} = C^T \left(\frac{I}{K} + AA^T + BB^T + CC^T + DD^T \right)^{-1} T \tag{9}$$

$$W_{dt} = D^T \left(\frac{I}{K} + AA^T + BB^T + CC^T + DD^T \right)^{-1} T \tag{10}$$

Testing feature matrix is A^{test} , B^{test} , C^{test} and D^{test} are the testing output matrices of three hidden layers. By substituting formulas (7)–(10) and the above four testing matrices into (4), the testing output function of WLN can be written as

$$f(A^{\text{test}}) = (A^{\text{test}}A^T + B^{\text{test}}B^T + C^{\text{test}}C^T + D^{\text{test}}D^T) \left(\frac{I}{K} + AA^T + BB^T + CC^T + DD^T \right)^{-1} T \tag{11}$$

Radial Basis Function (RBF) $K(x, y)$ are used to map features into high-dimensional feature space.

$$f(A^{\text{test}}) = (A^{\text{test}}A^T + K(A^{\text{test}}, A) + K(B^{\text{test}}, B) + K(C^{\text{test}}, C)) \left(\frac{I}{K} + AA^T + K(A, A) + K(B, B) + K(C, C) \right)^{-1} T \tag{12}$$

Therefore, it is necessary to train the weight matrix of each hidden layer according to the input training feature matrix A in order to obtain the output matrix of each hidden layer.

In formula (12), the input weight matrix W_a is initialized randomly. The training process of WLN is as follow:

- 1) Compute the output matrix B using the random weight matrix W_a and formula (1), compute expected output C and D by (13) and (14).

$$C' = \tanh(A) \tag{13}$$

$$D' = \tanh(B) \tag{14}$$

- 2) Compute weight matrix W_{at} , W_{bt} , W_{ct} and W_{dt} by (15). Update expected output C'' and D'' by (16).

$$\begin{bmatrix} W_{at} \\ W_{bt} \\ W_{ct} \\ W_{dt} \end{bmatrix} = [A \ B \ C' \ D']^\dagger T \tag{15}$$

$$[C'' \ D''] = (T - AW_{at} - BW_{bt}) \begin{bmatrix} W_{ct} \\ W_{dt} \end{bmatrix}^\dagger \tag{16}$$

- 3) Compute weight matrix W_b by (17). Update actual output C by (18)

$$W_b = B^\dagger (\tanh^{-1}(C'') - A) \tag{17}$$

$$C = \tanh(BW_b + A) \tag{18}$$

- 4) Compute B^{test} and C^{test} using obtained weight matrix W_a and W_b .
- 5) Compute the network prediction $f(A^{\text{test}})$ by substituting $[A, B, C]$ and $[A^{\text{test}}, B^{\text{test}}, C^{\text{test}}]$ into (12).

The traditional ELM initializes weight matrix randomly. The randomly initialized matrix can cause instability of the hidden layer output matrix, increasing the uncertainty of subsequent training and making the network unable to achieve the best performance. In the structure of HWLN, the HIK [13] $H(x, y)$ is introduced between input layer A and hidden layer B of WLN instead of random initialization. As shown in (19), A is input into the kernel function to calculate B , which improves the robustness of the network while avoiding random initialization. Experimental results show that HIK has more obvious effect on the performance of WLN than the random initialization.

$$B = H(A, A) \tag{19}$$

$$B^{\text{test}} = H(A^{\text{test}}, A) \tag{20}$$

We rewrite formula (12) as:

$$f(A^{\text{test}}) = (A^{\text{test}}A^T + B^{\text{test}}B^T + K(B^{\text{test}}, B) + K(C^{\text{test}}, C)) \left(\frac{I}{K} + AA^T + BB^T + K(B, B) + K(C, C) \right)^{-1} T \tag{21}$$

In the HWLN, formulas (19) and (20) are used to compute B and B^{test} . Other training steps are consistent with WLN. Finally, $[A, B, C]$ and $[A^{\text{test}}, B^{\text{test}}, C^{\text{test}}]$ are substituted into (21) to generate network prediction value. After all the parameters are derived, the calculation process is as shown in Fig. 5.

