Neurocomputing 142 (2014) 282–290

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Adaptive spatial partition learning for image classification

Bingyuan Liu, Jing Liu*, Hanqing Lu

National Laboratory of Pattern Recognition, CASIA, Beijing 100190, China

ARTICLE INFO

Article history:
Received 25 November 2013
Received in revised form
21 February 2014
Accepted 24 March 2014
Communicated by Qingshan Liu
Available online 23 May 2014

Keywords:
Image classification
Spatial information
Group sparsity

ABSTRACT

Spatial Pyramid Matching is a successful extension of bag-of-feature model to embed spatial information of local features, in which the image is divided into a sequence of increasingly finer girds, and the grids are taken as uniform spatial partitions in ad-hoc manner without any theoretical motivation. Obviously, the uniform spatial partition cannot adapt to different spatial distribution across image categories. To this end, we propose a data-driven approach to adaptively learn the discriminative spatial partitions corresponding to each class, and explore them for image classification. First, a set of over-complete spatial partitions covering kinds of spatial distribution of local features are created in a flexible manner, and we concatenate the feature representations of each partitioned region. Then we adopt a discriminative learning formulation with the group sparse constraint to find a sparse mapping from the feature representation to the label space. To further enhance the robustness of the model, we compress the feature representation by removing the dimensions corresponding to those unimportant partitioned regions, and explore the compressed representation to generate a multi-region matching kernel prepared to train a one-versus-others SVM classifier. The experiments on three object datasets (*i.e.* Caltech-101, Caltech-256, Pascal VOC 2007), and one scene dataset (*i.e.* 15-Scenes) demonstrate the effectiveness of our proposed method.

1. Introduction

In recent years, the bag-of-feature (BoF) model [5] becomes very popular in image classification systems. The BoF model starts from well-engineered local features, such as SIFT [18], HOG [6] and color invariant descriptors [25], quantizes them into distinct visual words, and then computes a histogram representation as the image representation. The BoF-based representation describes an image as an orderless collection of local features, while the spatial layout of the features is completely neglected.

To overcome this problem, one popular extension, called as Spatial Pyramid Matching (SPM) [15], has been shown effective for image representation. It requires to first partition each image into a fixed sequence of increasingly finer uniform grids (*e.g.* 1×1, 2×2, 4×4), and then concatenates the BoF features in each grids forming a high dimensional image feature, while the grids in the same level are equally treated for concatenation. However, images in different classes often have different spatial distribution about target and background context, contributing different discriminative abilities to recognize a certain category. For instance, as shown in Fig. 1, the airplanes tend to locate in the middle of the image and the shown spatial partitioned regions (learned by the proposed method) separate the object and background properly, providing more reasonable and semantical spatial information than traditional SPM. In these cases which are common for natural images, the uniform spatial partitions and their equal treatment in SPM fail to reflect reasonable spatial information. We believe that the optimal spatial partition for classification should be learned with some discriminative priors, in order to segment an image into some semantically meaningful regions and crudely indicate the discriminative object regions.

To address above issues, this paper proposes a data-driven approach to learn the discriminative spatial partitions adaptable to each image class and explore them for image classification. In our work, we first create various spatial partitions of images as many as possible, and concatenate the feature of each partitioned region as the image representation. Second, we attempt to train a linear classifier by sparsely mapping the concatenated representations of samples into the label space. In particular, we deem that only a few partitioned regions among all are helpful to image classification and present a group sparse constrained discriminative formulation, in which the feature dimensions corresponding to the same partitioned region are defined as a group. Furthermore, we adopt a leave-one-out scheme to measure the importance of each partitioned region, *i.e.* calculating the training error increase after neglecting the features corresponding to one partitioned region, and consider the regions with the largest increase as the important ones. We compress the concatenated representation by only

* Corresponding author. Tel.: +86 10 82544507; fax: +86 10 82544594.
E-mail address: jliu@nlpr.ia.ac.cn (J. Liu).

http://dx.doi.org/10.1016/j.neucom.2014.03.057
0925-2312/

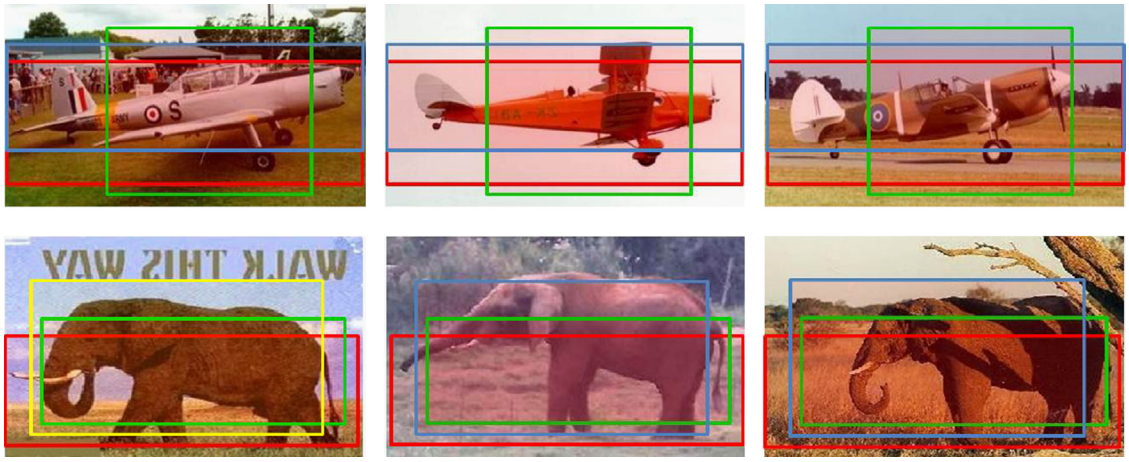

Fig. 1. Discriminative spatial partitions learned by our method, which provide better spatial information than SPM.

remaining the dimensions corresponding to the most important regions and explore the compressed representation to estimate a multi-region matching kernel for a strong SVM classifier to each class. We evaluate our algorithm on three challenging object datasets (*i.e.* Caltech-101, Caltech-256, and Pascal VOC 2007), and one scene dataset (*i.e.* 15-Scenes), showing the effectiveness of our method in comparison with some related works.

The rest of the paper is organized as follows. Section 2 reviews the related work of BoF models with spatial information. In Section 3, we elaborate our proposed method of adaptive spatial partition learning for image classification. The experimental evaluation is given in Section 4, and we conclude in Section 5.

