He *EURASIP Journal on Image and Video Processing* (2019) 2019:38
https://doi.org/10.1186/s13640-019-0425-8

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Image classification based on sparse coding multi-scale spatial latent semantic analysis

Tao He

Abstract

In the face of huge amounts of image data, how to let the computer simulate human cognition of images and automatically classify images into different semantic categories have become a key issue in image semantic analysis. Image classification is based on some attribute of the image, and it is divided into pre-set categories. For human beings, image classification is not difficult but there is a series of problems in using computers to classify images: (1) images contain a large amount of information, which is complex, diverse, and indescribable; and (2) there is a huge difference between the physical expression of images and the conceptual information known by human beings. The traditional sparse coding method loses the spatial information when classifying images. In this paper, spatial pyramid multi-partition method is used to add spatial information restriction to the feature. The proposed multi-scale spatial latent semantic analysis method based on sparse coding has higher average classification accuracy than many existing methods, which verifies its effectiveness and robustness. Experiments also show that the classification accuracy of this paper is 2.1% higher than that of sparse coding for image classification (ScSPM) and the classification performance is 3.1% higher than that of ScSPM when the number of training images is 40. Compared with other methods, the classification performance of the proposed method is improved significantly.

Keywords: Sparse coding, Multi-scale space, Latent semantic analysis, Image classification, Image segmentation

1 Introduction

With the rapid development of high-speed Internet, the development of information storage and transmission technology, and the popularity of digital equipment, the acquisition and storage of digital color images become easier and the number of image data people come into contact with is increasing at an unprecedented rate. Faced with a huge amount of image data, how to simulate the cognitive mechanism of human understanding of images and automatically classify images into different semantic categories according to the way people understand become a key issue. In addition, image classification not only includes people's overall understanding of an image, but also provides the context in which objects appear in the image, which provides a basis for further identifying other content in the image [1]. Therefore, image classification has become a hot topic in the field of computer vision and multimedia information processing.

According to the different ways of describing images in traditional image classification methods, image classification methods can be divided into two categories, one is based on global features and the other is based on middle-level semantic information. Traditional visual dictionary models usually use K-means clustering and other vector quantization coding methods to encode local features to generate visual codebook, that is, visual dictionary [2, 3]. Then, the local feature of the image is assigned to the nearest visual word to form a visual vocabulary histogram expressing the content of the image. However, the methods of image representation and classification by vector quantization encoding have the following problems: (1) visual codebook generated by vector quantization encoding will lead to the loss of spatial information, (2) visual vocabulary histogram construction leads to serious quantization errors, and (3) this kind of image representation will show good classification effect only when using non-linear kernel SVM classifier. The training and testing time complexity of non-linear SVM classifier are O ($n2$-$n3$) and O (n), respectively. N is the number of training images, which

Correspondence: het@sziit.edu.cn
Shenzhen Institute of Information Technology, Shenzhen, China

will undoubtedly reduce its practicability; it is difficult to apply to a large-scale dataset classification.

Visual neurophysiology studies have shown that mammalian visual cortex neurons have directional, localized, and band-pass characteristics and the visual perception system only needs a few neurons to obtain the main information in the image and to form a sparse signal representation [4]. M. Kumar et al.'s [5] physiological experiments showed that the responses of primary visual cortex neurons to natural images as stimuli were sparsely distributed. D. Voegeli [6] suggests that in the early stages of the visual cortex, sparse coding (SC) [7–10] is achieved by discovering local statistical rules of natural images and thereby reducing redundancy of neuronal signals. The middle part of the visual cortex may be represented by a small number of special atoms. That is to say, a small number of base vectors are selected from the over-complete dictionary by sparse coding to represent the input information.

Researchers have also done a lot of work to solve the problem of missing spatial information of visual words in traditional visual dictionary models. Spatial pyramid matching (SPM) divides an image into multiple blocks of the same size at multiple scales. The visual lexical histogram descriptions of each image block are generated to compensate for the lack of word space information [11]. S. Rousseau and others extend the initial spatial pyramid matching method and integrate the spatial position information of visual words into the visual dictionary model [12]. Zhao Chunhui and Zhao Zhongqiu proposed an image classification method based on sparse coding and multi-scale space to solve the problem of serious quantization error. This method not only compensates for spatial information, but also improves the robustness of image content expression [13, 14]. To solve the problem of serious quantization error, S. Zhang anf S. Zhang.et.al proposed a soft allocation method [15–17], which effectively reduced the quantization error and enhanced the accuracy of image expression. Weinshall and others combined the soft allocation strategy with the latent Dirichlet allocation model (LDA) and proposed a soft allocation LDA model [18, 19]. In order to obtain non-linear image expression vectors, we can still get better image classification performance in the case of linear kernel SVM classifier. J. Zhu, et al . used fast sparse coding algorithm to generate SIFT descriptor-based visual dictionary and sparse vector combined with spatial pyramid matching (SPM) algorithm [20–22] and proposed sparse coding-based spatial pyramid matching method for image classification. This method uses linear kernel SVM classifier (can be used). Reducing the training time complexity of the classifier to $O(n)$ can still achieve better classification performance, thus effectively reducing the training time of the classifier and enhancing the practicability of the classification method. Through further study of sparse coding results, Y. Gao et al. [23, 24] proposed a locality-constrained linear coding (LLC) method based on sparse coding, which further improved the coding performance. The premise of high classification accuracy is that the local feature and its k-nearest neighbors are in the same subspace [25]. However, the k-nearest neighbor algorithm is very sensitive to noise, so it is difficult to ensure that the k-nearest neighbors are in the same subspace; therefore, Zhuang et al. [26, 27] proposed a non-negative sparse local linear coding (NSLLC) algorithm based on LLC algorithm to encode local features. Moreover, compared with sparse representation, non-negative sparse representation is more suitable for image classification tasks. However, all the sparse coding models mentioned above belong to the sparse representation model of signal reconstruction error minimization. They aim at minimizing signal reconstruction error and ignore the importance of discriminability to image classification task.

