A Robust Image Classification Scheme with Sparse Coding and Multiple Kernel Learning

Dongyang Cheng[1], Tanfeng Sun[1,2,3], and Xinghao Jiang[1,2,*]

[1] School of Information Security Engineering Shanghai Jiao Tong University,
Shanghai 200240, China
{dycheng,tfsun,xhjiang}@sjtu.edu.cn
[2] National Engineering Lab on Information Content Analysis Techniques, GT036001,
Shanghai 200240, China
[3] Department of Electrical and Computer Engineering, New Jersey Institute of Technology,
Newark 07102, USA

Abstract. In recent researches, image classification of objects and scenes has attracted much attention, but the accuracy of some schemes may drop when dealing with complicated datasets. In this paper, we propose an image classification scheme based on image sparse representation and multiple kernel learning (MKL) for the sake of better classification performance. As the fundamental part of our scheme, sparse coding method is adopted to generate precise representation of images. Besides, feature fusion is utilized and a new MKL method is proposed to fit the multi-feature case. Experiments demonstrate that our scheme remarkably improves the classification accuracy, leading to state-of-art performance on several benchmarks, including some rather complicated datasets such as Caltech-101 and Caltech-256.

Keywords: Sparse coding, MKL, Feature fusion.

1 Introduction

Nowadays, image classification has captured a lot of interest in computer vision. The common classification schemes mainly consist of two parts: image representation and classification.

With regard to image representation models, Bag of Words (BoW) model with following three modules has been widely used and shows good performance: (i) Region selection and representation; (ii) Codebook generation and feature quantization; (iii) Frequency histogram based image representation. Specifically, the codebook consisting of entries of visual words is used to reconstruct the input local features. The process to generate the codebook and quantize features governs the quality of image representation. But the frequently used k-means method may lead to severe information loss since it assigns each feature to only one visual word in the codebook.

[*] Corresponding author.

Y.Q. Shi, H.J. Kim, and F. Pérez-González (Eds.): IWDW 2012, LNCS 7809, pp. 520–529, 2013.

After the image is represented as a histogram of visual words, a classifier will be required to make the decision that which category the histogram belongs to. Kernel based classifiers such as support vector machine (SVM) are now widely used by many researchers for their wonderful performance. For SVM, the input histograms are mapped to a higher dimensional space by kernel function, in which they can be easily classified in a linear way. However, the sensitiveness of kernel function to categories will increase the fluctuation in accuracy, resulting in a relatively unsatisfying overall performance.

Many works have been done to improve the classification performance. Yang *et al.* [1] applied sparse coding instead of k-means since it can learn the optimal codebook and reduce the information loss. Zhang *et al.* [2] proposed a framework by leveraging an improved sparse coding method, low-rank and sparse matrix decomposition techniques. Linear SVM classifier is used for classification. Gao *et al.* [3] proposed a robust Laplacian sparse coding algorithm for feature quantization which generated more discriminative sparse codes. Naveen *et al.* [4] presented a new framework which was built upon a way of feature extraction that generates largely affine-invariant features and an AdaBoost based classifier. From the perspective of classifier, multiple kernel learning (MKL) can increase the stability of overall performance by learning a linear combination of a series of kernel functions. Bosch *et al.* [5] combined different features by using a weighted linear combination of kernels, where the weights were learnt on a validation set. Lampert *et al.* [6] proposed a method to combine the efficiency of single class localization with a subsequent decision process that worked jointly for all given object classes.

In this paper, we devise a novel image classification scheme by adopting sparse coding and multi-feature MKL, which can ameliorate the image representation and classification phase respectively. The improved multi-feature MKL is proposed based on original MKL, in order to adapt to multi-feature case. Specifically, SIFT and SURF descriptors are extracted and then converted into sparse vectors precisely by the trained dictionaries. The images can be represented by these vectors using max-pooling method which is proved to be more robust than others. After that, the two descriptors are combined into a single vector. Finally, multi-feature MKL approach is implemented to train and test those histograms, generating stable results due to the auto adjustment of the linear combination of kernel functions for each feature.