3.3 Method procedure

As shown in Fig. 1, the SPM feature of the N -th image is S_N . The local position feature is Z_N . The global contour feature is L_N . The final feature vector of the image is S'_N . And The classifier is HWLN. The method is summarized as follows:

Algorithm

Input: $N(N = 1, 2, \dots, k)$ images

Output: Classification accuracy

1. Use SPM and LSAC methods to extract SPM feature S_N , and use l_2 -norm to normalize.
2. Calculate the local position feature Z_N and the global profile feature L_N of each image. For details, see the paper [27].
3. Concatenate S_N, Z_N, L_N of each image into S'_N .

$$S'_N = [S_N^T, Z_N^T, L_N^T]^T \tag{22}$$

4. Use HWLN to classify S'_N . For details, see 3.2.
-

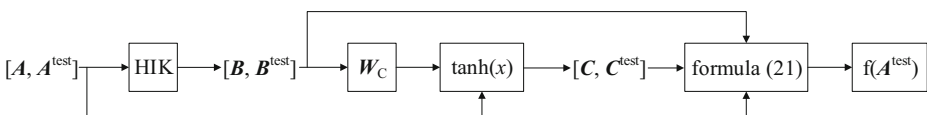


Fig. 5 The calculation process of HWLN

4 Experimental results

In our paper [27], the SPM features, local position features and global contour features are multiplied by their respective weights and then classified by SVM. The combination of the three features plays an active role in improving the accuracy rate. In this section, the experiments are performed from two aspects. First, the SVM in the libsvm [6] toolbox of MATLAB is used to classify the combined features to prove the effectiveness of the extracted features, which are not weighted. The kernel of SVM is HIK. Second, the HWLN is used to classify the combined features to prove the effectiveness of the proposed method.

All the experiments are performed according to the steps and parameters of in the paper [24], where the Dense-SIFT extraction parameters are single-layer 16×16 pixel patch and an interval of 8 pixels. In this paper, all of the experiments are repeated for 10 times. Each experiment randomly selects training images and test images. According to the average classification result of each experiment, the average of these 10 experimental results is calculated as the final experimental result. The experiments are performed on a PC with an Intel Core i5–8400 2.80GHz CPU and 8GB RAM running MATLAB 2014a.

4.1 Experiment of feature extraction

4.1.1 Caltech 101 database

The experiment of this section is first performed on the Caltech 101 database. The objects in the images of database are centered and occupied most of the area. Therefore, the database is used to test the effectiveness of the extracted features. In this experiment, 101 categories are used, randomly selecting 30 images from each category for training, and randomly selecting 50 from each of the remaining images for testing. If there are less than 50 remaining images, all the remaining images are selected.

As shown in Table 1, the extraction of SPM features takes 226 min. The local position features are extracted in 5 min. Setting the decomposition parameters of the NSCT to [1, 0, 2], the global contour features of the seven contour images take 48 min. From the experimental results in Table 1, it can be seen that adding local position features or global contour features in SPM can improve the accuracy of classification in the case that the codebook dimension is not changed. Adding two features at the same time can increase the accuracy by 14%.

In Fig. 6, the partial images with higher accuracy in the experiment are listed in the first row, and the partial images with lower accuracy are listed in the second row. Analyzing the experimental results and the corresponding images, it can be found that the category images with high accuracy have characteristics of simple background and clear outline of the objects. Most of those categories with low accuracy are wild animals. Because of their need for

Table 1 Experimental results of three features on Caltech 101 and MSRC

Database	Methods	Codebook size	Contour codebook	Contour images	Time (minutes)	Accuracy (%)
Caltech 101	SPM	400	–	–	226	64.13
	SPM + local+global	400	30	7	279	77.93
MSRC	SPM	400	–	–	98	87.61
	SPM + local+global	400	30	7	108	90.3

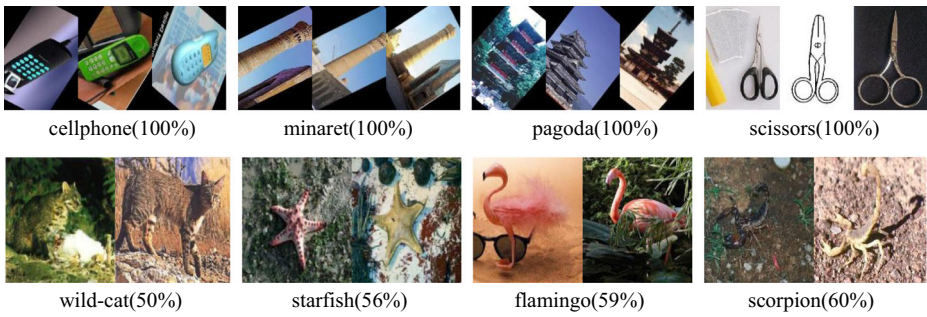


Fig. 6 Partial image with higher accuracy and lower accuracy

survival, wild animals often hide their bodies and integrate them with the environment. The images with high accuracy in this database are simple background images. If the images are replaced with complex background images, is the method still valid? In order to answer this question, the following experiment is conducted on the MSRC database.