Aiming at the shortcomings of the existing methods, in order to solve the shortcomings of the previous methods, and taking into account the spatial distribution of regional semantics in similar images, there are often some rules; in this paper, an image classification method based on sparse coding and multi-scale spatial latent semantic analysis is proposed. This method uses spatial pyramid to divide the image into spatial layers and local regions to obtain the spatial relationship between the local blocks. Then, it uses SC to soft-quantify each local block to form a co-occurrence matrix. Then, it uses the PLSA model to mine the latent semantic information of each local block to obtain its latent semantic information distribution. Finally, the latent semantic information collected at different scales is concatenated by weights to get the final feature description of the image. The feature takes into account not only the local latent semantic information of images, but also the spatial information of images. Experimental results show that the algorithm can generate multi-scale spatial latent semantic information with high classification performance. The classification accuracy is 2.1% higher than that of ScSPM, and the classification performance is 3.1% higher than that of ScSPM when the number of training images is 40. Compared with other methods, the classification performance of the proposed method is improved significantly.

2 Proposed method

Image can be regarded as a collection of several local regions. The method proposed in this paper is to classify images according to their spatial distribution and local latent semantic information.

2.1 Local feature extraction

At present, features are widely used in the field of computer vision, such as scene classification, object recognition, and motion detection, because of their unique characteristics, rich information, robustness, and scalability. According to the description, the extraction process of the feature, as shown in Fig. 1, contains four steps.

(1) Detection of extreme points in scale space—establishing differential Gauss scale space for input images, and search in this scale space, the critical point is determined by the cable extreme point.
(2) Accurate location of key points—Taylor expansion of differential Gaussian function and interpolation operation are used to accurately locate key points. At the same time, they can filter out low contrast points and strong edge response points.
(3) Key point direction parameter assignment—statistical key point neighborhood pixel gradient direction histogram designates key points for two parameters of main direction and auxiliary direction.
(4) SIFT describes sub-generation—the neighborhood of the key point 16 × 16 is divided into 16 4 × 4 small neighborhoods, and then, a histogram of the gradient direction of eight lattices is computed in each small neighborhood; thus, a 128-dimensional feature descriptor is constructed for each key point and the scale space of a two-dimensional image can be obtained by convolution of image and Gauss kernel. Set $I(x, y)$ is the original image, $L(x, y, \sigma)$ is the transformed image, and

$$L(x, y, \sigma) = G(x, y, \sigma) \times I(x, y) \tag{1}$$

Among them, the scale $G(x, y, \sigma)$ is the variable two-dimensional Gauss function.

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{2}$$

In formula (1), (x, y) is the image position coordinate and σ represents the scale space factor; the larger σ, the more the image is smoothed and the larger the corresponding scale. Small scale describes the detail features of images and the general features in large scale.

Thus, the pyramid can be built for an input image. The pyramid is divided into O groups, each with S layers, and the next group of images is sampled from the previous group of images.

When detecting the extremum of scale space, each pixel must be compared with all of its adjacent points, including the neighborhood pixels of the same scale and the neighborhood pixels of 9 × 2 corresponding to the next and next adjacent scales, which totals to 26 pixels, so as to ensure that both the scale space and the two-dimensional image space are extremum points.

Because the difference of Gaussians (DoG) value is sensitive to noise and edge, it is necessary to fit the local extremum points detected in DoG scale space by three-dimensional quadratic function to determine the position and scale of the key points accurately and to remove the unstable points of edge response and low contrast points, so as to enhance the stability and anti-noise ability of the feature points. The two Taylor expansion of the DoG function is as follows:

$$D(x) = D + \frac{\partial D^{\mathrm{T}}}{\partial x} x + \frac{1}{2} x^{\mathrm{T}} \frac{\partial^2 D^{\mathrm{T}}}{\partial x^2} x \tag{3}$$

Among them, $x = (x, y, \sigma)^{\mathrm{T}}$ is the exact position of the key point and the position between the extremum points in the scale space and the scale offset vector determined in the first step; the derivative of Eq. (4) is obtained, and let $\frac{\partial D}{\partial x} = 0$, precise key points can be obtained as follows:

$$\hat{x} = \left(-\frac{\partial^2 D}{\partial x^2} \right)^{-1} \frac{\partial D}{\partial x} \tag{4}$$

In order to enhance the stability of key points and improve the anti-noise ability, it is necessary to remove the points with unstable edge response and low contrast. Among them, the point with unstable edge response has a larger principal curvature in the direction across the edge and a smaller principal curvature in the direction perpendicular to the point with unstable edge response. The two principal curvature values are compared with the following Hessian matrix eigenvalues.

