Image Classification Using Sparse Coding and Spatial Pyramid Matching

Xiaofang Wang[1,2], Jun Ma[1], Ming Xu[3]

[1] School of Computer Science and Technology, Shandong University, Jinan 250101, China
[2] Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan 250022, China
[3] Dareway Software co., ltd., Jinan 250101, China
ise_wangxf@ujn.edu.cn, majun@sdu.edu.cn, xm@dareway.com.cn

Abstract - Recently, the Support Vector Machine (SVM) using Spatial Pyramid Matching (SPM) kernel has achieved remarkable successful in image classification. The classification accuracy can be improved further when combining the sparse coding with SPM. However, the existing methods give the same weight of patches of SPM at different levels. Clearly the discriminative powers of SPM at different levels are distinct and there are correlation relationships among the sparse coding bases vectors, which usually have negative influence on the classification accuracy. This paper assigns different weights to the patches at different levels of SPM, and then proposes a new spatial pyramid matching kernel. Furthermore, the Principle Component Analysis (PCA) is employed to reduce the dimension of the feature vectors in order to decrease correlation among vectors and speed up the SVM training process. The preprocessing can enhance the discriminative ability of the new kernel as well. Experiments carried out on Caltech101 and Caltech256 datasets show that the new SPM kernel outperforms the existing methods in terms of the classification accuracy.

Index Terms - Sparse coding, Spatial pyramid matching, Support vector machine, Image classification.

1. Introduction

In recent years, Bag-of-Visual-Words (BoVW) model has been extremely popular in image classification. The method splits an image into several patches using different sampling techniques, and then represents the image as a group of disorder descriptors extracted from local patches. These descriptors are quantized into discrete "visual words", which compose the dictionary of visual words. Through some statistics of all sample patches contained in the image, a compact histogram representation is calculated. Then various classification methods can be adopted to classify images.

Traditional BoVW model consists of four stages: feature extraction, dictionary learning, image representation and image classification. In recent years, many extension works based on traditional BoVW model have been done.

In feature extraction stage, it has been verified that dense sampling strategy outperforms coarse sampling method [1]. The comparison of various local descriptors in [2] has shown that SIFT [3] can achieve the best match performance under different transformation. Later, literature [4] confirmed that SIFT also exhibits excellent performance in object recognition field.

In dictionary learning stage, generative models are proposed in [5-6] which are based on the co-occurrences of visual words, and discriminative models are used in [1,7] instead of traditional unsupervised K-means clustering method to learn the dictionary. In [8], the image features are represented using sparse coding, and a group of basis vectors are obtained by training the image features. These basis vectors constitute the dictionary. For the size of the dictionary, [9] has shown that a larger visual word dictionary performs better than smaller dictionary, and this is further be conformed in [10] where a large visual dictionary obtained by K-means clustering can get better classification performance.

In image representation stage, literature [11] uses the space pyramid matching kernel (KSPM) to model the spatial position of local features, and then the image histogram representation is obtained. A sparse coding based space pyramid matching (ScSPM) method has been proposed in [8], together with max pooling strategy to pool features in various scale and locations in image space to obtain the final image encoding vectors. In the classification stage, support vector machine (SVM) is widely used due to its robustness to high dimensional feature and sparse data, and achieves better classification results in BoVW model. But the selection of SVM kernel function can affect the final classification performance to some extent. In [12], local features are used to design kernel function for the first time and achieve superior classification performance than global color histogram. The kernel functions that have been proved to be effective include histogram intersection kernel [13], generalized histogram intersection kernel [14], chi-square kernel [15], the traditional RBF kernel and linear kernel [8]. The performance of the kernel function varies due to different experimental settings. Overall, the time complexity of nonlinear Mercer kernel is much higher than the linear kernel. Literature [8] proposed a linear kernel based method ScSPM that greatly reduced the time complexity, and achieved better results.

These outreach efforts to BoVW model greatly enrich the study of image classification, among which ScSPM is one of the best classification methods. However, during the space pyramid creating procedure, different levels in space pyramid are not given different weights and there is a certain correlation between different levels, which may have an impact on the classification results.