4.1.2 MSRC database

This section experiments on the MSRC database, Microsoft Research Cambridge Object Recognition Image Database [31]. Images in MSRC is very similar to images in MSRC-21 [26]. Compared with the MSRC-21 database, images in the MSRC database have been classified into categories and are more convenient to use. Compared with other databases, the background of the database is more complex (as shown in Fig. 7). Even in the same category of images, each image has a great change in the camera perspective and appearance [26]. Therefore, the experimental results of this database can be used to answer the above question.

MSRC has a total of 18 large categories of object images, and some large categories also contain several subcategories. 18 categories are selected in this section, which include: bicycles (single), birds, building, cars (side view), chairs, chimneys, clouds, cow, doors, flowers, knives, leaves, planes, sheep, signs, spoons, trees, windows. Three categories are selected from the MSRC-21 database: boat, face, and office. A total of 21 categories of object images are selected for experimentation. 30 images are randomly selected for training in each category, and 30 images are selected for testing in the remaining images. A total of 1260 images are used for experiment.

As shown in Table 1, the extraction of SPM features takes 98 min. The local location features are extracted in 2 min. Setting the decomposition parameters of NSCT to [1, 0, 2], the global contour features of the 7 contour images are extracted in 8 min. Adding two features at



Fig. 7 MSRC database images

the same time can increase the accuracy by 2.69%. Even with cluttered backgrounds, our features have high classification accuracy for each categories of object image.

4.2 Experiment of feature classification

In Section 4.1, it has been proved that the extracted local features and global features are valid. Next, the extracted features are input into the HWLN to test the classification performance. In this section, experiments are performed on the Caltech 101, MSRC, and Caltech 256 databases. The experimental setup is the same as in Section 4.1.

We first verify the effectiveness of the WLN structure. The WLN weight matrix W_a is initialized randomly. After removing the C layer and the D layer in the WLN, the structure is denoted as WLN^- . The three structures WLN, KELM and WLN^- are compared. In the three structures, the 21 different values of K and RBF parameter γ are $\{2^{-10}, 2^{-9}, \dots, 2^9, 2^{10}\}$. In order to decrease the computation, the number of nodes in layer B and D in the WLN is set to 3000. The number of nodes in layer A and C is the same as the dimension of the input feature vector. As shown in a, b and c of Fig. 8, a large number of experiments have verified that when the value of K is larger and the RBF parameter γ is smaller, the classification effect of the three classifiers is better. Therefore, $K = 10$ and $\gamma = 0.5$ are set in all three classifiers. The accuracy of ELM for combined features is only about 1%, so they are not compared.

The effectiveness of the WLN structure is verified from three aspects. As shown in Table 2. WLN^- is compared with traditional KELM. The accuracy of WLN^- is increased by 4.54% and