2. Related work

The BoF model, although is simple and directly borrowed from text retrieval community, has been proven useful and effective to represent an image for many computer vision tasks, such as object recognition [24], scene classification [2], image annotation and retrieval [17]. The standard BoF extracts a set of local patch descriptors, assigns each descriptor to the closest entry in a visual codebook which is learned offline by clustering a large sampling set of descriptors with K-means and then pool them to an image-level histogram signature. However, the ignorance of spatial information hinders its final performance. To overcome the limitation, many subsequent researches to incorporate spatial information have been done from the following two directions.

One direction is to incorporate the local spatial layout in image, *i.e.*, the relative positions or pairwise positions of the local features. Savarese et al. [26] explore the combination of correlograms and visual words to represent spatially neighboring image regions. In [16], an efficient feature selection method based on boosting is proposed to mine high-order spatial features, while Morioka and Satoh [20] propose to jointly cluster feature space to build a compact local pairwise codebook capturing correlation between local descriptors and in [21] incorporate the spatial orders of local features. In [23], a data mining method is proposed to automatically find spatial configurations of local features occurring frequently on instances of a given object class.

Since images often have spatial preferences, another direction is to incorporate global spatial layout property, *i.e.*, the absolute positions in image, which is also our focus in this paper. Lazebnik et al. [15] pioneer the direction of exploiting spatial layout property and propose SPM. In SPM, the image is divided into uniform grids at different scales (*e.g.* $1\times 1, 2\times 2, 4\times 4$), and the features are concatenated over all cells. Yang et al. [31] and Wang et al. [30] show that incorporating sparse coding or locality-constrained coding into the SPM model improves the performance. More recently, the combinations of SPM with super vector [33] and fisher vector [22] models are demonstrated effective to obtain a good representation. Different from SPM, Krapac et al. [14] propose to encode the spatial layout by Gaussian mixture model using the Fisher kernel framework. In [3], local features of an image are first projected to different directions creating a series of ordered bag-of-features.

Several parameters in SPM model, such as the number of pyramid levels and the weights of the grid at each level, are chosen in an ad-hoc manner without any optimization [15]. Thus the model is not adaptable to different situations and the performance is highly dependent on the experiences and datasets. To address this issue, Harada et al. [11] propose to form the image feature as a weighted sum of semi-local features over all pyramid levels and the weights are automatically selected to maximize a discriminative power. To design better spatial partition, Sharma and Jurie [28] define a space of grids where each grid is obtained by a series of recursive axis aligned splits of cells and propose to learn the spatial partition in a maximum margin formulation. In this paper, we both consider the weights of different regions and the spatial partition style. The most related work to ours is [13], which adopts the idea of over-complete and formulate the problem in a multi-class fashion with sparse regularization for feature selection. However different categories actually have different spatial distribution, thus we adopt the idea of group sparsity to adaptively learn a class-specific spatial partition style and then explore a semantically compression and multi-region matching kernel to classify each category.

3. Adaptive spatial partition learning

In this section, we describe the details of our proposed framework for image categorization, which is shown in Fig. 2.

3.1. Over-complete spatial partition

As shown in Fig. 2, starting with an input image I_n, we densely extract local features (*e.g.* SIFT or HOG) and encode the features by learned codebook. Instead of spatial pyramid partition, we create an over-complete spatial partition set to compile the spatial information, and use the spatial pooling function to concatenate representations of all the regions.

Spatial pyramid partition plays an important role in many state-of-the-art image classification systems. They adopt an increasingly

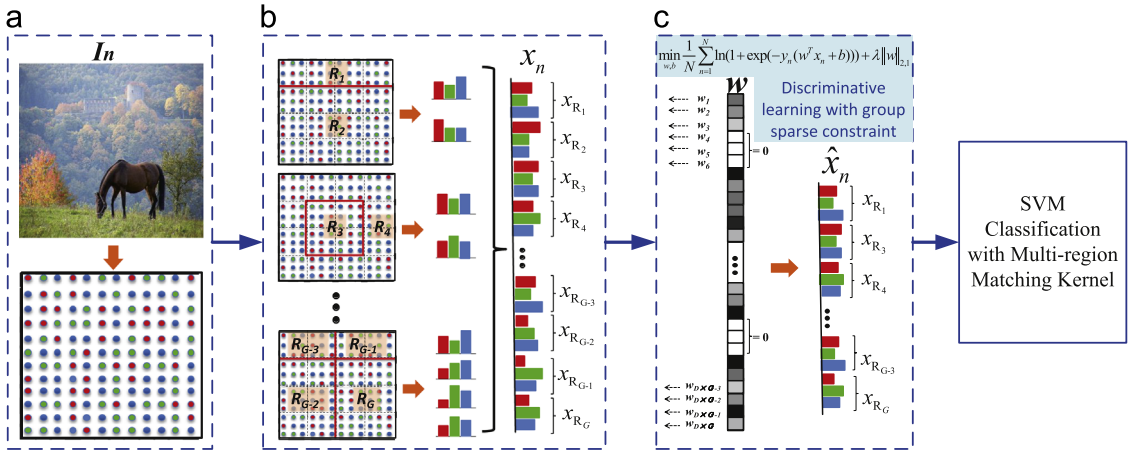

Fig. 2. Our image classification pipeline. See Section 3 for details.

finer uniform grids (e.g. $1\times 1, 2\times 2, 4\times 4$). Obviously, the uniform spatial partition cannot adapt to different spatial distribution across image categories. In order to overcome this constraint, we propose to construct the spatial partitions in a more flexible scheme, involving as many geometric properties of the local features as possible. We think an ideal partition should be able to divide the image into some semantic regions, *i.e.* object region and background context region.

First, we employ randomly distributed horizontal and vertical grids to divide the image into rectangular grids (the dotted grids as shown in Fig. 2). These grids are considered as the candidate grids to generate a certain kind of spatial partition. Then a type of spatial partition is created by randomly choosing a subset of the candidate grids (the solid grids as shown in Fig. 2). By covering all the possible combination of the candidate grids, an over-complete set of spatial partition is established. In this way, we are able to incorporate various spatial information of images as many as possible. We use R_i to denote the i-th partitioned region and G to denote the number of all the partitioned regions.

Note that our method is also feasible by providing more kinds of spatial partitions, *e.g.* circle partitions. For simplicity, we only apply the straight lines to establish the over-complete spatial partition set in our implementation. Obviously, the spatial partitioned regions created in this way are highly redundant and only a small subset of the spatial partition is discriminative and proper to encode the common spatial layout property.