In this paper, inspired by the traditional KSPM, we propose an improved sparse coding based space pyramid matching algorithm (ScKSPM). Combining with BoVW model, we coding the extracted SIFT features, assign different

weight to sparse coding at different level and design a new spatial pyramid matching kernel for image classification. Since the higher dimension of feature vector can increase the computational complexity and the correlation between feature dimensions will have an impact on the classification results, we perform Principle Component Analysis (PCA) on feature set before taking SVM classification. The PCA process of image features can largely reduce the computation time and remove the correlation between feature dimensions. Experiments on Caltech101/256 and Pascal VOC 2006 dataset show that, the method based on new spatial pyramid matching kernel can achieve higher accuracy in the image classification. It is noteworthy that, the dimension reduction process before performing SVM classification can improve the performance of new proposed kernel and reduce the computation time.

2. Sparse Coding Presentation of Image

Given a set of nature images, in order to get a more precise presentation of image, Olshausen & Field (1996) [16] proposed a method called sparse coding. This method is based on the assumption that every single image *I(x,y)* in the image set can be represented in terms of a linear superposition of (not necessarily orthogonal) basis functions, $\phi(x, y)$:

$$I(x, y) = \sum_i a_i \phi_i(x, y) \tag{1}$$

The image code is determined by the choice of basic functions ϕ_i, and each image has different coefficient vector a.

Combined with the extraction of SIFT descriptor for each image in the image set, we denote the set of SIFT features extracted from the whole image set by *X*, i.e. $X = [x_1, ..., x_M]^T \in \mathbb{R}^{M \times D}$. Hence, the problem of seeking the basic functions can be further quantified as an optimization problem as described below:

$$\min_{A,\Phi} \sum_{m=1}^{M} \left(\left\| x_m - a_m \Phi \right\|^2 + \lambda \left| a_m \right| \right) \tag{2}$$

where x_m represent the *m*th feature in the feature set, Φ is the basis function set, a_m is the coefficient vector corresponding to the *m*th feature and $|a_m|$ denotes the *L1*-norm of vector a_m. Honglak Lee et.al has given an efficient algorithm to solve the above optimization problem in [17].

3. Spatial Pyramid Matching Kernel

In the feature extraction phase, we represent an image as a set of descriptors, and then we can compute a single feature vector based on some statistics of the descriptors codes. In traditional KSPM, all the feature vectors representing the whole image set are quantized into *M* different discrete features using traditional clustering methods, and the quantization is based on the assumption that feature vectors match with each other only when they are of the same type. Then the spatial pyramid is built by split the image into different spatial level. Then combine with the pyramid match kernel [18], various weights are assigned to feature vectors in different spatial level. Put these entire piece together, we obtain the feature vector of the input image.

Yang et.al presented a sparse coding based spatial pyramid matching model which is different from the traditional SPM. They constructed a linear spatial pyramid matching kernel. In the feature extraction phase, a group of sparse coding for the image set is obtained. Based on the traditional SPM process, a spatial pyramid is built. Then the max pooling strategy [19] is adapted to pool features in each level. The final feature vector of the input image is obtained by linearly combine the pooling vectors.

4. An Improved Spatial Pyramid Matching Algorithm Based on Sparse Coding

In this paper, we proposed an improved spatial pyramid matching algorithm based on sparse coding (ScKSPM for short). The model is illustrated in Fig. 1.

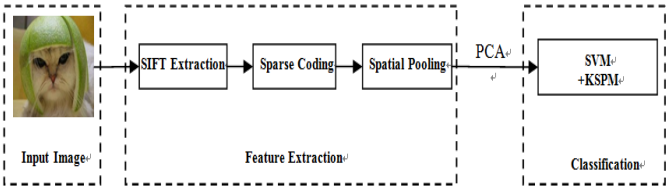

Fig. 1 The model of ScKSPM

Given a input image set, we extract SIFT features for each image using dense sampling method in feature extraction phase, and get the sparse coding dictionary using methods described in section 2. Then we use the max pooling strategy to pool the sparse coding set. Let *X* is the feature vectors and let *A* be the sparse coding matrix obtained by equation 2. Suppose the basis vector set Φ in the training stage is obtained and fixed, we use the pooling function described below to calculate the image feature.

$$z_j = \max\{|a_{1j}|, |a_{2j}|, ..., |a_{Mj}|\} \tag{3}$$

where *z* is the feature of image, z_j is the *j*-th element of *z*, a_{ij} is the matrix element at *i*-th row and *j*-th column of *A*, and *M* is the number of local descriptors in the region.