$$H = \begin{bmatrix} D_{\mathrm{xx}} & D_{\mathrm{xy}} \\ D_{\mathrm{yx}} & D_{\mathrm{yy}} \end{bmatrix} \tag{5}$$

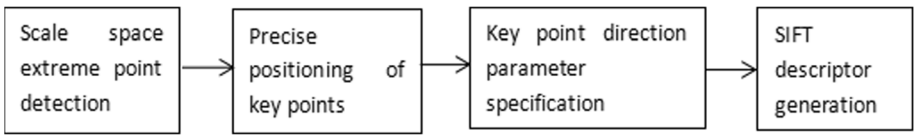

Fig. 1 SIFT feature extraction algorithm flow

where the partial derivative is the partial derivative of the DoG operator at the extreme point, set α to H is the maximum eigenvalue, β to H is the minimum eigenvalue, let $\alpha = r\beta$, and

$$Tr(H) = D_{xx} + D_{yy} = \alpha + \beta$$
$$\mathrm{Det}(H) = D_{xx}D_{yy} - \left(D_{xy}\right)^2 = \alpha\beta \tag{6}$$

2.2 Sparse coding

In the feature encoding stage of the word bag model, the hard assignment encoding method in the previous section can effectively realize the mapping from local feature points to visual words, but this one-to-one mapping method is too strict. Relevant studies have shown that although lexicographic reasonableness is ensured in word allocation from local features to nearest neighbors, i.e., the selected words are most relevant to local features, the ambiguity of dictionaries is not taken into account, i.e., the correlation between local features and other dictionary words. In recent years, the over-complete sparse representation of images is a research hotspot in the field of image recognition. The sparse coding theory introduced in the modeling of local features aims to preserve the correlation between a local feature and its most relevant visual words.

Natural images are very complex signals, usually not in sparse themselves, and contain a variety of morphological and structural components. In natural images, many morphological structures, such as wheel galleries, edges, and shapes, form an over-complete dictionary in the form of base vectors (that is, the dimension of base in the dictionary is much smaller than the number of base). Then, the sparse coding representation of an image in a dictionary is obtained by sparse constraints under certain reconstruction errors.

$$\arg\ \min_{D,C} \sum_{i=1}^{n} \|x_i - Dc_i\|^2 + \lambda\|c_i\|_0,\ \ \text{s.t.}\|c_i\|^2 \le 1 \tag{7}$$

Among them, the first one is to reconstruct the construction error; the second is the sparse regularization term. Sparse terms are very important because they not only make the objective function have a unique solution, but also constrain the sparsity of the coding, to ensure that the input x_i is represented only by the more significant feature patterns.

Set up $X = [x_1, x_2, \cdots, x_n] \in R^{m \times n}$ is the local characteristics of the image, and $D = [d_1, d_2, \cdots d_k] \in R^{m \times k}$ is the dictionary or codebook. Among them, the base vector D is usually over-perfect, that is, $k > > n$, and $C = [c_1, c_2, \cdots, c_n] \in R^{k \times n}$ is the sparse coding coefficient. Parameters λ are adjusted to reconstruct error terms and sparse terms. $\|^*\|$ subscript default indicates 2-norm. $\|^*\|_0$ is the L_0 norm, a number defined as a non-zero element of a vector (or matrix).

L_0 norm minimization is a kind of NP-hard problem, which usually replaces L_0 norm with L_1 norm, transforms the original problem into a convex problem, and then solves it by convex optimization method. The sparse representation model is as follows:

$$\arg\ \min_{D,C} \sum_{i=1}^{n} \|x_i - Dc_i\|^2 + \lambda\|c_i\|_1,\ \ \text{s.t.}\|c_i\|^2 \le 1 \tag{8}$$

In detail, the sparse coding learning process can be decomposed into two parts for optimization.

(1) The construction of over-complete sparse representation dictionary.

The fixed X formula (7) is transformed into a constrained least squares problem. The expression is as follows.

$$\arg\ \min_{D} \sum_{i=1}^{n} \|x_i - Dc_i\|^2 + \lambda\|c_i\|_0,\ \ \text{s.t.}\|c_i\|^2 \le 1 \tag{9}$$

The problem can be solved by gradient projection or transformation to dual space.

(2) Solving sparse representation under dictionary.

The fixed dictionary D (7) is transformed into an unconstrained least squares problem. The expression is as follows.

$$\arg\ \min_{C} \sum_{i=1}^{n} \|x_i - Dc_i\|^2 + \lambda\|c_i\|_1 \tag{10}$$

The problem can be solved by orthogonal matching pursuit (OMP) or feature-sign search (FSS).

The traditional space pyramid matching and sparse coding methods are combined. The specific process includes two stages: training and coding. When training, SIFT features are extracted from known images and encoded. When coding, the same method is used to obtain the sparse coding of an image representing $C = [c_1, c_2, \cdots, c_m] \in k \times m$. Then, the image is divided into pyramids and the coding vectors in each region are max pooling, that is, the maximum value of all coding vectors in a region is taken to form a k-dimensional vector to represent the characteristics of the region. Finally, they form a $(1 + 4 + 16 + \cdots) \bullet K$ dimension characteristic vector, that is, the ScSPM feature.

Essentially, sparse coding is used to solve the sparse representation of the eigenvector matrix under the base vector by regularization constraint L1 norm under certain reconstruction errors. The coding coefficients of sparse coding are only a few non-zero, and most of them are zero. It is obvious that sparse coding has several disadvantages.