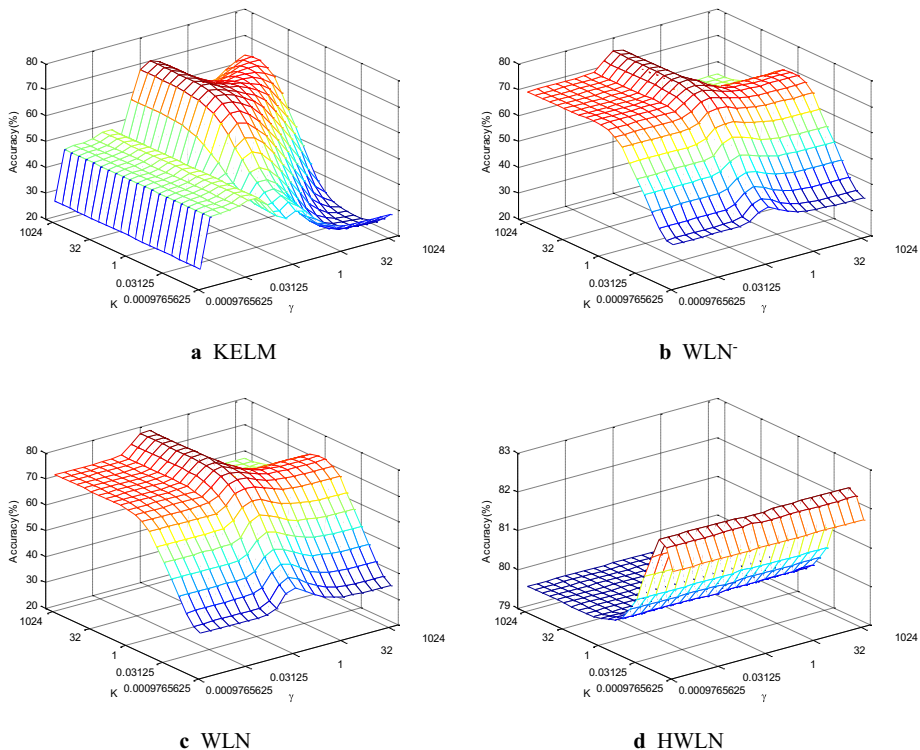


Fig. 8 Relationship between parameter selection and change trend of accuracy on Caltech 101

Table 2 Classification accuracy of WLN, WLN⁻ and KELM

Methods	Caltech 101		MSRC(%)	
	Accuracy(%)	Training time(s)	Accuracy(%)	Training time(s)
SPM + local+global+KELM	70.25	28.5	85.08	0.39
SPM + local+global+WLN ⁻	74.79	32	88.3	0.52
SPM + local+global+SVM	77.93	3.13	90.3	0.13
SPM + local+global+WLN	78.89	262.14	91.43	24.06

3.22% on the 101 database and the MSRC database, which fully prove the effectiveness of the combination of linear and nonlinear. WLN is compared with WLN⁻. The accuracy of WLN is increased by 4.1% and 3.13% respectively, which proves the effectiveness of the weighting structure. WLN is compared with SVM. The accuracy of WLN is increased by 0.96% and 1.13% respectively, which shows that rationally combining spatial information features can improve accuracy. KELM only needs to calculate the RBF of sample matrix A , and WLN⁻ needs to calculate randomly initialize W_a to calculate the hidden layer output matrix B and the product of A and A^T , so the training time is improved compared to KELM. Based on the WLN⁻, the WLN also needs to calculate the hidden layer output matrix C , and calculate the RBF of B and C , so the training time will be significantly increased. Although the time to train network parameters is significantly increased, WLN provides a training method for the multi-layer feed-forward neural network under the premise of avoiding repeated iterative training of the feedback network. It has become possible to increase the accuracy by increasing the number of network layers.

Then, HIK is added to the WLN structure to verify the effectiveness of HWLN. In HWLN, eq. (19)–(20) is used instead of random initialization of the weight matrix W_a . In the structure of HWLN, the 21 different values of K and γ are $\{2^{-10}, 2^{-9}, \dots, 2^9, 2^{10}\}$. The number of nodes in layers B and D is 3031, which is the same as the dimension of the output feature vector of HIK. As shown in Fig. 8d, the smaller the K in the HWLN is, the higher the classification accuracy is. The change in γ has little effect on the accuracy. Figure 8d is compared with the other three figures. When K and γ are changed, the accuracy of the HWLN is changed to a small extent, which indicates that the HWLN has strong robustness. For fair comparison, the value of γ is set to be consistent with γ of WLN. Therefore, $K = 0.0026$ and $\gamma = 0.5$. As shown in Table 3, HWLN is compared with WLN. It can be seen that the accuracy of HWLN is increased by 3.31% on the Caltech 101 and by 3.13% on the MSRC, which proves the effectiveness of HIK instead of random initialization of the weight matrix. When extracting features, the local features, contour features and SPM features are combined into sample matrix A , and HIK is used to map A into B . This avoids random initialization of W_a and

Table 3 Classification accuracy of HWLN, WLN and SVM

Methods	Caltech 101		MSRC(%)	
	Accuracy(%)	Training time(s)	Accuracy(%)	Training time(s)
SPM + local+global+SVM [27]	77.93	3.13	94.13	0.13
SPM + local+global+WLN	78.89	262.14	91.43	24.06
SPM + local+global+HWLN	82.2	193	95.56	6.9