2 Proposed Scheme

As two main parts in image classification scheme, image representation and classification can substantially affect the classification performance. On one hand, a good kernel method for classification is necessary, for it provides an intuitive and principled tool for learning from high-dimensional vectors that represent images. On the other hand, the performance of kernel method strongly depends on the data representation of images, which means an accurate image representation algorithm is indispensable. Our paper is to enhance the image classification accuracy through the amelioration of both parts.

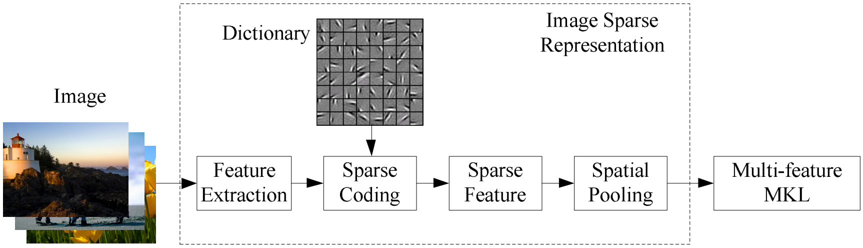


Fig. 1. Framework of proposed scheme (ScMMKL)

Fig. 1 is the framework of the proposed scheme which mainly consists of sparse coding and multi-feature MKL (ScMMKL). The extended multi-feature MKL is defined theoretically for feature fusion method.

The process of our algorithm is as follows:

1. 128-dimentional D-SIFT and 64-dimensional D-SURF descriptors are extracted from the images.

2. Dictionaries are learned based on those features using sparse coding method. This step is of most importance in image sparse representation phase because a better dictionary yields more accurate image representation.

3. Each feature point is denoted as a sparse vector based on the dictionaries trained previously.

4. Represent the image as a single vector using spatial pooling method. Thus an image can be represented as a 128-dimentional (SIFT) or 64-dimentional (SURF) vector after the pooling. Then the two vectors are combined together.

5. The last step of our algorithm is the multi-feature MKL. Kernel combinations are determined for each feature and the final decision can be generated.

2.1 Implementation of Sparse Representation

Comparing with k-means, sparse coding method represents images more precisely, for it describes each feature as a linear combination of basic vectors with minor

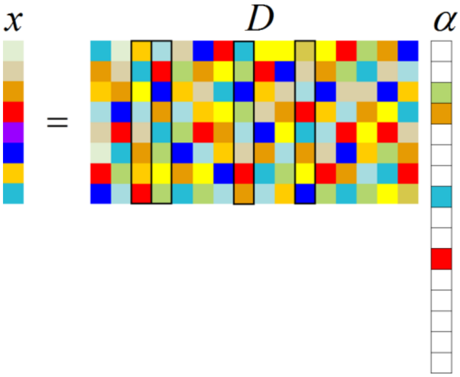


Fig. 2. Visually explanation of sparse representation

quantization loss. Besides, the high dimensional space used to represent features can lead to an easier classification. Fig. 2 is the visually explanation of this procedure.

There are two main steps to apply sparse coding to image representation: dictionary learning and sparse representation. These two steps are similar to codebook generation and vector quantization in traditional BoW model using k-means.

In the phase of dictionary learning, a small set of images should be selected from the whole image dataset randomly. For each image, D-SIFT and D-SURF features are extracted. Then the set of SIFT or SURF descriptors $X = [x_1, x_2, \dots, x_k]$ is used to optimize an empirical cost function to train the dictionary:

$$f_k(D) = \frac{1}{k} \sum_{i=1}^{k} \ell(x_i, D) \tag{1}$$

Where $D \in \mathrm{R}^{m \times n}$ is a dictionary and $\ell(x, D)$ is a loss function of which smaller values yield better dictionaries. The loss function can be defined as the optimal value of the ℓ_1 sparse coding problem:

$$\ell(x, D) = \min_{\alpha \in \mathrm{R}^n} \frac{1}{2} \|x - D\alpha\|_2^2 + \lambda \|\alpha\|_1 \tag{2}$$

Where λ is a regularization parameter. Here, we base our algorithm on [7] which is one of the most efficient dictionary learning algorithms.