(1) The encoding speed is slow. In the process of coding, the feature coding of each block in the image needs to be iteratively computed by the regularization constraint L1 norm, which is computationally expensive and memory-consuming.
(2) Traditional sparse coding codes independently encode each feature local descriptor, ignoring the correlation between local descriptors, such as spatial relationship or structural layout relationship.
(3) Sparse coding adopts independent coding method for features, and the coding of the same image may change because of the influence of illumination changes, occlusion, and other noises. In addition, sparse coding also ignores the correlation between local descriptors and good coding representation should make similar image features have similar coding to minimize intra-class differences.

2.3 Sparse representation of images

The assumption of image sparse representation is that the image signal can be approximated well by a linear combination of a small number of base vectors in an over-complete dictionary and can be achieved by optimizing the following expression:

$$\min_{B,X} \|Y-BX\|_2^2 + \lambda\|X\|_1$$

$$s.t.\left\|B_j\right\|_2 \le 1 \quad \forall j \tag{11}$$

Among them, $Y = \{y_1, y_2, \cdots, y_N\} \in R^{\mathrm{d}\times \mathrm{N}}$ is the training sample and $B \in R^{\mathrm{d}\times \mathrm{K}}$ is an over-complete dictionary to learn. B_j indicates B in the j column, and $X = [x_1, x_2, \ldots, x_N] \in R^{\mathrm{K}\times \mathrm{N}}$ is the sample Y in the dictionary B sparse representation coefficient. The formula to sum is not convex. However, when any one of the fixed sum is considered, the expression is transformed into a convex optimization problem, which can be solved by alternating optimization methods. Therefore, we can transform the formula into two sub-problems.

(1) Dictionary learning

When fixed X, the formula (11) is transformed into the constrained least squares problem.

$$\min_{B} \|Y-BX\|_2^2 + \lambda\|X\|_1$$

$$s.t.\left\|B_j\right\|_2 \le 1 \quad \forall j \tag{12}$$

The problem can be solved by gradient projection or transformation to dual space.

(2) Sparse decomposition

When fixed B, formula (11) can be transformed into the following unconstrained least squares problem:

$$\min_{X} \|Y-BX\|_2^2 + \lambda\|X\|_1 \tag{13}$$

The problem can be solved by orthogonal matching pursuit or feature symbol search algorithm. Natural images have been proved to have sparse structure, so sparse coding is suitable for image representation. Sparse coding consists of two steps, dictionary learning and sparse decomposition, which correspond to the construction of visual vocabulary and sparse vector representation in image representation.

Sparse coding model-based image local semantic concept representation has the following advantages: First, sparse coding model map image features high-dimensional space, compared with low-dimensional vector, and high-dimensional vector is more conducive to image classification. Second, each feature point in the image is represented by a number of base vectors in an over-complete dictionary, which reduces the quantization error and makes the image more accurate. Finally, the non-zero coefficients of sparse representation actually reveal the classification relationship of signals. Therefore, it is very advantageous for scene classification task to obtain image representation using sparse coding model.

However, these sparse coding models aim at minimizing signal reconstruction error. For image scene classification, it is more important to find a discriminant representation than to minimize the reconstruction error. Therefore, if discriminant analysis can be added to the sparse coding model to enhance the discriminant of image sparse vector representation, it will play a great role in improving the performance of scene classification. In addition, through more in-depth study of sparse coding, some researchers believe that locality is more important than sparsity. Therefore, if the locality of sparse coding model can be further considered, it will be of great significance to improve the performance of image sparse representation and the effective utilization of image spatial information. We will analyze the shortcomings of the existing sparse coding models in detail in the fourth chapter and improve it.

2.4 PLSA image local feature semantic extraction

The main idea of the probabilistic latent semantic analysis model is to analyze the co-occurrence of words in a document set and to take the probability distribution of words as the theme $z_k(k = 1, 2, \cdots, K)$. The principle of the model is shown in Fig. 2.

Figure 2 shows the PLSA graph model representation. The black box in the graph represents the repeated generation of M documents and N words in each document. The solid parts d and w are observed variables, and the

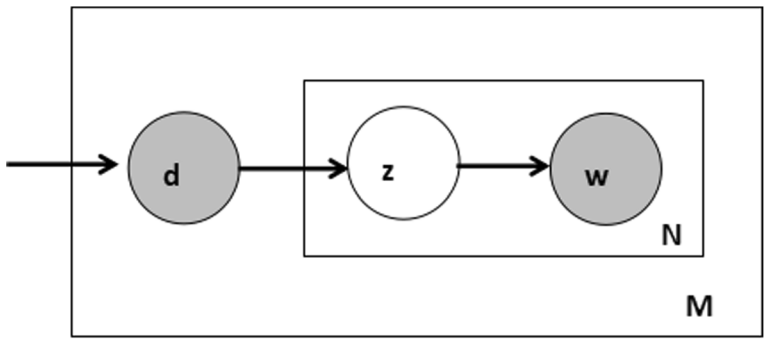

Fig. 2 PLSA graph model representation

hollow part z represents unknown variables that need to be predicted by the model. Given a collection of documents $D = \{d_1, d_2, \ldots, d_M\}$, the words in the document are taken from the glossary $W = \{w_1, w_2, \ldots, w_v\}$ and the co-occurrence frequency matrix of documents and words $N = [n(d_i, w_j)]$ can be obtained. Among them, $n(d_i, w_j)$ is the statistical documents, d_i is the word, and w_j is the number that appears. Using z represents latent semantic topics, assuming that words in a document are generated by latent semantic topics, and the entire document generation process is as follows:

(1) Select a document whose probability is expressed as $P(d)$.
(2) Choose a hidden theme Z, making $P(z|d)$ satisfy the polynomial distribution.
(3) Under the condition of known subject, conditional probability $P(w|z)$ of word w satisfies polynomial distribution. The joint distribution of words and documents generated by the above generation process can be expressed as:

$$P(w,d) = \sum_{z \in Z} P(w,d,z) = P(d) \sum_{z \in Z} P(w|z)P(z|d) \tag{14}$$

Because of $P(w,d) = P(d)P(w|d)$, according to formula (14), $P(w|d)$ can be written as:

$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \tag{15}$$

According to Fig. 2, Eq. (15) can be considered a matrix decomposition process in which the word distribution in each document is composed of convex combinations of latent semantic topics, weights $P(w|z)$ calculate the conditional distribution of words in known subjects, and $P(w|z)$ has nothing to do with specific documents.

By iterating the following maximum logarithmic likelihood function with EM algorithm, the conditional distribution in PLSA model can be estimated as $P(w|z)$ and $P(z|d)$.

$$L = \log P(D,W) = \sum_{d \in D} \sum_{w \in W} n(d,w) \log P(w,d) \tag{16}$$

As can be seen from the above, the model predicts the semantic topic of unknown documents by analyzing the content of known documents and learning model parameters from these documents.

Because the computational complexity of LDA model is too complex, the PLSA model can meet the needs of image scene classification tasks. Therefore, the concept of image local semantics is expressed based on the PLSA model.

The PLSA model aims to extract topics from documents. Similarly, after introducing the visual dictionary model into the field of image scene classification, the PLSA model can be used to extract semantic topics from images and then the images can be represented as the distribution of latent semantic topics. Local semantic concept representation based on the PLSA model needs to construct the visual word description of the image first and then use the PLSA model to mine the latent semantic topics in the image.

After the visual word description of the image is constructed, the PLSA model regards the whole image as a document and uses the d representation, and regards the visual words in the image as the words in the document and uses the w representation.

Then, statistics of the frequency of each visual word is in the image. The co-occurrence frequency matrix $N = [n(d_i, w_j)]$ of image and visual word can be obtained. Among them, $n(d_i, w_j)$ represents in images d_i and the number of occurrences of visual words w_j. Suppose, Z represents the set of latent semantic topics in an image, then the extraction of latent semantic topics

based on the PLSA model can be divided into the following two stages:

(1) Training phase—the PLSA model is used to train all the images in the training image set through the EM algorithm until the algorithm converges to get the latent semantic topic $P(w|z)$. $P(w|z)$ describes the distribution of visual words in the latent semantic theme of an image.
(2) Inference phase—for test images, we keep $P(w|z)$ the same and also use the PLSA model to iterate through the EM algorithm until convergence and to get potential semantic topics $P(z|d)$. $P(z|d)$ represents the probability of an image containing latent semantic topics.

For an image d, supposing the number of latent semantic topics in it is T, the latent semantic topics are extracted based on the PLSA model. We can get a T dimension eigenvector $[P(z_1|d), P(z_2|d), \ldots, P(z_T|d)]$, this is the latent semantic theme feature of the image, and then the SVM classifier can be used to realize the image scene classification.

3 Experimental results

Experiments are carried out on an image library. The image sources in the library include image sets, image search engines, and personal photo albums. It is the most commonly used benchmark set in the field of scene classification. According to the general experimental settings, 100 training samples were randomly selected and the remaining images were taken as test samples. Firstly, the OB-representation features of the eight types of scene images are extracted and the SVM model is trained to get three scene themes (street, tall building, inside city, highway), field themes (forest, mountain, open country), and beach themes (coast). Then, three SVM classifiers are learned by using low-rank coding, respectively. In the test, the response value of the object is extracted from the new input image and the scene topic label is obtained by input training. Then, the suitable single SVM classifier is selected, and the encoding representation matrix is obtained by using the f LCLR encoding method, and the classification result is obtained by inputting the single SVM classifier.

The experimental features are characterized by dense SIFT. The extraction method is as follows: firstly, the image is divided into 16×16 pixel size and 8 pixel interval image blocks, then each image block is divided into 4×4 sub-regions, and then the gradient histogram of 8 directions on each sub-region is calculated as seed points. Finally, the seed points on 4×4 sub-regions are connected to get 128-dimensional SIFT feature vector. Because of the classification of multi-class scenes, SVM uses "one-to-many" strategy to construct multiple classifiers, that is, each scene category is trained to its own SVM classifier. The algorithm runs in Visual C++ 6.0 and MATLAB 7.0. The hardware is configured as a processor P42.6G, 8 memory computer. Parameters are determined by cross validation. At the same time, in order to ensure the objectivity of the results, 10 random experiments were carried out independently in each database and the average classification accuracy and standard variance were used as evaluation indicators.