Similar to the procedure to build traditional spatial pyramid, we split the image space into *l* level and assign various weights to each level, e.g. the weight of *l*-th level is $\frac{1}{2^{L-l}}$, where *L* represents the total number of levels in image space. For different location in different space level, we calculate the pooling feature for each grid. Comparing to calculate the mean value of the feature vectors in each grid, feature vectors calculated by max pooling function can be more robust to local transformation. For the feature vector z_i representing image I_i, we adopt the improved space pyramid matching kernel

$$\kappa^L(z_i, z_j) = \sum_{d=1}^{D} \min(z_i(d), z_j(d)) = \sum_{d=1}^{M} \sum_{l=0}^{L} \frac{1}{2^{L-l}} \left(\sum_{s=1}^{2^l} \sum_{t=1}^{2^l} \min(z_i^l(s,t,d), z_j^l(s,t,d)) \right) \quad (4)$$

where z represents the max pooling feature of image I_i at *(s-t)*-th segment of *l*-th level, D is the dimension of final feature vector z, and M is the dimension of feature vector in each segment of pyramid space. The spatial pyramid matching kernel is proved to be Mercer kernel in [18]. When adopting M basis vectors and L level space pyramid, the resulting vector has dimensionality $M \sum_{l=0}^{L} 4^l$. During the experiment in section V, a larger dictionary is adopted, where *M=2048* and *L=2*, and then the dimension of the feature vector z is *43008*. Taking the high computational complexity and the correlations existed among feature dimensions into account, we perform PCA to feature vector z, in order to reduce the kernel computation time and remove the correlations among feature dimensions.

5. Experiments and Performance Analysis

A. Experiment settings

In this paper, we perform experiments on Caltech-101[20] and Caltech-256[21]. We use Dense-SIFT in feature extraction procedure. Each image is densely sampled to extract SIFT feature. The sample region is 16*16, and the step is 6 pixels. In the image sparse coding stage, we random choose 50000 SIFT descriptors which are extracted in feature extraction procedure and use these chosen descriptors as training sample. Using equation 2, we get the basis vector set Φ , which contains 2048 basis vectors. We perform PCA before SVM classification; only remain the first 1024 dimensions. In the multi-class classification stage, we adopt one-vs-one classification strategy. The accuracy is the average of the classification accuracy for each class. We perform the experiment for 5 times, and take the mean of all these 5 experiment results as the final result. We use LIBSVM [22] as our SVM classifier.

We realize the improved spatial pyramid matching kernel ScKSPM and carry out comparisons with the existing SPM methods on Caltech-101 and Caltech-256. The methods used for comparison are: (1) KSPM: the popular nonlinear kernel SPM that uses spatial pyramid histograms and Chi-square kernels; (2) LSPM: the simple linear SPM that uses linear kernel on spatial pyramid histograms; (3) ScSPM: the linear SPM that use linear kernel on spatial pyramid pooling of SIFT sparse codes. Some of our presented results are drawn from [8]. Also, we further compare the performance and the running time of ScKSPM with or without PCA.

B. Performance comparison of SPM methods

We followed the common experiment setup for Caltech-101, training on 15 and 30 images per category respectively and testing on the rest. For Caltech-256 dataset, we train on 15, 30, 45, and 60 images per category respectively and test on the rest. Detailed comparison results are shown in Table 1 and Table 2.

As shown in Table 1 and Table 2, along with the increase of the number of training samples, the classification performance of these SPM methods have different degrees of improvement. For different data sets, the performances of these methods are also different. Over all, the KSPM method that uses Chi-square kernels outperforms LSPM method that based on linear kernel, while ScSPM that uses linear kernel on sparse codes achieves a much better performance than the former two SPM methods. Our ScKSPM method outperforms the ScSPM method by more than 3 percent. In the cases of 45 and 60 training images per category, KSPM and LSPM was not tried due to its very high computation cost for training.

TABLE I Classification accuracy (%) comparison on Caltech-101

Numbers of training samples	SPM methods			
	KSPM	LSPM	ScSPM	ScKSPM
15	56.44	53.23	67.0	**67.64**
30	63.99	58.81	73.2	**73.90**

TABLE II Classification accuracy (%) comparison on Caltech-256

Numbers of training samples	SPM methods			
	KSPM	LSPM	ScSPM	ScKSPM
15	23.34	13.20	27.23	**29.75**
30	29.51	15.45	34.02	**36.60**
45	—	—	37.46	**38.36**
60	—	—	40.14	**41.97**

C. PCA dimension reduction

In this section, we examine the impact of PCA process on the performance of sparse coding vector extracted in feature extraction stage by comparing the performance and the running time of ScSPM and ScKSPM before and after PCA process. We conduct the experiments on Caltech-101, where we randomly select 30 images from each class for training. The classification results are shown in Table 3.