Table 4 Classification accuracy of SPM, VLAD and FV on Caltech 101

Methods	Features	Codebook size	Accuracy (%)
SVM	SPM	400	64.13
	VLAD	256	72.41
	FV	256	64
	SPM + VLAD + FV	–	72.04
HWLN	SPM	400	70.76
	VLAD	256	78.37
	FV	256	77.73
	SPM + VLAD + FV	–	81.48

calculation of B . Therefore, the training time of HWLN will be decline. The results of artificially weighted features are listed in the first row of Table 3. The weights of the three features on Caltech 101 are [1], and the weights on MSRC are [1, 12, 0.2]. HWLN is compared with artificially weighted features. Compared to the accuracy of the artificially weighted features, the HWLN is increased by 4.27% on the Caltech 101 and by 1.43% on the MSRC. It can be shown that the improvement effect of HWLN on the accuracy rate is more obvious than that of artificial weighting.

We also select VLAD [20] features and FV [35] features for experiments to verify that the HWLN structure can also reasonably use other combined features. In Table 4, SPM, VLAD and FV are classified by classifiers respectively. After the three features are combined, the combined features are classified. The VLAD feature has the highest accuracy when SVM is used for classification. When HWLN is used for classification, the accuracy of each of the three features is higher than that of SVM. After the three features are combined, the accuracy rate is significantly improved. However, the classification accuracy of SPM + VLAD + FV + SVM does not exceed our method. This can prove that the rational use of spatial information on the basis of SPM feature is effective for improving accuracy.

Some of the classification accuracy on the Caltech 101 database are listed in Table 5. Compared with the improved method based on BoK, our method has achieved better classification results on the Caltech 101 database. In order to challenge more image categories, 150 objects are randomly selected from the Caltech 256 database for experimentation, 30 images are randomly selected for training, 30 images are selected for testing, and a total of 9000 images are selected for experimentation. The experimental results are shown in Table 6.

Table 5 Experimental results of several methods on the Caltech 101 database

Methods	Codebook size	Accuracy(%)
SPM [24]	200	64.60
SPS _{ad} + [22]	200	68.40
LSAC [29]	2048	72.58
LLC [37]	2048	73.44
YCbCr-SIFT+LSC + ELM [28]	1024	78
Zhu [41]	1000	74.3
SDCD+PHOW [30]	1024	75.37
Local+Global R ² FP [39]	–	76
ScSPM+DVM [14]	–	77.7
Supervised Deep [12]	2048	79.7
Koniusz et al. [23]	600	83.9
SPM + local+global+HWLN	400	82.2

Table 6 Experimental results of several methods on the Caltech256 database

Methods	Codebook size	Accuracy(%)
SPS _{ad} + [22]	4000	39.9
LLC [37]	4096	41.19
YCbCr-SIFT+LSC + ELM [28]	1024	44.42
Zhu [41]	1000	42.42
Supervised Deep [12]	2048	41.7
SPM + local+global+HWLN	1024	48

Through the rational use of the spatial information between the features, the classification accuracy of the object images has indeed been significantly improved.

In Table 5, the accuracy of method [23] is higher than that of the method in this paper. It is an improved method based on BoK, and it is a method with a higher accuracy in non-CNN classification methods. Koniusz et al. [23] uses higher-order symbiotic pooling method to make better use of existing image features, but does not effectively utilize the spatial information between features. Our method mainly focuses on the rational use of image spatial information. Compared with other improved methods based on BoK, our method utilizes spatial information locally and globally. The accuracy of the classification of the object image is improved while requiring only a low codebook dimension.

5 Conclusion

In this paper, we present an image classification method that includes feature extraction and classification. The extracted local spatial information and global spatial information are concatenated to SPM features to generate final features, which can represent these images. In the structure of HWLN, hidden layer output features are used to weight input features. The input layer and each hidden layer are directly connected with the output layer to combine linear and nonlinear features. The HIK is used to implement mapping of input features to hidden output features. The HWLN is used to perform a series of weighting combinations on the extracted three features to improve classification accuracy. Our method effectively improves the classification accuracy of the object image. Therefore, the rational use of spatial information on the basis of SPM feature can effectively improve the classification accuracy.