3.2. Discriminative partition learning with structure regularization

By the method described above, we get G spatial partitioned regions to represent possible spatial distribution. Then we obtain the vector x representing the image by concatenating the BoF feature of each spatial partitioned region: $x=[x_{R_1}, x_{R_2}, \ldots, x_{R_G}]$, where $x_{R_i}(i=1,2,\ldots,G)$ denotes the BoF feature of the R_i region. If the size of the visual codebook $B=[b_1, b_2, \ldots, b_D]$ is D, the dimension of x is $D\times G$. Since both the visual codebook and spatial partitioned regions are created in an over-complete scheme, the resulting features are usually very high-dimensional. It is conventional to train a linear classifier using the high-dimensional feature x above, but the performance is very weak because of the high redundancy. Thus we propose to adaptively learn the discriminative spatial partitions in a discriminative formulation with group sparse regularization, while the discrimination of each partition is measured by its sparse parameters.

In this paper, we present our model in the context of linear binary classifier with a one-versus-others classification fashion for each class. Given $X=\{x_1, x_2, \ldots, x_N\}$ the learning of a linear classifier leads to the following optimization problem:

$$\min_{w,b} \frac{1}{N}\sum_{n=1}^{N} l(w^T x_n + b, y_n) + \lambda R(w) \tag{1}$$

where vector w and scalar b are the parameters to be estimated, x_n is the feature vector of the n-th sample, $y_n \in \{-1, 1\}$ is the label of the n-th sample, $l(w^T x_n + b, y_n)$ is a certain non-negative convex loss, N is the number of training images, $R(w)$ is a regularizer and $\lambda \in R$ is the regularization coefficient. We choose the binomial negative log likelihood as the loss function:

$$l(w^T x_n + b, y_n) = \ln(1+\exp(-y_n(w^T x_n + b))) \tag{2}$$

which leads to the logistic regression classifier.

In addition, we expect the classifier to select the most discriminative spatial partitions because of the high redundancy. Noted that only a few partitioned regions among all are discriminative and helpful for image classification, which inspires us to perform semantical compression with the group sparsity prior. Recent analysis and application of the mixed norm regularization [1,27,4] show that under certain conditions the sparse coefficient vector w enjoys the group sparsity property, encouraging the content-based structured feature selection in high-dimensional feature space. Thus we adopt the idea of structured sparsity [32], and train the binary linear classifier via the following optimization problem:

$$\min_{w,b} \frac{1}{N}\sum_{n=1}^{N} \ln(1+\exp(-y_n(w^T x_n + b))) + \lambda \|w\|_{2,1} \tag{3}$$

where $\|w\|_{2,1}$ denotes l_2/l_1 norm regularizer, incorporating group sparsity property . We set the feature groups here as the dimensions corresponding to a certain partitioned region (x_{R_i}), encouraging the model to mine and select discriminative spatial partitions. The number of the groups is G and the dimension of features in each group is D.

The formulation of the l_2/l_1 mixed-norm regularizer is

$$\|w\|_{2,1} = \sum_{i=1}^{G} \|w_{R_i}\|_2 \tag{4}$$

where w_{R_i} is the i-th group of parameters corresponding to the i-th spatial region R_i. This motivates dimensions in the same group to be jointly zero. Thus the optimization procedure tends to select a much smaller but more discriminative subset within the over-complete representations. Beyond the regular l_1 norm regularizer, the sparsity is now imposed on spatial region level rather than merely on feature level.

Although the optimization problem for l_2/l_1 regularized logistic regression is convex, the non-smooth penalty function makes the

optimization highly nontrivial. We adopt the efficient algorithm proposed in [19,12]. The dual of the proximal problem associated with the norm can be reformulated as a quadratic min-cost flow problem, which is able to be efficiently computed in polynomial time. The algorithm is very suitable for our problem because of its efficiency and ability to scale up to millions of variables. Also a well implemented open toolbox called SPAMS (SPArse Modeling Software)[1] based on the algorithm developed by Julien Mairal is convenient and effective to solve our problem.

3.3. Feature compression

The linear classification trained by Eq. (3) can be directly utilized to classify a given image. However, this classifier is weak because of the high redundancy of our representation. Thus we propose to compress feature dimensions according to the w learned by Eq. (3), leading to a semantically more compact and discriminative representation.

The learned parameters w in Eq. (3) can be regarded as importance weights of dimensions corresponding to each spatial region. Based on the idea, we adopt a leave-one-out scheme to measure the importance of each partitioned region. The importance value of a particular partitioned region is measured as the increase of the training error with the dimensions corresponding to the region removed, while the training error of our model over all training data given w is defined as

$$Error_0 = \frac{1}{N}\sum_{n=1}^{N} \ln(1+\exp(-y_n(w^T x_n + b))) \tag{5}$$

where w and b are the solutions of Eq. (3). The error of neglecting the dimensions corresponding to the j-th partitioned region, *i.e.*, setting $w_{R_j} = 0$, is denoted as $Error_j$. Then, the importance of the j-th region is calculated by measuring the training error increase after neglecting the region dimensions as

$$S_j = \frac{Error_j - Error_0}{Error_0} \tag{6}$$

The larger S_j indicates the neglected dimensions of the j-th region are more important and discriminative for the category task. We then perform the feature compression on x by only remaining the dimensions corresponding to the regions with positive spatial importance S_j which is common for the particular class. By this way, the feature compression is performed semantically to remove redundant dimensions adapting to the particular class.

3.4. Multi-region matching kernel for SVM classification

In traditional SPM, the grids in the same level are equally treated to get the final kernel. However, the weights of different spatial regions should be different (usually the object region should be paid more attention than background context region). To this end, the values of $S_j (j = 1, 2, \ldots, G)$ are employed to roughly indicate the importance of different regions and define a multi-region matching kernel with the weighted sum of the separate region kernels, prepared to train a SVM classifier:

$$\mathcal{K}(\hat{x}^1, \hat{x}^2) = \sum_{m=1}^{M} S_m \cdot K(\hat{x}^1_{R_m}, \hat{x}^2_{R_m}) \tag{7}$$

where $\hat{x}^1$ and $\hat{x}^2$ are feature vectors of two images after compression, $\hat{x}^1_{R_m}$ and $\hat{x}^2_{R_m}$ represent the corresponding dimensions of the m-th region, M is the remaining number of regions and S_m is the importance weight of the region. The kernel K may be any kernel function[2] and a normalization operation is usually needed. With the multi-region matching kernel, we turn to train a one-versus-others classifier for each class to further enhance the classification performance.