Given the trained dictionary, the image can be denoted by the pooling of all descriptors which are sparsely represented. In detail, every descriptor can be represented as a sparse vector using:

$$x_i \approx D\alpha_i \tag{3}$$

Where $D \in \mathrm{R}^{m \times n}$ is the dictionary, and $\alpha_i \in \mathrm{R}^n$ is the sparse representation of descriptor x_i. Least Angle Regression (LARS) algorithm [8] is used to solve this problem.

In order to represent an image with a single vector P, a pooling method needs to be applied. Among the commonly used pooling methods such as average pooling, max pooling and square root pooling, the max pooling procedure is well established by biophysical evidence in visual cortex [9] and is empirically justified by many algorithms applied to image categorization. So in our case, we also use max pooling denoted as follows:

$$p_j = \max\{|\alpha_{1j}|, |\alpha_{2j}|, \dots, |\alpha_{kj}|\} \tag{4}$$

Where p_j is the j-th element in vector P and α_{ij} is the j-th element in the i-th descriptor α_i. Thus, vector P is the sparse representation of the image.

2.2 Multi-feature MKL

As a typical kernel method, the performance of SVM is sensitive to feature type and kernel parameters, while MKL could generate the optimal result through combination of different kernels. In our scheme, two features are involved and a new MKL method (multi-feature MKL) should be developed to achieve multi-feature classification.

For the original MKL, the objective is to optimize jointly over a linear combination of kernels:

$$k^{*} = \sum_{m=1}^{F} \beta_m k_m(x,x') \tag{5}$$

Where F is the number of kernels and $\beta_m > 0,\ \sum_{m=1}^{F} \beta_m = 1$. The objective function can be denoted as follows [10]:

$$\min_{\alpha,\beta,b}\left(\frac{1}{2}\sum_{m=1}^{F} \beta_m \alpha^{\mathrm{T}} K_m \alpha + C\sum_{i=1}^{N} L(y_i, b + \sum_{m=1}^{F} \beta_m K_m(x)^{\mathrm{T}} \alpha)\right) \tag{6}$$

Where $L(y,t) = \max(0, 1 - yt)$ denotes the Hinge loss, C is the misclassification penalty, parameters $\alpha \in \mathrm{R}^N$ and $b \in \mathrm{R}$ are of an SVM. The decision function is like this:

$$F_{\mathrm{MKL}}(x) = \mathrm{sign}\left(\sum_{m=1}^{F} \beta_m (K_m(x)^{\mathrm{T}} \alpha + b)\right) \tag{7}$$

Where $K_m(x)$ is the kernel response of the m-th kernel for a given sample x and $K_m(x)^{\mathrm{T}}$ is the transposition of the vector $K_m(x)$.

However, the original MKL could only be used in single feature classification. For our scheme, different kernel combinations will be learned for each feature and:

$$k_{\mathrm{MF}} = \sum_{i=1}^{n} c_i k_i^{*} \tag{8}$$

Where k_{MF} is the kernel combination for multiple features, n is the number of features, c_i is the coefficient for each kernel combination and k_i^* is the kernel combination for feature i. The multi-feature MKL function can be defined based on (5):

$$k_{\mathrm{MF}} = \sum_{i=1}^{n}\left(c_i \sum_{j=1}^{m_i} \beta_{ij} k_{ij}(x_i, x_i')\right) \tag{9}$$

Where m_i is the number of kernels combined for feature i, k_{ij} is the j-th kernel in feature i and β_{ij} is the coefficient for k_{ij}. In order to obtain the best performance, we need to consider the weight of both features and kernels. So the constraint for kernel coefficients should be changed to $\sum_{i=1}^{n} c_i \sum_{j=1}^{m_i} \beta_{ij} = 1$. Take (9) into (7) and we can get the final MKL decision function for our algorithm:

$$F_{\mathrm{MKL}} = \mathrm{sign}\left(\sum_{i=1}^{n} c_i \left(\sum_{j=1}^{m_i}\left(\beta_{ij}(K_{ij}(x_i)^{\mathrm{T}}\alpha + b)\right)\right)\right) \tag{10}$$

Through the above definition, this extended multi-feature MKL can be directly used for any multi-feature problem.