In terms of scene image classification, using SVM to classify the three extracted features achieves good classification results in 15-Scene database. However, the accuracy of HWLN does not exceed SVM. In the next step, our work will be focused on analyzing the reasons why HWLN has no advantage in scene image classification. Reasonable improvements are made to the reasons, so that our classification method can simultaneously adapt to the classification task of scene and object images.

Acknowledgements This work is supported by National Natural Science Foundation of China (No. 51641609), Natural Science Foundation of Hebei Province of China (No. F2015203212).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Ahmed KT, Irtaza A, Iqbal MA (2017) Fusion of local and global features for effective image extraction. *Appl Intell* 47(2):526–543
2. Anwar H, Zambanini S, Kampel M (2014) Encoding spatial arrangements of visual words for rotation-invariant image classification. In: German Conference on Pattern Recognition, 443–452
3. Avila S, Thome N, Cord M et al (2013) Pooling in image representation: the visual Codeword point of view. *Comput Vis Image Underst* 117(5):453–465. <https://doi.org/10.1016/j.cviu.2012.09.007>
4. Boiman O, Shechtman E, Irani M (2008) In defense of nearest-neighbor based image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR. 1–8
5. Bosch A, Zisserman A, Munoz X (2007) Image classification using random forests and ferns. In: Computer Vision, ICCV, pp 1–8
6. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2(3):389–396. <https://doi.org/10.1145/1961189.1961199>
7. Csurka G, Dance CR, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In Workshop on statistical learning in computer vision, ECCV, 1–22
8. Cunha ALD, Zhou JP, Do MN (2006) The nonsubsampling contourlet transform: theory, design, and applications. *IEEE Trans Image Process* 15(10):3089–3101. <https://doi.org/10.1109/TIP.2006.877507>
9. Deng WY, Ong YS, Zheng QH (2016) A fast reduced kernel extreme learning machine. *Neural Netw* 76: 29–38
10. Frome A, Singer Y, Malik J (2007) Image retrieval and classification using local distance functions. In: Advances in neural information processing systems, 417–424
11. Frome A, Singer Y, Sha F, Malik J (2007) Learning globally-consistent local distance functions for shape-based image retrieval and classification. *IEEE International Conference on Computer Vision*
12. Goh H, Thome N, Cord M, Lim JH (2014) Learning deep hierarchical visual feature coding. *IEEE Trans Neural Netw Learn Syst* 25(12):2212–2225
13. Grauman K, Darrell T (2005) The pyramid match kernel: Discriminative classification with sets of image features. *International Conference on Computer Vision*. 1458–1465
14. Gui J, Liu T, Tao D, Tan T (2016) Representative vector machines: a unified framework for classical classifiers. *IEEE Trans Cybernet* 46(8):1877–1888
15. Hu J, Shen L, Sun G (2017) Squeeze-and-Excitation Networks. arXiv preprint arXiv:1709.01507
16. Huang GB, Zhu QY, Siew CK (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. In: *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, 2004. IEEE, 985–990
17. Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. *Neurocomputing* 70(1–3):489–501
18. Huang FJ, Boureau YL, LeCun Y (2007) Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: *Computer Vision and Pattern Recognition, CVPR*, 1–8
19. Huang GB, Zhou H, Ding X, Zhang R (2012) Extreme learning machine for regression and multiclass classification. *IEEE Trans Syst Man Cybern Part B Cybern* 42(2):513–529
20. Jégou H, Douze M, Schmid C, Pérez P (2010) Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition. CVPR*. 3304–3311
21. Juneja M, Vedaldi A, Jawahar CV, Zisserman A (2013) Blocks that shout: Distinctive parts for scene classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 923–930
22. Khan R, Barat C, Muselet D, Ducottet C (2015) Spatial histograms of soft pairwise similar patches to improve the bag-of-visual-words model. *Comput Vision Image Understand* 132:102–112. <https://doi.org/10.1016/j.cviu.2014.09.005>
23. Koniusz P, Yan F, Gosselin P, Mikolajczyk K (2017) Higher-order occurrence pooling for bags-of-words: visual concept detection. *IEEE Trans Pattern Anal Mach Intell* 39(2):313–326. <https://doi.org/10.1109/TPAMI.2016.2545667>
24. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition New York*, 2169–2178
25. Li G, Niu P, Duan X, Zhang X (2014) Fast learning network: a novel artificial neural network with a fast learning speed. *Neural Comput & Applic* 24(7–8):1683–1695
26. Li WS, Dong P, Xiao B, Zhou L (2016) Object recognition based on the region of interest and optimal bag of words model. *Neurocomputing* 172(8):271–280. <https://doi.org/10.1016/j.neucom.2015.01.083>