4. Experiments and results

In the experiments, we implement and compare our method mainly with KSPM [15] (the popular kernel SPM), ScSPM [31] (the method that uses sparse coding and linear SVM) and some other works considering discriminative spatial information [28,13]. Although our method can be combined with diverse coding and pooling algorithms, we apply the sparse coding and max pooling strategy for fair comparison. We use a single local descriptor type, the popular SIFT descriptor, by densely extracting local patches of 16×16 pixels computed over a grid with spacing of 8 pixels and then the local features are encoded with sparse coding. For all the experiments, we fix the codebook size as 1024. Considering simplicity and efficiency, we construct our over-complete spatial regions using three horizontal lines and three vertical lines. We also randomly generate some additional spatial grids to incorporate more flexible spatial information. Totally, a set of 100 different spatial partitioned regions is employed as our over-complete spatial partition set. The adaptive spatial partition learning model is then trained to obtain the parameters w for each class. Finally, we train the SVM classifier by the compressed features with the multi-region matching kernel. The trade-off parameters to the group sparsity regularization term and the SVM regularization term are chosen via 5-fold cross validation on the training data.

We demonstrate the effectiveness of our method on three diverse object databases, Caltech-101, Caltech-256, Pascal VOC 2007 and one scene database, 15-Scenes. Following the common benchmarking procedures, we repeat the experimental process by 5 times with different randomly selected training and testing images to obtain reliable results. The final results are reported by the mean and standard deviation of the classification rates. The extensive comparisons and analysis are presented in the following subsections.

4.1. Results on Caltech-101 dataset

We start our experiments with an in-depth analysis of our method on the dataset of Caltech-101 [8]. The Caltech-101 dataset contains 9144 images totally from 102 different categories, including 101 object categories and 1 additional background category, with high shape variability. The number of images per category varies from 31 to 800. We follow the common experiment setup for Caltech-101, training on 15 and 30 images per category and testing on the rest.

The performance comparison results are shown in Table 1. The ASPL (adaptive spatial partition learning) in the table denotes the method that directly use the spatial partition learning model trained by Eq. (3) as the final classifier and ASPL+SVM denotes our method described in Section 3.4 of training a stronger SVM classifier using the compressed feature with multi-region matching kernel. It is indicated that the best result is obtained by using the final compressed representations with multi-region matching kernel, which outperforms the traditional method, *i.e.* ScSPM [31], by a margin of roughly 4% according to our implementation. To evaluate the effect of our spatial partition learning model, we compare our method with the scheme of randomly selecting

[1] The toolbox is available at http://spams-devel.gforge.inria.fr/.

[2] In experiments, we use χ^2 kernel for 15-Scenes dataset, and linear kernel for Caltech-101, Caltech-256 and Pascal VOC 2007 datasets.

the same number of spatial regions from the over-complete set instead of the spatial partition learning process, which is denoted as Rand+SVM in Table 1. It shows that our method outperforms the randomly selecting scheme, indicating that our method does help to select more important and discriminative spatial partitions for image classification. Our method also performs better than [13], in which they perform feature selection in a multi-class fashion to learn discriminative spatial partition for the whole categories. In our method we adaptively learn category-specific spatial partitions because different categories obviously have different spatial distributions, and the feature compression and multi-region matching procedure also improve the performance. In Fig. 3 , we show some examples of the classes that our method increase most and some bad examples that our method decrease the performance. It is shown that our method obviously improve the classification accuracy under the condition that the objects have strong spatial prior nature.

In our method, there is one free parameter λ in Eq. (3) needed to determine when we learn the discriminative spatial partitions, which controls the sparsity of the solution. The bigger λ is, more sparse the solution will be, thus in the feature compression process described in Section 3.3 more spatial partitioned regions will be removed, leading to more compact final image representations. Fig. 4 shows the effect of the tradeoff coefficient λ by investigating the classification accuracy with different values of λ. The accuracies are bad when the parameter is small or large, since small λ may cause redundancy of the features, while large λ may bring about loss of useful information. Thus the best performance was achieved with well balanced parameter. Empirically, we found that keeping the tradeoff term λ to be around 0.5 yields good results. In the traditional spatial pyramid representation, the final image representation is in the dimension of 21,504 (the size of the codebook is 1024 and the spatial partition setup is 1×1, 2×2, 4×4). In our implementation when λ is 1.5 the dimension of the final feature is 16,384, which is much less than spatial pyramid representation and performs better in the recognition task. This indicates that a more compact and discriminative representation can be obtained by our model, as we better explore the sparse nature of the features. While the best recognition performance is achieved with λ set as 0.5, the final dimension of the image representation is 27,648.

Fig. 5 is an encouraging demonstration to show what we learned by the spatial partition learning method. We also draw the importance maps of the spatial region using the accumulated importance value of the partitioned regions obtained by our method described in Section 3.3, as shown in Fig. 5(c). It shows that the learned spatial partition properly capture the structural information of the classes. For example, the airplanes in the Caltech-101 dataset tend to locate in the middle of a image and our learned spatial regions shown in Fig. 5 capture this property by paying more attention to the middle regions. In the dolphin category, the objects appear mostly in the middle and top of the image, thus these regions are more significant to recognize the dolphin, which is well modeled by our method. It is demonstrated that the spatial importance we get really provide better spatial information than traditional method. However, for some classes, our method has not capture the spatial distribution of the target well because of the large diversity of the object and the unconstraint of the location. Overall, our method provides a more discriminative class adaptive spatial information than traditional SPM and some other related works.

Table 1
Classification rate (%) comparison on Caltech-101.