TABLE III The influence of PCA on the classification accuracy (%) comparison on Caltech-101

	ScSPM	ScKSPM
no PCA	73.2	71.0
PCA	71.7	**73.9**

From Table 3 we can see that, the PCA process harms the classification performance of ScSPM with linear kernel. In the contrary, the performance of ScKSPM has been improved for nearly 3% by performing PCA process. That indicates through

the dimensional reduction process in PCA, the correlation among dimensions in feature vector have been removed, which benefits the calculation of kernel function proposed in this paper. It is worth noting that, the PCA process can significantly reduce the calculation time of both methods. We implement all method on the same machine and see that, on Caltech-101, the calculation time of ScKSPM with PCA process is 15min, which is comparable to ScSPM with linear kernel.

6. Conclusion and Future Work

In this paper, based on sparse coding, we proposed a novel spatial pyramid matching kernel for image classification. We assign different weights for sparse coding vectors in different levels in image representation stage and then adopt SVM with the proposed kernel function for image classification. The experiments on Caltech101 and Caltech256 datasets show that the proposed kernel function performs better than previous kernel functions. Further, we show that through PCA process, we can significantly reduce the running time of the kernel computation time and improve the performance of proposed kernel function. That indicates the basis vectors extracted in feature extraction stage correlate with each other, while the PCA process can remove this correlation and further improve the classification of the proposed kernel function.

We conducted experiments on traditional datasets to check the performance of our proposed kernel function. However, for large scale web dataset, the huge number of images and rich visual information make us not only consider the accuracy of classification, but also the classification efficiency. Efficient methods for feature extraction and sparse coding training should be further investigated in the future.

Acknowledgment

This work is supported by the National Nature Science Foundation of China (No.61272240, 60970047, 61103151), the Doctoral Fund of Ministry of Education of China (No.20110131110028), the Natural Science Foundation of Shandong province (No.ZR2012FM037), the Science Foundation of University of Jinan (No.XKY1316).

References

[1] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In ICCV, 2005.

[2] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. PAMI, 27:1615–1630, 2005.

[3] D.G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60:91–110, 2004.

[4] K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In ICCV, 2005.

[5] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In CVPR, 2008.

[6] P. Quelhas, F. Monay, J. Odobez, D. G.-P. T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In ICCV, 2005

[7] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transaction on Image Processing, 2006.

[8] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In CVPR, 2009

[9] M. Marsza lek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning representations for visual object class recognition. Pascal VOC 2007 challenge workshop. ICCV, 2007

[10] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. A comparison of color features for visual concept classification. In CIVR, 2008

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In CVPR, 2006

[12] Chapelle O, Haffner P, and Vapnik V. SVMs for histogram-based image classification[J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1055-1064.

[13] Barla A, Odone F, and Verri A. Histogram intersection kernel for image classification[C]. Proceedings of the International Conference on Image Processing, Barcelona, Catalonia, Spain, Sept. 14-17, 2003, Vol. 2: 513-516.

[14] Boughorbel S, Tarel J, and Boujemaa N. Generalized histogram intersection kernel for image recognition[C]. Proceedings of the International Conference on Image Processing, Image Processing, Genoa, Italy, Sept. 11-14, 2005: 161-164

[15] Bosch A. Image classification for a large number of object categories[D]. [Ph.D. dissertation], Uuniversity of Girona, 2007

[16] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381:607–609, 1996

[17] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In NIPS, 2006

[18] K. Grauman and T. Darrell. Pyramid match kernels: Discriminative classification with sets of image features. In Proc. ICCV, 2005

[19] Y-L. Boureau, F. Bach, Y. LeCun, J. Ponce. Learning Mid-Level Features For Recognition. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[20] L. Fei-Fei, R. Fergus and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE. CVPR 2004, Workshop on Generative-Model Based Vision. 2004

[21] Griffin, G. Holub, AD. Perona, P. The Caltech-256, Caltech Technical Report 7694. California Institute of Technology, 2007.

[22] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. Software available at

http://www.csie.ntu.edu.tw/?cjlin/libsvm.