3 Experimental Result

3.1 Experimental Setup

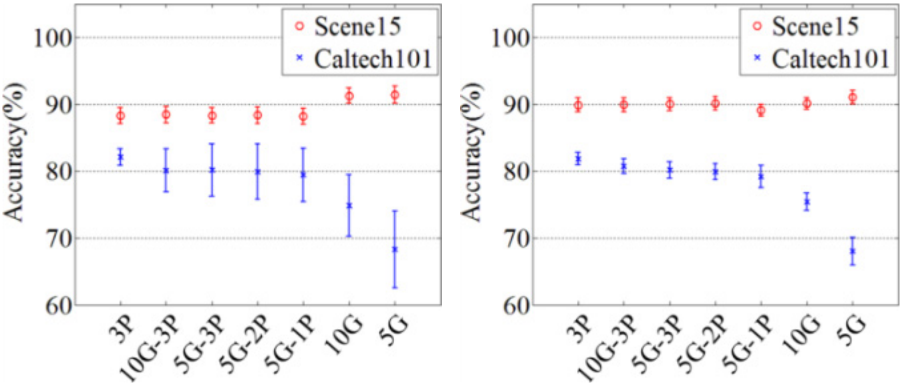
We evaluate the proposed approach ScMMKL on three public dataset: Scene-15, Caltech-101 and Caltech-256. In the phase of sparse coding, the dictionaries trained for SIFT and SURF are $R^{128\times500}$ and $R^{64\times300}$ respectively. The outputs of sparse coding are a series of sparse codes, each representing one image.

For multi-feature MKL, we base our algorithm on SimpleMKL [10] and extended original MKL to a multi-feature one. Certain parts of sparse codes are combined with labels (1 for positive and -1 for negative) to generate the training matrix and some other sparse codes without labels are used for testing. For detailed parameters, ref. [11] proposed that high values of C in (6) turned out to work better and C = 100 is found to perform the best in our case. Moreover, some iteration processes with corresponding stop criterion should be utilized to gain optimal parameters. Through experiments with small sample size, the duality gap with parameter of 0.01 is more suitable for our scheme.

3.2 Kernel Selection Experiment

The following experiment is designed to determine which kernel combination works best for our scheme. As Gaussian and polynomial kernels are most commonly used, seven kernel combinations are taken into consideration: 3P, 5G, 10G, 5G+1P, 5G+2P, 5G+3P and 10G+3P, where G and P denote Gaussian and polynomial kernel respectively.

Fig. 3 shows the result for SIFT and SURF features on Scene-15 and Caltech-101. It can be seen that after eliminating polynomial kernels, the accuracy becomes slightly better for Scene-15, while the performance for Caltach-101 is extremely poor. Take SIFT feature for instance, the results for Scene-15 and Caltech-101 are 88.1% and



(a) SIFT (b)SURF

Fig. 3. Result of different combination of kernels

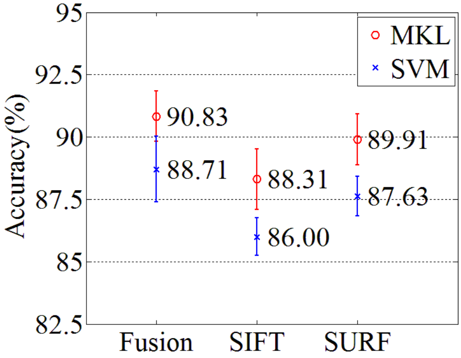
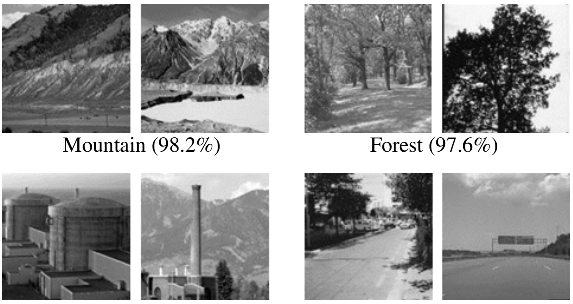