Algorithms	15 training	30 training
KSPM [15]	56.40	64.6 ± 0.80
ScSPM [31]	67.0 ± 0.45	73.2 ± 0.54
LCC+SPM [30]	65.43	73.44
Jia et al. [13]	–	75.3 ± 0.70
ScSPM	66.38 ± 0.30	72.32 ± 0.35
ASPL	65.04 ± 0.21	70.52 ± 0.20
Rand+SVM	65.93 ± 0.55	71.39 ± 0.84
ASPL+SVM	$\mathbf{69.51\pm0.30}$	$\mathbf{76.72\pm0.38}$

4.2. Results on Caltech-256 dataset

The Caltech-256 [10] dataset is an extension of the Caltech-101. It holds 29,780 images in 256 object categories where the number of images in each category varies from 31 to 800. This dataset is much more challenging as it possesses much higher intra-class variability and higher object location variability compared with

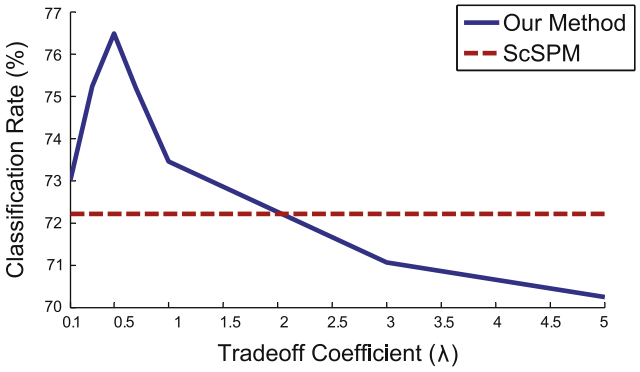

Fig. 4. The classification performance comparisons with varying tradeoff coefficient λ.

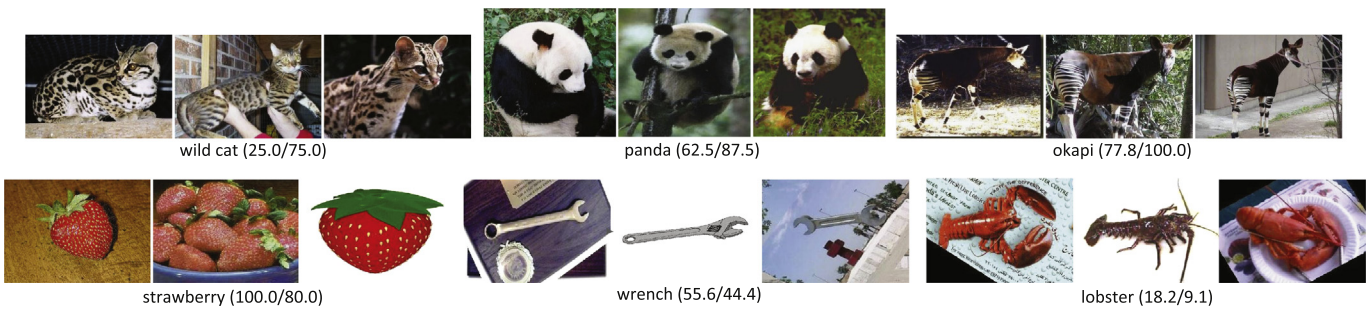

Fig. 3. Examples of the Caltech-101 set. Top: the top 3 categories where our method improves most. Bottom: the 3 categories where our method decreases performance. The numbers in the brackets indicate the classification rate (ScSPM/our method).

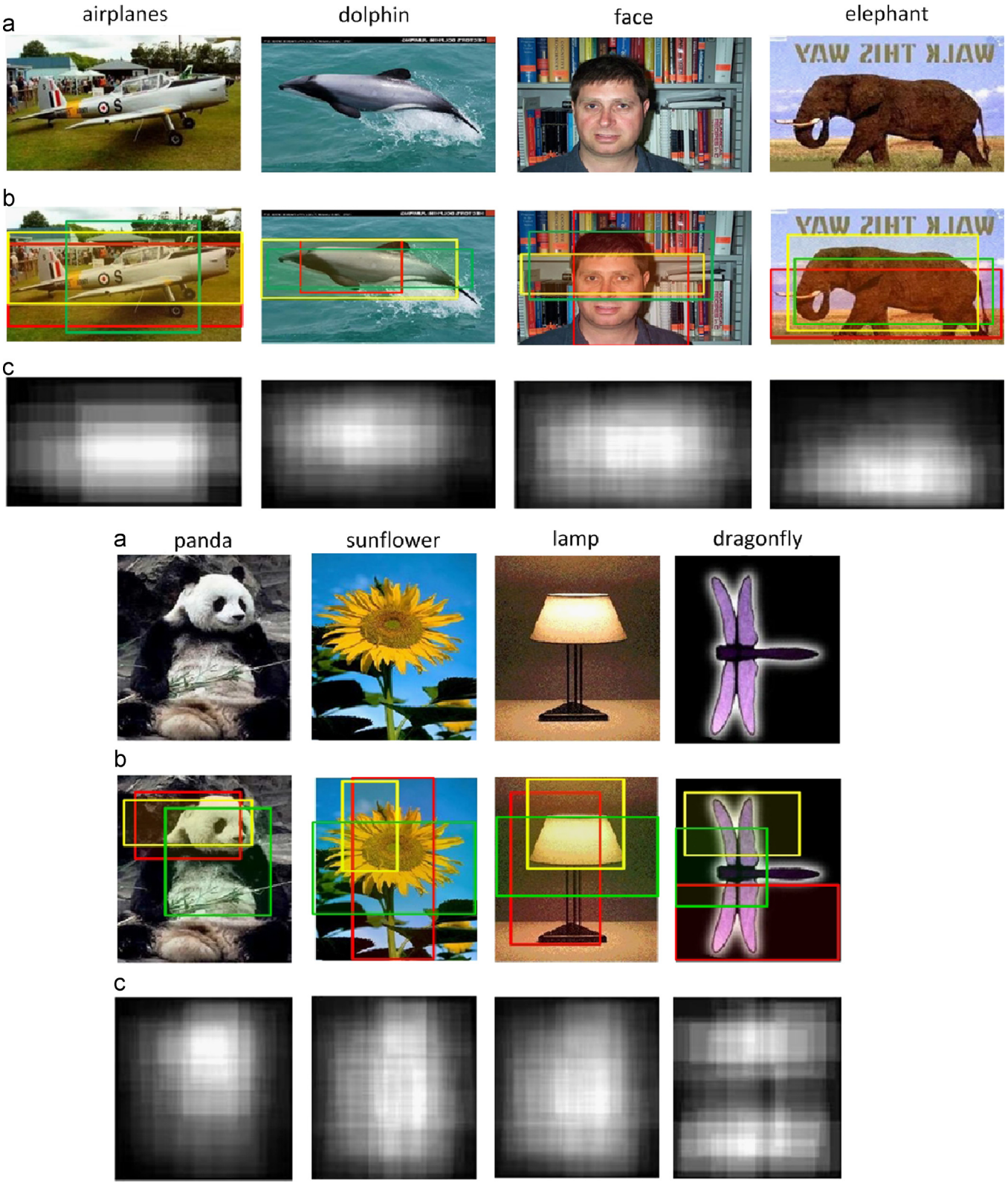

Fig. 5. Demonstrations showing the effectiveness of our method to learn discriminative spatial partition of the Caltech-101 dataset. (a) Image examples for the classes. (b) The top 3 important regions according to the weights learned in our method. (c) Spatial importance maps obtained by our method for the particular class. The lighter, the more important the region is.

