DISCRIMINANT SPARSE CODING FOR IMAGE CLASSIFICATION

Bao-Di Liu, Yu-Xiong Wang, Yu-Jin Zhang, Yin Zheng

Tsinghua National Laboratory for Information Science and Technology, Tsinghua University
Department of Electronic Engineering, Tsinghua University
Beijing 100084, China
lbd08@mails.tsinghua.edu.cn, albertwyx@gmail.com, zhang-yj@mail.tsinghua.edu.cn,
y-zheng09@mails.tsinghua.edu.cn

ABSTRACT

Recently, dictionary learned by sparse coding has been widely adopted in image classification and has achieved competitive performance. Sparse coding is capable of reducing the reconstruction error in transforming low-level descriptors into compact mid-level features. Nevertheless, dictionary learned by sparse coding does not have the ability to distinguish different classes. That is to say, it is not the optimum dictionary for the classification task. In this paper, based on the global image statistics, a novel discriminant dictionary learning method combining linear discriminant analysis with sparse coding is proposed to obtain a more discriminative dictionary while preserving its descriptive abilities and a block coordinate descent algorithm is proposed to solve the optimization problem. Experimental results show that our algorithm has capabilities to learn dictionary with more discriminative power and achieves superior performance.

Index Terms— Sparse coding, image classification, dictionary learning, linear discriminant analysis

1. INTRODUCTION

Recently, image classification, which aims at associating images with semantic labels automatically, has become quite a significant topic. The typical framework adopted by the majority of existing image classification systems is discriminative model [1, 2, 3, 4, 5, 6, 7]. Initially, bag of words model [7] (also called codebook or codeword, i.e. dictionary), which treats an image as a collection of ”Visual Words”, is the most commonly used method in image classification. Although it achieves satisfactory results, bag of words model has two drawbacks. One is that the spatial information for classification is lost because of unordered ”Visual Words”, thus severely limiting the classification performance. The other is that each feature only corresponds to one word, so this hard decision will cause too large reconstruction error. For the former, the spatial pyramid matching method proposed by Lazebnik et al. [2] has achieved remarkable success, and thus becomes an indispensable step for image classification. For the latter, in order to solve the visual word ambiguity, Van Gemert et al. [3] suggested kernel-codebook, and Wang et al. [6] recommended locality-constrained linear coding which utilizes the linear combination of N-neighborhood bases to represent features. Furthermore, Yang et al. [4] proposed sparse coding based dictionary learning, which represents features by the sparse linear combination of several bases, and achieved state-of-the-art performance.

However, the reconstruction error criterion takes effect mainly in measuring the mapping expression when transforming low-level descriptors into compact mid-level features. For the classification task, merely abasing the reconstruction error is far from enough. The optimum dictionary should have the ability to distinguish different classes. Hence, Lazebnik et al. [8] incorporated discriminative information by minimizing the loss of mutual information between features and labels during the quantization step. Mairal et al. [9] proposed a sparse coding based discriminative dictionary training approach with respect to the sparse codes rather than the pooling results, so it requires each code to be labeled, and ignores global image statistics.

In this paper, to obtain a more discriminative dictionary while preserving its descriptive abilities, we combine linear discriminant analysis (LDA) with sparse coding. To be specific, the Fisher linear discriminant analysis information is embedded as regularization term into the original objective function of the sparse coding method so as to effectively reduce the within-class scatter as well as increasing the between-class scatter. Based on this formula, a novel block coordinate descent algorithm is proposed to solve the minimization problem. Therein we try to optimize one single variable at a time, and thus the closed-form solution furthest decreasing the corresponding objective function is obtained based on the convexity of a much simpler univariate parabolic function.

The rest of this paper are organized as follows. Section 2 overviews the framework and formulation of sparse coding based image classification. How to incorporate the discriminant information into the existing schematism and construct the corresponding objective function is elaborated in Section 3. The solution to the optimization problem of proposed objective function and the implementation of the algorithm are demonstrated in Section 4. Section 5 shows experimental results and analysis. Discussions and conclusions are drawn in Section 6.

2. NOTATIONS AND RELATED WORK

This section firstly reviews the framework for image classification. Then, some notations and formulae used throughout this paper are illustrated.

2.1. Framework of sparse coding based image classification

The framework of sparse coding based image classification include five major steps [4]: Feature extraction, Dictionary learning, Image

This work was supported in part by National Natural Science Foundation of P. R. China under Grant 61171118.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

coding, Pooling combined with spatial pyramid matching (SPM), and classifier training. Figure1 shows the flow chart of the detailed framework.