Fig. 4. Classification accuracy on Scene-15 dataset


Mountain (98.2%)

Forest (97.6%)

Industrial (78.4%)

Highway (73.8%)

Fig. 5. Sample images in Scene-15 dataset

82.1% using 3 polynomial kernels. With only Gaussian kernels, though the accuracy rises up to 91.5% for Scene-15, the data for Caltech-101 are unacceptable with 68.3% accuracy and 5.75% standard deviation. This happens to SURF feature as well. In consideration of stronger practicability, the combination of 3 polynomial kernels is selected for both features.

3.3 Scene-15 Dataset

The Scene-15 dataset has 4,485 images in 15 categories. Experimental process is repeated for 10 times with randomly selecting training and testing images to obtain reliable results. Each category is treated as the test database in turn, and one versus rest scheme is employed. According to common practice, both of the chosen training set and testing set include 100 images. The final results are reported by the mean and standard deviation of classification rates per category which are recorded in each run.

Fig. 4 is the result of single vs. multi feature and SVM vs. MKL comparison on Scene-15. It's obvious that feature fusion and multi-feature MKL are better design choices.

Fig. 5 shows some sample images from classes with highest and lowest classification accuracies in Scene-15 dataset. Our scheme performs better for categories like mountain and coast because the meaningful part takes a large percentage of the whole image, while the complicated background objects in industrial and highway may lead to misjudgments.

Table 1 gives the performance comparison of our approach and some other methods proposed in [1] [2] [3] [12]. The first 3 algorithms are based on sparse coding and SVM with the same experimental parameters comparing with ours. Ref. [12] is an improved BoW model using k-means and SVM.

As can be seen from the table, our scheme with sparse coding and multi-feature MKL generates a satisfying performance in image classification. The ascendency of sparse coding to k-means is apparent comparing with [12], because it can represent images more precisely with less quantization loss. Besides, our scheme outperforms [1] by 10% due to the superiority of multi-feature MKL with combinations of kernels. Though the accuracies in [2] [3] are fairly high, our scheme still achieves 1% improvement.

Table 1. Comparison on Scene-15 dataset

Method	**Accuracy**
ScSPM[1]	80.28±0.93
LScSPM[3]	89.75±0.50
LR-Sc+SPM[2]	90.03±0.70
Improved BoW[12]	79.0
ScMMKL	**90.83±1.01**

3.4 Caltech-101 Dataset

The Caltech-101 dataset contains 102 classes with high intra-class appearance and shape variability. In this dataset, we randomly choose 15/30 images per category for training, another 15 and up to 30 images for testing.

Table 2 gives the performance comparison of the method proposed in this paper and some other literatures [1] [2] [4]. [4] adopted an improved sparse coding method and an AdaBoost based classifier. As is shown in the table, our scheme outperforms the LR-Sc+SPM [2] by more than 11.5% for 15 training and 10% for 30 training. The superiority over [1] and [2] is reasonable because the accuracies in [1] and [2] fluctuate due to the large volume of categories in Caltech-101 and the sensitiveness of single kernel function to categories, while the kernel combination in MKL can stabilize the performance since it can adjust the weight of each kernel automatically to gain the optimal result. Our scheme also achieves a 3%-4% improvement compared with [4] whose classifier performs better than SVM.

Table 2. Comparison on Caltech-101 dataset

Method	**15 training**	**30 training**
ScSPM[1]	67.00±0.45	73.20±0.54
LR-Sc+SPM[2]	69.58±0.97	75.68±0.89
Naveen *et al.*[4]	78.38	83.28
ScMMKL	**82.93±1.42**	**86.32±0.88**

3.5 Caltech-256 Dataset

The Caltech-256 dataset has 29,780 images of 257 classes. The intra-class variance including object location is much bigger than Caltech 101 and makes it a very challenging dataset so far for object recognition. The image number for training and testing are set to 15, 30 and 45 as usual practice, and the experiment is repeated for 5 times under each allocation.