4 Discussion

Figure 3 shows some of the sample images for each scenario of scene 13 and scene 15. To be fair, in each random experiment, 100 images of different categories were randomly selected as training sets and the remaining

Fig. 3 Partial examples of scene 13 and scene 15 for each scenario

Fig. 4 Part of the sample image of the Caltech-101 image dataset

images were taken as test sets. The Caltech-101 image dataset (part of its example image is shown in Fig. 4) contains 121 categories and 10,101 pictures. To compare with previous methods, 15 images were randomly selected from each category for training and the remaining images were tested.

1. Verify the necessity of PLSA latent semantic information extraction through scene 13 and scene 15 datasets. Table 1 shows the classification results based on sparse encoding space pyramid matching (ScSPM), principal component analysis (PCA), and ScSPM combined with PLSA
   The experimental results in Table 1 show that the classification accuracy of ScSPM+PLSA method is improved by 1.9% and 2.4%, respectively, compared with the simple SCSPM and SCSPM+PCA method on scene 15 dataset. The results fully show that the latent semantic information obtained by learning the PLSA model in each local area can improve the classification accuracy of images and verify the importance of PLSA in the image classification model in this paper.
2. The influence of latent semantic number of Fig. 5 on classification accuracy (dictionary size 1024, $L = (0)$, $L = (0, 1)$, $L = (0, 1, 2)$) represents three levels of spatial pyramid matching, respectively
   The influence of latent semantic number on classification accuracy is analyzed through scene 13 dataset experiment. Figure 5 shows the trend of classification accuracy with the increase of latent semantic number. As can be seen from Fig. 5, within a certain range, the classification accuracy will be improved with the increase of the number of topics, but when the number of topics exceeds a certain range, the classification accuracy will be reduced, and when the number of topics is 50, the classification accuracy reaches the maximum. When the spatial pyramid is set to three layers, the classification accuracy is the highest, which fully shows that multi-scale spatial matching is conducive to discovering more spatial location information of image targets and improving the classification accuracy.
3. Comparing the performance of the other four algorithms (spatial pyramid matching kernel (KSPM), sparse coded spatial pyramid matching (ScSPM), probabilistic latent semantic analysis (PLSA), and subject space relations (SR-PLSA), Fig. 6 shows the comparison of the classification accuracy between this method and other better methods in scene 13 and scene 15. Figure 7 shows the comparison of the classification accuracy between this method and other better methods in Caltech-101

Table 1 Comparison of classification accuracy of ScSPM, ScSPM, and PCA and their combination PLSA (%)

Method	Scene 13 dataset	Scene 15 dataset
ScSPM	85.31 ± 0.69	80.81 ± 0.67
ScSPM+PCA	85.01 ± 0.64	80.51 ± 0.42
ScSPM+PLSA	86.54 ± 0.51	82.61 ± 0.62

The average classification accuracy of this method on scene 15 dataset is 83.12%. Figure 6 is a confusion matrix generated when scene 15 image set is classified. As can be seen from Fig. 6, the classification rate of this method is 86.75% on scene 13 image and 83.12% on scene 15 image. As can be seen from Fig. 7, when the number of training images is 20, the classification

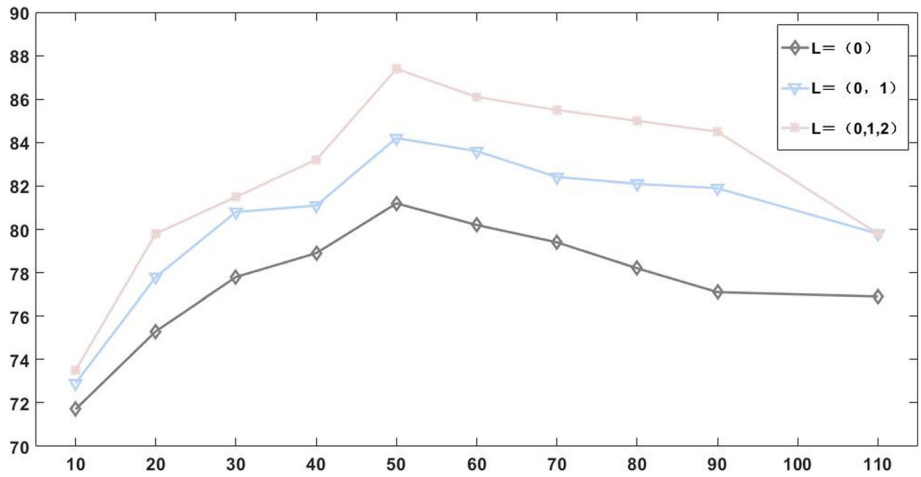

Fig. 5 The number of potential topics

accuracy of this paper is 2.1% higher than that of ScSPM, and when the number of training images is 40, its classification performance is 3.1% higher than that of ScSPM. Compared with other methods, the classification performance of the proposed method is improved significantly. Experimental results show the effectiveness and robustness of the proposed image classification method based on sparse coding multi-scale spatial latent semantic analysis.