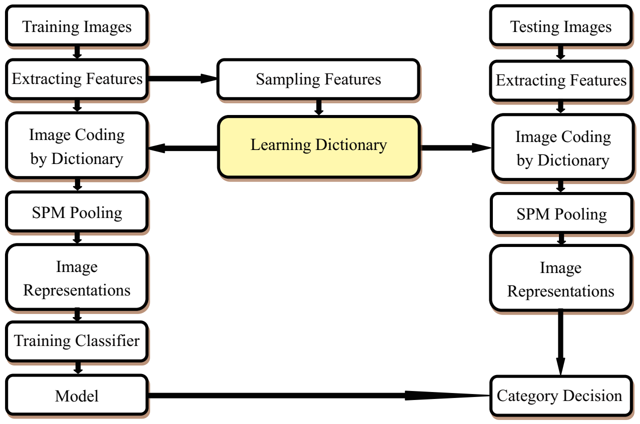

Fig. 1. The framework of sparse coding based image classification. There are five major steps: Feature extraction; sparse coding for dictionary learning; Image coding by dictionary; Pooling combined with spatial pyramid matching; and classifier training. Obviously, dictionary learning is the most important part and dominant the performance of classification.

2.2. Notations and formulae for image classification

Let $\boldsymbol{X} \in \mathbb{R}^{D\times N}$ be the local descriptors randomly extracted from training images for learning dictionary, D be the dimension of $\boldsymbol{X}$, and N be the number of samples in $\boldsymbol{X}$. For classification task, Let C represent the number of classes, N_0 represents the number of images extracted from each class for learning dictionary and training classifier, M represents the number of features randomly extracted from each image for learning dictionary. Then we have $N = C \times N_0 \times M$. $\boldsymbol{B} \in \mathbb{R}^{D\times K}$ is the dictionary, K is the size of the dictionary. $\boldsymbol{S} \in \mathbb{R}^{K\times N}$ is the local descriptors' codes under given dictionary $\boldsymbol{B}$.

In this paper, for the convenience of deriving the solution to the ensuing discriminant sparse coding, we impose additional nonnegativity constraint to the sparse codes. Then sparse coding can be formulated as minimization of the following objective function:

$$\begin{aligned} f(\boldsymbol{B},\boldsymbol{S}) \;&=\; \|\boldsymbol{X}-\boldsymbol{B}\boldsymbol{S}\|_{\mathrm{F}}^2 + 2\alpha\,\|\boldsymbol{S}\|_1 \\ s.t.\ \boldsymbol{S} \;&\geq\; 0,\;\; \|\boldsymbol{B}_{\bullet i}\|_2 = 1, \forall i = 1,2,...,K \end{aligned} \tag{1}$$

Here $\boldsymbol{A}_{\bullet n}$ and $\boldsymbol{A}_{k\bullet}$ denote the n_{th} column and k_{th} row vectors of matrix $\boldsymbol{A}$, respectively. α is a regularization parameter to control the tradeoff between fitting goodness and sparseness. Dictionary $\boldsymbol{B}$ can be obtained by solving (1).

After obtaining the dictionary, for each image $\boldsymbol{I}$, we assume that $\boldsymbol{X}^I \in R^{D\times L}$ represents the features extracted from $\boldsymbol{I}$, and $\boldsymbol{S}^I \in R^{K\times L}$ represents the corresponding sparse codes. Let f_c and f_p denote coding and pooling operators, $\boldsymbol{Z}^I \in R^{K\times 1}$ denotes the image representation. Then the coding step and pooling step can be formulated as:

$$\boldsymbol{S}^I = f_c(\boldsymbol{X}^I) \qquad \boldsymbol{Z}^I = f_p(\boldsymbol{S}^I) \tag{2}$$

3. DISCRIMINANT SPARSE CODING

Although sparse coding algorithm can effectively reduce the reconstruction error and has achieved competitive performance for image classification, some other factors, such as within-class scatter and between-class scatter should be also considered. Specifically, we should take the strategy to debase the within-class scatter and increase the between-class scatter.

3.1. Discriminant information

Let Z_{ci} represent the i_{th} image in the c_{th} class. Then Z_{ci} can be written as follows,

$$\boldsymbol{Z}_{ci} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{S}_{\bullet m}^{ci} \tag{3}$$

Let $\overline{\boldsymbol{Z}_c}$ represent the mean vector in the c_{th} class:

$$\overline{\boldsymbol{Z}_c} = \frac{1}{N_0}\sum_{i=1}^{N_0} \boldsymbol{Z}_{ci} \tag{4}$$

The within-class scatter D_W can be formulated as:

$$D_W = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_0}\sum_{i=1}^{N_0}\left(\boldsymbol{Z}_{ci}-\overline{\boldsymbol{Z}_c}\right)^T\left(\boldsymbol{Z}_{ci}-\overline{\boldsymbol{Z}_c}\right) \tag{5}$$

The corresponding between-class scatter D_B is

$$D_B = \frac{1}{C(C-1)}\sum_{c=1}^{C}\sum_{d=1}^{C}\left(\overline{\boldsymbol{Z}_c}-\overline{\boldsymbol{Z}_d}\right)^T\left(\overline{\boldsymbol{Z}_c}-\overline{\boldsymbol{Z}_d}\right) \tag{6}$$

3.2. Proposed discriminant sparse coding algorithm

Elevating the between-class scatter and suppressing the within-class scatter will undoubtedly improve the performance of classification. In this paper, D_W and D_B are embedded into (1) as regularization terms forming the following formula,

$$\begin{aligned} f(\boldsymbol{B},\boldsymbol{S}) \;&=\; \|\boldsymbol{X}-\boldsymbol{B}\boldsymbol{S}\|_F^2 + 2\alpha\|\boldsymbol{S}\|_1 + \beta D_W - \gamma D_B \\ s.t.\boldsymbol{S} \;&\geq\; 0\,, \|\boldsymbol{B}_{\bullet i}\|_2 = 1, \forall i = 1,2,...,K \end{aligned} \tag{7}$$

where β and γ are the regularization factors that adjust the weight of within-class scatter and between-class scatter, respectively. In other words, these parameters are used for balancing the trade-off between the descriptive and discriminative abilities.