Caltech-101. Following the common experiment setup for Caltech-256, we tried our algorithm on 15, 30, 45, and 60 training images per class respectively and tested on the rest.

The performance comparison results are shown in Table 2. In the more challenging dataset than Caltech-101, our ASPL+SVM also consistently leads the performance and outperforms the baseline ScSPM by more than 4 percent for all the cases. In Fig. 6, we show some examples of the classes that our method increases most and some bad examples that our method decreases the performance.

Table 2
Classification rate (%) comparison on Caltech-256.

Algorithms	15 train	30 train	45 train	60 train
KSPM [10]	28.34	34.10	–	–
KC [9]	–	27.17 ± 0.46	–	–
ScSPM [31]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55	40.14 ± 0.91
ScSPM	26.82 ± 0.73	33.14 ± 0.46	36.10 ± 0.60	39.01 ± 0.53
ASPL	26.00 ± 0.40	32.41 ± 0.50	34.03 ± 0.11	37.82 ± 0.15
ASPL+SVM	**30.03 ± 0.30**	**37.65 ± 0.38**	**40.02 ± 0.25**	**44.02 ± 0.21**

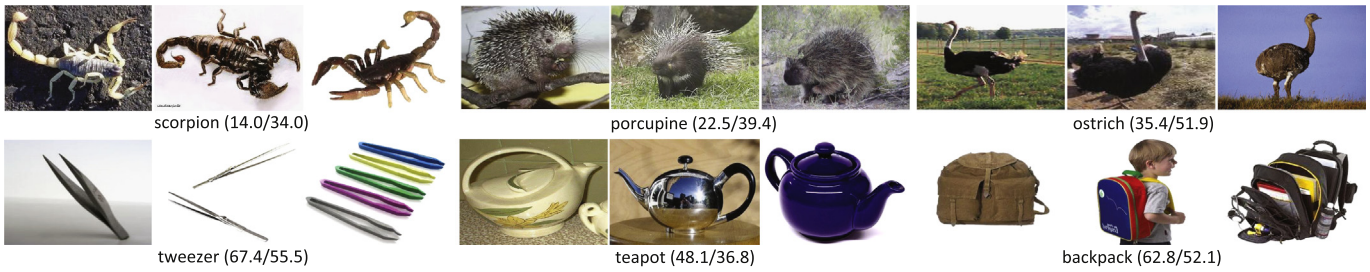

Fig. 6. Examples of the Caltech-256 set. Top: the top 3 categories where our method improves most. Bottom: the 3 categories where our method decreases performance. The numbers in the brackets indicate the classification rate (ScSPM/our method).

4.3. Results on Pascal VOC 2007 dataset

The PASCAL Visual Object Challenge (VOC) datasets are widely used as testbeds for evaluating algorithms for image understanding tasks and provide a common evaluation platform for both object classification and detection. This database is considered to be an extremely challenging one because all the images are daily photos obtained from Flicker where the size, viewing angle, illumination, appearances of objects and their poses vary significantly, with frequent occlusions. The PASCAL VOC 2007 dataset [7] consists of 9963 images from 20 classes, which are divided into "train", "val" and "test" subsets, i.e. 25% for training, 25% for validation and 50% for testing. The classification performance is evaluated using the Average Precision (AP) measure, a standard metric used by PASCAL challenge. It computes the area under the Precision/Recall curve, and the higher the score, the better the performance. We use the train and val sets for training our model and report the mean average precision for the 20 classes on the test set as the performance measure, following the standard protocol of this database. Very high performance has been reported by using multiple features and combining diverse models [29]. For efficiency, here we just adopt one single feature, SIFT, just to evaluate our method.

In Fig. 7, we show our scores for all 20 classes in comparison with our baselines, ScSPM and ASPL. The performance of our method for most of the classes and on an average(43.02 vs. 41.17) is higher than the traditional SPM method. Fig. 8 shows the improvement of our method for each category. It is shown that for some classes, such as sheep, aeroplane and car, the performance increase is larger due to the well captured spatial nature of the objects. However, for some classes, such as dining table and pottedplant, the improvement of our method is limited. This is mainly on account of the highly diversity of the images in the database.

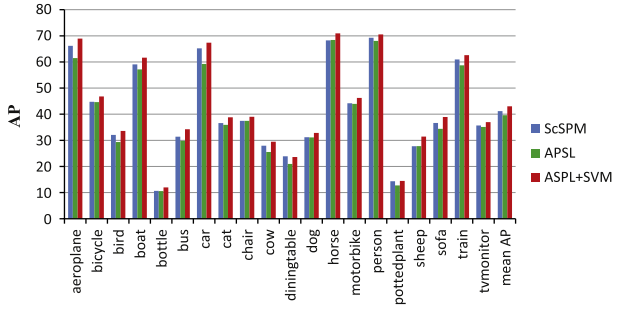

Fig. 7. The AP for all the classes of the VOC 2007 database. We compare our method (ASPL+SVM) with ASPL and traditional SPM.

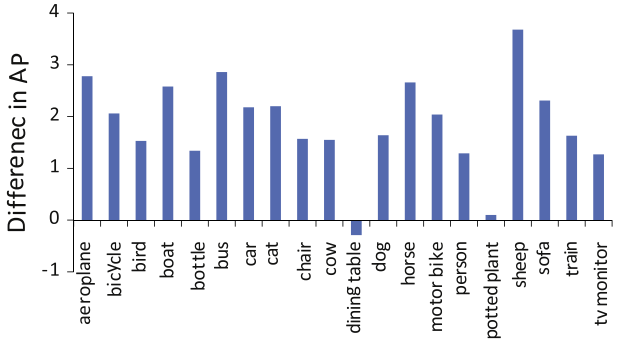

Fig. 8. The difference in AP for all the classes of the VOC 2007 database between our method and the traditional ScSPM.