27. Li YQ, Wu C, Li HB (2018) Image classification method combining local position feature with global contour feature[J]. *Acta Electron Sin* 46(7):1726–1731. <https://doi.org/10.3969/j.issn.0372-2112.2018.07.026>
28. Li Q, Peng Q, Chen J, Yan C (2018) Improving image classification accuracy with ELM and CSIFT. *Comput Sci Eng* 99:1–1
29. Liu LQ, Wang L, Liu XW (2011) In defense of soft-assignment coding. *Proceedings of the International Conference on Computer Vision*. 2486–2493. <https://doi.org/10.1109/CVPR.2010.5540039>
30. Mansourian L, Abdullah MT, Abdullah LN, Azman A, Mustaffa MR, Applications (2018) An effective fusion model for image retrieval. *Multimed Tools Appl*, 77 (13):16131–16154
31. Microsoft Research Cambridge Object Recognition Image Database, <https://www.microsoft.com/en-us/download/details.aspx?id=52644>
32. Nilsback ME, Zisserman A (2008) Automated flower classification over a large number of classes. In: *Computer Vision, Graphics & Image Processing*, 722–729
33. Perronnin F, Dance C (2007) Fisher kernels on visual vocabularies for image categorization. In: *IEEE conference on computer vision and pattern recognition*, 1–8
34. Perronnin F, Sánchez J, Mensink T (2010) Improving the fisher kernel for large-scale image classification. In: *European conference on computer vision*, 143–156
35. Sánchez J, Perronnin F, Mensink T, Verbeek J (2013) Image classification with the fisher vector: theory and practice. *Int J Comput Vis* 105(3):222–245
36. Van Gemert JC, Veenman CJ, Smeulders AW, Geusebroek JM (2009) Visual word ambiguity. *IEEE Trans Patt Anal Mach Intell* 32(7):1271–1283
37. Wang JY, Yang JC, Yu K, et al (2010) Locality-constrained linear coding for image classification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3360–3367
38. Wang S, Lu J, Gu X, Yang J (2016) Semi-supervised linear discriminant analysis for dimension reduction and classification. *Pattern Recogn* 57:179–189
39. Xiong W, Zhang L, Du B, Tao D (2017) Combining local and global: rich and robust feature pooling for visual recognition. *Pattern Recogn* 62:225–235
40. Zafar B, Ashraf R, Ali N, Ahmed M, Jabbar S, Chatzichristofis SA (2018) Image classification by addition of spatial information based on histograms of orthogonal vectors. *PLoS One* 13(6):e0198175
41. Zhu QH, Wang ZZ, Mao XJ, Yang YB (2017) Spatial locality-preserving feature coding for image classification. *Appl Intell* 47(1):148–157
42. Zou J, Li W, Chen C, Du Q (2016) Scene classification using local and global features with collaborative representation fusion. *Inf Sci* 348:209–226



Chao Wu received the B.S. degree in Automation Engineering from Luoyang Institute of Science and Technology, China, in 2015. He is currently pursuing Ph.D. degree at Yanshan University. His main research interest is Pattern Recognition and Machine Learning.



Yaqian Li is an assistant professor in the institute electrical engineering from the Yanshan University, Qinhuangdao, China. She has received her Ph.D. degree in the institute national des sciences appliqués de Rouen in 2010. Her current research interests lie in the area of Pattern Recognition.



Zhibiao Zhao received the B.S. degree in Measurement and Control Technology and Instrument from Yanshan University, China, in 2012. He is currently pursuing Ph.D. degree at Yanshan University. His main research interest is Artificial Intelligence.



Bin Liu is a professor in the college of information science and technology from the Yanshan University, Qinhuangdao, China. He has received his Ph.D. degree in the institute electrical engineering from the Hebei University of Technology, Tianjin, P.R. China in 2009. His current research interests lie in the area of low power wireless sensor networks and topology control.