Table 3 gives the comparison results with [1] [2] [3] [4]. Significant improvement with 30%-40% gap can be seen from the table. With more categories in this complicated dataset, the ascendency of MKL with strong adaptability is more pronounced.

There is another reason for this advantage compared with [1] [2] [3] which use SPM kernel. It's notable that the linear SPM kernel takes spatial information into consideration, but the high intra class variability and object location variability in Caltech-256 result in the totally different backgrounds of objects. The image is divided into several patches by SPM, but some patches may have no correlation with target objects. Therefore, the consideration of background by SPM kernel may lead to misclassification and drag final accuracy down.

It's noteworthy that as the category number increases, the performance of our approach has a small fluctuation, providing a scheme with strong stability and practicability.

Table 3. Comparison on Caltech-256 dataset

Method	**15 training**	**30 training**	**45 training**
ScSPM[1]	27.73±0.51	34.02±0.35	37.46±0.55
LScSPM[3]	30.00±0.14	35.74±0.10	38.54±0.36
LR-Sc+SPM[2]	35.31±0.70	N/A	N/A
Naveen *et al.*[4]	39.42	45.83	49.3
ScMMKL	**71.47±1.32**	**74.44±0.63**	**78.26±0.76**

4 Conclusion

In this paper, we proposed an image classification scheme with sparse coding and multi-feature MKL techniques, which improves the image representation and classification phases simultaneously. Furthermore, feature fusion scheme is used and the original MKL is redefined to adapt to multiple feature case, providing theoretical and experimental support to the extension of MKL. Experimental result shows that ScMMKL has a state-of-art performance on several public datasets with strong adaptability and stability.

Acknowledgment. The work of this paper is sponsored by the National Natural Science Foundation of China (No. 61071153, No. 61272249), the National New Century Excellent Talents Support Plan of Ministry of Education, China (No. NECT-10-0569). It is also under the Project of International Cooperation and Exchanges supported by Shanghai Committee of Science and Technology (No. 12510708500).

References

1. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR, pp. 1794–1801. IEEE, Miami Beach (2009)
2. Zhang, C., Liu, J., Tian, Q.: Image classification by non-negative sparse coding, low-rank and sparse decomposition. In: CVPR, pp. 1673–1680. IEEE, Colorado Springs (2011)
3. Gao, S., Tsang, I., Chia, L., Zhao, P.: Local features are not lonely-Laplacian sparse coding for image classification. In: CVPR, pp. 3555–3561. IEEE, San Francisco (2010)
4. Naveen, K., Li, B.: Discriminative Affine Sparse Codes for Image Classification. In: CVPR, pp. 1609–1616. IEEE, Colorado Springs (2011)
5. Bosch, A., Zisserman, A., Munoz, X.: Image classification using rois and multiple kernel learning. Intl. J. Computer Vision (2008)
6. Lampert, C., Blaschko, M.: A multiple kernel learning approach to joint multi-class object detection. In: Proceedings of the 30th DAGM Symposium on Pattern Recognition, pp. 31–40 (2008)
7. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online Learning for Matrix Factorization and Sparse Coding. Journal of Machine Learning Research 11, 19–60 (2010)
8. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Annals of Statistics 32(2), 407–499 (2004)
9. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: CVPR, pp. 994–1000. IEEE, San Diego (2005)
10. Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: More efficiency in multiple kernel learning. In: ICML, pp. 775–782. ACM, Corvalis (2007)
11. Gehler, P.V., Nowozin, S.: On feature combination for multiclass object classification. In: ICCV, pp. 221–228. IEEE, Kyoto (2009)
12. Hao, J., Jie, X.: Improved Bags-of-Words Algorithm for Scene Recognition. In: ICSPS, pp. 279–282. IEEE, Dalian (2010)