4.4. Results on 15-Scenes dataset

We finally experiment with a popular scene classification benchmark, 15-Scenes dataset. This dataset includes 15 classes of different scenes (*e.g.* kitchen, coast, highway), containing totally 4485 gray-scale images with the number of each category ranging from 200 to 400. Fig. 9 shows some example images of the 15-Scenes dataset. Following the common experiment setup of the dataset [15,31], we take 100 images per class for training and the rest for testing. The other parameters are transposed from the former experiments.

Table 3 shows the detailed comparison results. It is shown that the ASPL+SVM method achieves better performance than traditional KSPM and ScSPM. The evaluation on the 15-Scenes dataset demonstrates that our method also adapt to the scene categories to improve the classification accuracy beyond the object recognition. Fig. 10 shows the confusion table between the 15 scene categories. Confusion occurs

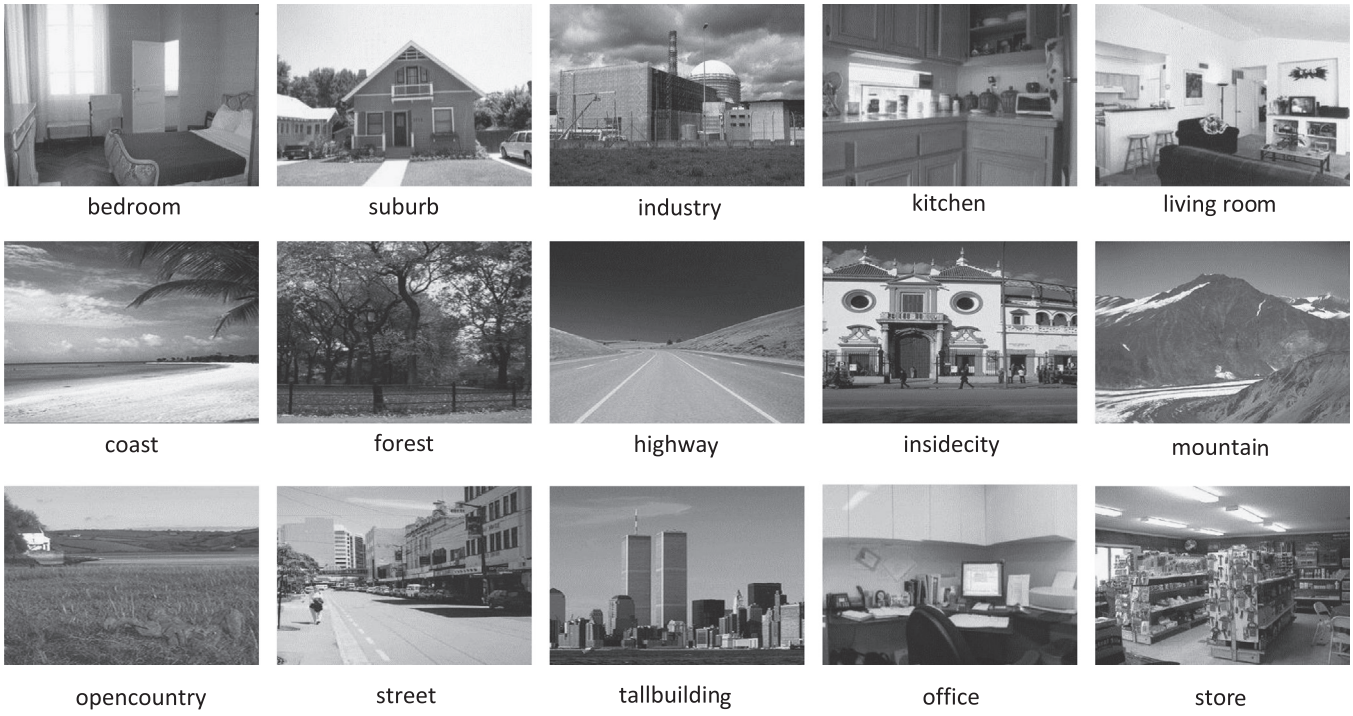

Fig. 9. Example images for the 15-Scenes dataset.

Table 3
Classification rate (%) comparison on 15-Scenes.

Algorithms	Classification rate
KSPM [15]	81.40 ± 0.50
ScSPM [31]	80.28 ± 0.93
Sharma and Jurie [28]	80.10 ± 0.60
ScSPM	79.20 ± 0.53
ASPL	77.85 ± 0.54
Rand+SVM	77.90 ± 0.82
ASPL+SVM	$\mathbf{82.19 \pm 0.10}$

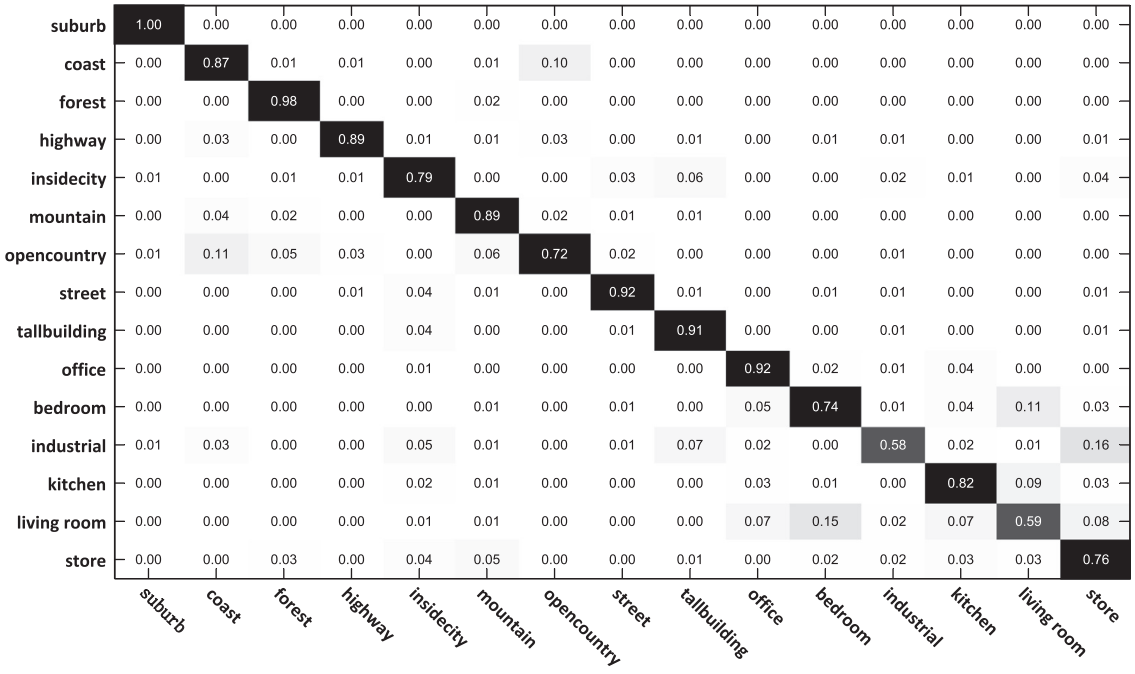

Fig. 10. Confusion table for the 15-Scenes dataset. Average classification rates for individual classes are listed along the diagonal. The entry in the i-th row and j-th column is the percentage of images from class i that are misidentified as class j.

between the indoor classes (*e.g.* bedroom, kitchen, and living room) and also between some natural classes (*e.g.* coast and opencountry).