5 Conclusions

The key problem of image scene classification is how to bridge the "semantic gap" between the underlying features and the high-level semantics. It is an important research idea to solve the core problem of scene classification by extracting the local invariant features of images and constructing the local semantic concept representation of images. Sparse coding theory is introduced into the study of image local semantic concept representation and has achieved high accuracy in image scene classification. However, the existing sparse coding models aim at minimizing signal reconstruction error. For image scene classification, it is more important to find a discriminant representation than to minimize the reconstruction error. Therefore, this paper proposes an image classification method based on sparse coding multi-scale spatial latent semantic analysis. Spatial pyramid matching of image segmentation is used to extract the spatial position information of the target, and feature soft quantization based on sparse coding is used to form a co-occurrence matrix, which improves the accuracy of the original feature representation. Finally, the PLSA model is used to mine the local latent semantic information, and each local semantic information is

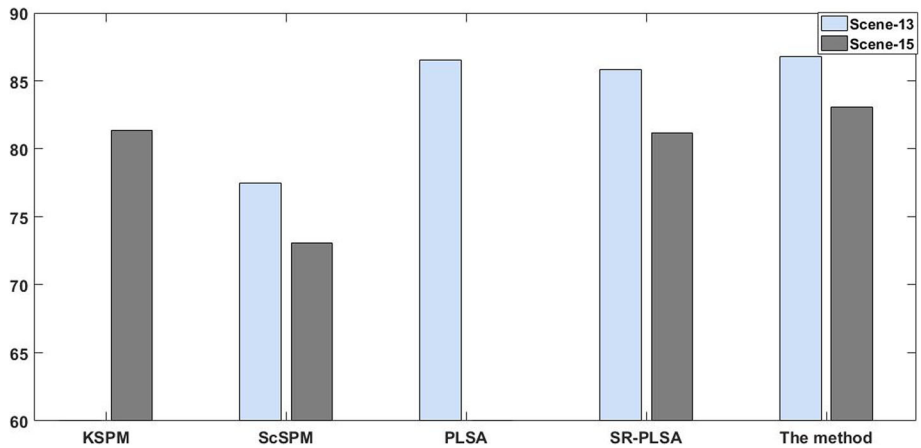

Fig. 6 Comparison of classification accuracy of this method and other classification methods on scene datasets (%)

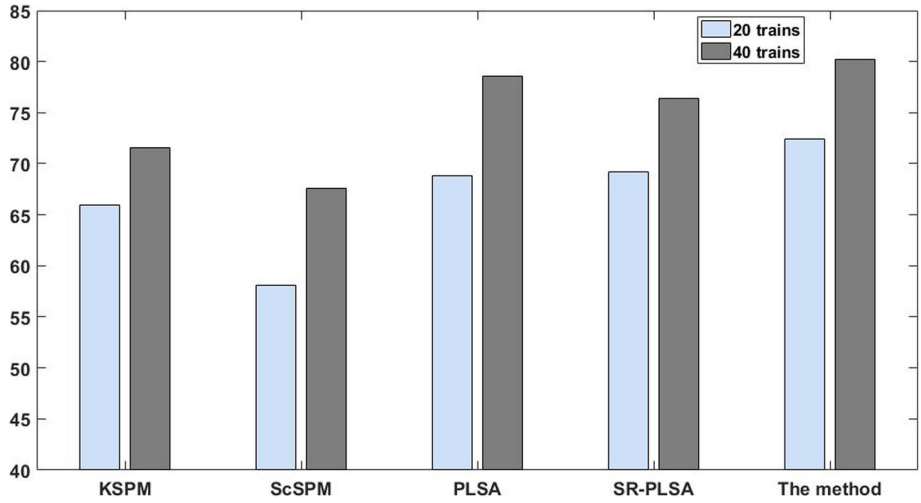

Fig. 7 Comparison of classification accuracy of this method and other classification methods on Caltech-101 database (%)

concatenated to obtain the image multi-scale spatial latent semantic information. Experimental results show that the proposed method has higher classification accuracy than the existing better image classification methods, and the three modules of spatial pyramid matching, sparse coding to construct a co-occurrence matrix, and PLSA dimensionality reduction are indispensable in this method, so that the image can be more accurately represented and the performance of image classification can be improved together.

Abbreviations

LDA: Latent Dirichlet allocation; LLC: Logical link control; NSLLC: Non-negative sparse local linear coding; PLSA: Probabilistic latent semantic analysis; SC: Sparse coding; ScSPM: Sparse coding for image classification; SIFT: Scale-invariant feature transform; SPM: Spatial Pyramid Matching; SVM: Support vector machine

Acknowledgements

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Funding

This work was supported by Shenzhen Science and Technology Program (JCYJ20170306095849825, JCYJ20170306095735097), Cultivation Project of Shenzhen Institute of Information Technology (ZY201715).

Availability of data and materials

Please contact author for data requests.

Authors' contributions

The author, TH, took part in the discussion of the work described in this paper. He wrote all versions of the paper, did experiments of it, and revised the paper. The author read and approved the final manuscript.

Authors' information

Tao He was born in Jingxi, Guangxi, P.R. China, in 1973. He graduated from Shanghai University and received the doctor of computer science in 2009. Now, He works in Shenzhen Institute of Information Technology. His research interests include software engineering and software formalization.

Competing interests

The author declares that he has no competing interests.