5. Conclusion

In this paper, we address the issue of the discriminative spatial partition learning and propose a data-driven approach to discover the most discriminative partitioned regions to the particular class. Different from traditional SPM applying manually defined spatial pyramid partitions, the proposed approach constructs a more flexible spatial partition set and adopt the idea of structured sparsity to learn discriminative spatial partitions. Then we propose to measure the importance of the spatial partitioned regions by the learned sparse parameters and train SVM classifier with a multi-region matching kernel. Our method outperforms traditional SPM method on three object datasets (*i.e.* Caltech-101, Caltech-256, Pascal VOC 2007), and one scene dataset (*i.e.* 15-Scenes), and the experiments have shown its effect to adaptively capture the spatial information of the images belonging to different categories.

Acknowledgements

This work was supported by 973 Program (2010CB327905) and National Natural Science Foundation of China (61272329, 61273034, 61202325).

References

[1] S. Bengio, F. Pereira, Y. Singer, D. Strelow, Group sparse coding, in: NIPS, 2009.
[2] A. Bosch, A. Zisserman, X. Munoz, Scene classification using a hybrid generative/discriminative approach, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2008).
[3] Y. Cao, C. Wang, Z. Li, L. Zhang, Spatial bag-of-features, in: CVPR, 2010.
[4] X. Chen, X.T. Yuan, Q. Chen, S. Yan, T.S. Chua, Multi-label visual classification with label exclusive context, in: ICCV, 2011.
[5] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: ECCV 2004 Workshop on Statistical Learning in Computer Vision.
[6] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, 2005.
[7] M. Everingham, L. Van Gool, C. Williams, J. Winn, A. Zisserman, The Pascal Visual Object Classes Challenge 2007 (VOC2007) Results (2007), 2007.
[8] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, Comput. Vis. Image Underst. 106 (2007) 59–70.
[9] J.C. van Gemert, J.M. Geusebroek, C.J. Veenman, A.W.M. Smeulders, Kernel codebooks for scene categorization, in: ECCV, 2008.
[10] G. Griffin, A. Holub, P. Perona, Caltech-256 Object Category Dataset (2007).
[11] T. Harada, Y. Ushiku, Y. Yamashita, Y. Kuniyoshi, Discriminative spatial pyramid, in: CVPR, 2011.
[12] R. Jenatton, J. Mairal, G. Obozinski, F. Bach, Proximal methods for sparse hierarchical dictionary learning, in: ICML, 2010.
[13] Y. Jia, C. Huang, T. Darrell, Beyond spatial pyramids: receptive field learning for pooled image features, in: CVPR, 2012.
[14] J. Krapac, J. Verbeek, F. Jurie, Modeling spatial layout with fisher vectors for image categorization, in: ICCV, 2011.
[15] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: CVPR, 2006.
[16] D. Liu, G. Hua, P. Viola, T. Chen, Integrated feature selection and higher-order spatial feature extraction for object categorization, in: CVPR, 2008.
[17] Y. Liu, D. Zhang, G. Lu, W.Y. Ma, A survey of content-based image retrieval with high-level semantics, Pattern Recognit. 40 (2007) 262–282.
[18] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[19] J. Mairal, R. Jenatton, G. Obozinski, F. Bach, Network flow algorithms for structured sparsity, in: NIPS, 2010.
[20] N. Morioka, S. Satoh, Building compact local pairwise codebook with joint feature space clustering, in: ECCV, 2010.
[21] N. Morioka, S. Satoh, Learning directional local pairwise bases with sparse coding, in: BMVC, 2010.
[22] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: ECCV, 2010.
[23] T. Quack, V. Ferrari, B. Leibe, L.J.V. Gool, Efficient mining of frequent and distinctive feature configurations., in: ICCV, 2007.
[24] P.M. Roth, M. Winter, Survey of Appearance-Based Methods for Object Recognition, Technical Report, Institute for Computer Graphics and Vision, Graz University of Technology, 2008.
[25] K.E.A. van de Sande, T. Gevers, C.G.M. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 1582–1596.
[26] S. Savarese, J. Winn, A. Criminisi, Discriminative object class models of appearance and shape by correlatons, in: CVPR, 2006.
[27] M. Schmidt, K. Murphy, G. Fung, R. Rosales, Structure learning in random fields for heart motion abnormality detection, in: CVPR, 2008.
[28] G. Sharma, F. Jurie, Learning discriminative spatial representation for image classification, in: BMVC, 2011.
[29] Z. Song, Q. Chen, Z. Huang, Y. Hua, S. Yan, Contextualizing object detection and classification, in: CVPR, 2011.
[30] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: CVPR, 2010.
[31] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: CVPR, 2009.
[32] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 68 (2006) 49–67.
[33] X. Zhou, K. Yu, T. Zhang, T.S. Huang, Image classification using super-vector coding of local image descriptors, in: ECCV, 2010.

Bingyuan Liu received his B.S. degree from Zhejiang University (ZJU), Hangzhou, China, in 2010. He is currently a Ph.D. candidate at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include image content analysis and deep representation learning.

Jing Liu received her B.E. degree in 2001 and M.E. degree in 2004 from Shandong University, and her Ph.D. degree from Institute of Automation, Chinese Academy of Sciences in 2008. Currently she is an associate professor in National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her research interests include machine learning, image content analysis and classification, and multimedia information indexing and retrieval.

Hanqing Lu received his B.E. degree in 1982 and his M.E. degree in 1985 from Harbin Institute of Technology, and Ph.D. degree in 1992 from Huazhong University of Sciences and Technology. Currently, he is a deputy director of National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include image and video analysis, medical image processing, and object recognition. He has published more than 200 papers in these fields.