Publisher's Note

Received: 21 October 2018 Accepted: 15 January 2019
Published online: 08 February 2019

References

1. B. Pan, Z. Shi, X. Xu, R-VCANet: a new deep-learning-based hyperspectral image classification method. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **10**(5), 1975–1986 (2017)
2. W.U. Jing-Hui, L.B. Tang, B.J. Zhao, et al., Visual dictionary and online multi-instance learning based object tracking. Syst. Eng. Electro. **37**(2), 428–435 (2015)
3. C. Hentschel, S. Stober, A. Rnberger, et al., in *Adaptive multimedia retrieval*. Automatic image annotation using a visual dictionary based on reliable image segmentation (2016), pp. 45–56
4. BC. Chen, YY. Chen, YH. Kuo, et al. Scalable Face Track Retrieval in Video Archives Using Bag-of-Faces Sparse Representation[J]. IEEE Transactions on Circuits & Systems for Video Technology (2017), **27**(7), 1595–1603
5. M. Kumar, Y.H. Mao, Y.H. Wang, T.R. Qiu, C. Yang, W.P. Zhang, Fuzzy theoretic approach to signals and systems: static systems. Inform. Sci. **418**, 668–702 (2017)
6. M. Woodhouse, P.R. Worsley, D. Voegeli, et al., The physiological response of soft tissue to periodic repositioning as a strategy for pressure ulcer prevention. Clin. Biomech. **30**(2), 166–174 (2015)
7. S. Gu, W. Zuo, Q. Xie, et al., in *IEEE International Conference on Computer Vision*. Convolutional sparse coding for image super-resolution (IEEE Computer Society, 2015), pp. 1823–1831
8. C. Bao, H. Ji, Y. Quan, et al., Dictionary learning for sparse coding: algorithms and analysis. IEEE Trans. Pattern Anal. Mach. Intell. **38**(7), 1356–1369 (2015)
9. X. Zhu, X. Li, S. Zhang, et al., Robust joint graph sparse coding for unsupervised spectral feature selection. IEEE Trans. Neural. Netw. **28**(6), 1263–1275 (2017)
10. W. Yang, K. Ricanek, F. Shen, Image classification using local linear regression. Neural Comput. & Applic. **25**(7–8), 1913–1920 (2014)
11. X. Peng, R. Yan, B. Zhao, et al., Fast low rank representation based spatial pyramid matching for image classification. Knowl.-Based Syst. **90**(C), 14–22 (2015)
12. S. Rousseau, P. Chainais, C. Garnier, in *IEEE International Conference on Image Processing*. Dictionary learning for a sparse appearance model in visual tracking (IEEE, 2015), pp. 4506–4510
13. Z.Q. Zhao, H.F. Ji, J. Gao, D.H. Hu, X.D. Wu, Sparse coding based multi-scale spatial latent semantic analysis for image classification. Chin. J. Comput. (2014)
14. Z. Zhao, L. Jiao, J. Zhao, J. Gu, J. Zhao, Discriminant deep belief network for high-resolution sar image classification. Pattern Recogn. **61**, 686–701 (2017)

15. S. Zhang, Z. Wei, Y. Wang, T. Liao, Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary. Futur. Gener. Comp. Syst. **81**(395–403) (2018)
16. P. Li, G. Song, H. Ji, J. Zhao, C. Wang, J. Wu, *A supply restoration method of distribution system based on Soft Open Point. Innovative Smart Grid Technologies - Asia* (IEEE, 2016), pp. 535–539
17. S. Zhang, W. Liu, X.L. Deng, Z. Xu, Kim-Kwang Raymond Choo, Micro-blog topic recommendation based on knowledge flow and user selection. J. Comput. Science **26**(512–521) (2018)
18. D. Weinshall, D. Hanukaev, G. Levi, in *International Conference on Machine Learning*. LDA topic model with soft assignment of descriptors to words (2013), pp. 711–719
19. Z. Zhang, W.A. Huo, Topic text network construction method based on PL-LDA model. Complex systems & complexity. Science **14**(1), 52–57 (2017) and 110
20. P. Karmakar, S.W. Teng, G. Lu, et al., in *International Conference on Digital Image Computing: Techniques and Applications*. Rotation invariant spatial pyramid matching for image classification (IEEE, 2016), pp. 1–8
21. J. Shi, Y. Li, J. Zhu, et al., Joint sparse coding based spatial pyramid matching for classification of color medical image. Comput. Med. Imaging Graph. **41**(1), 61–66 (2015)
22. Y. Gao, K. Katagishi, *Improved spatial pyramid matching for sports image classification*, IEEE Tenth International Conference on Semantic Computing (IEEE, 2016), pp. 32–38
23. H. Ni, Z. Guo, B. Huang, in *International Conference on Service Science*. Patent image classification using local-constrained linear coding and spatial pyramid matching (IEEE, 2015), pp. 28–31
24. M. Wang, Y. Ming, Q. Liu, et al., in *International Congress on Image and Signal Processing, Biomedical Engineering and Informatics*. Similarity search for image retrieval via local-constrained linear coding (2017), pp. 1–6
25. X. Li, L. Gao, X. Xing, et al., Kernel based latent semantic sparse hashing for large-scale retrieval from heterogeneous data sources. Neurocomputing **253**, 89–96 (2017)
26. L. Zhuang, A.Y. Yang, Z. Zhou, et al., Single-sample face cecognition with image corruption and misalignment via sparse illumination transfer. Int. J. Comput. Vis. **114**(2–3), 272–287 (2015)
27. S. Gao, K. Jia, L. Zhuang, et al., Neither global nor local: regularized patch-based representation for single sample per person face recognition. Int. J. Comput. Vis. **111**(3), 365–383 (2015)