Using (5) and (6), (7) can be simplified as follows,

$$\begin{aligned} f(\boldsymbol{B},\boldsymbol{S}) \;&=\; \|\boldsymbol{X}-\boldsymbol{B}\boldsymbol{S}\|_F^2 + 2\alpha\|\boldsymbol{S}\|_1 + pu \times \|\boldsymbol{S}\boldsymbol{U}\|_F^2 \\ &\quad -pv \times \|\boldsymbol{S}\boldsymbol{V}\|_F^2 + pw \times \|\boldsymbol{S}\boldsymbol{W}\|_F^2 \\ &s.t.\ \boldsymbol{S} \geq 0,\; \|\boldsymbol{B}_{\bullet i}\|_2 = 1, \forall i = 1,2,...,K \end{aligned} \tag{8}$$

Here, $pu = \dfrac{\beta}{NM}$, $pv = \dfrac{\beta(C-1)+2\gamma C}{N(N-N_0M)}$, $pw = \dfrac{2\gamma}{N(N-N_0M)}$, $\boldsymbol{U} \in \mathbb{R}^{N\times(N_0\times C)}$ is the feature-image label matrix. If the feature (i.e. the row index) belongs to the image (i.e. the column index), the corresponding elements are set to 1; otherwise 0. $\boldsymbol{V} \in \mathbb{R}^{N\times C}$ represents the feature-class label matrix. Similarly, if the feature (i.e. the row index) belongs to the class (i.e. the column index), the corresponding elements are set to 1; otherwise 0. $\boldsymbol{W} \in \mathbb{R}^{N\times 1}$ is a vector with all elements 1.

4. OPTIMIZATION OF THE OBJECTIVE FUNCTION

In this section, we focus on solving the minimization of objective function (8). While this optimization problem is not jointly convex in both $\boldsymbol{S}$ and $\boldsymbol{B}$, it is separately convex in either $\boldsymbol{S}$ or $\boldsymbol{B}$. So it can be decoupled into the following two optimization subproblems which can be solved by alternating minimizations. Finding the sparse codes is as follows,

$$\begin{aligned} \min_{\boldsymbol{S}\geq 0} f(\boldsymbol{S}) \;=\; & \|\boldsymbol{X}-\boldsymbol{B}\boldsymbol{S}\|_F^2 + 2\alpha\|\boldsymbol{S}\|_1 + pu \times \|\boldsymbol{S}\boldsymbol{U}\|_F^2 \\ & -pv \times \|\boldsymbol{S}\boldsymbol{V}\|_F^2 + pw \times \|\boldsymbol{S}\boldsymbol{W}\|_F^2 \end{aligned} \tag{9}$$

Learning bases is as follows,

$$\begin{aligned} \min f(\boldsymbol{B}) \;&=\; \|\boldsymbol{X}-\boldsymbol{B}\boldsymbol{S}\|_F^2 \\ s.t.\|\boldsymbol{B}_{\bullet i}\|_2 \;&=\; 1, \forall i = 1, 2, ..., K \end{aligned} \tag{10}$$

4.1. Finding sparse codes

The objective function in (9) can be rewritten as follows,

$$\begin{aligned} f(\boldsymbol{S}) \;=\; & Tr\left\{\boldsymbol{X}^T\boldsymbol{X} - 2\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{S} + \boldsymbol{S}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{S}\right\} \\ & +2\alpha\sum_{k=1}^{K}\sum_{n=1}^{N}\boldsymbol{S}_{kn} + Tr\left\{\boldsymbol{S}\boldsymbol{G}\boldsymbol{S}^T\right\} \end{aligned} \tag{11}$$

where $\boldsymbol{G} = pu\boldsymbol{U}\boldsymbol{U}^{\boldsymbol{T}} - pv\boldsymbol{V}\boldsymbol{V}^{\boldsymbol{T}} + pw\boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}}$. The objective function (11) in terms of $\boldsymbol{S}_{kn}$ reduces to (12) with $\boldsymbol{B}$ and $\{\boldsymbol{S}_{ij}, i = 1, ..., K, j = 1, ..., N\}/\boldsymbol{S}_{kn}$ fixed.

$$\begin{aligned} f(\boldsymbol{S}_{kn}) \;=\; & \boldsymbol{S}_{kn}^2\{1+pu-pv+pw\} + 2\boldsymbol{S}_{kn}\sum_{m=1,m\neq n}^{N}\boldsymbol{G}_{nm}\boldsymbol{S}_{km} \\ & +2\boldsymbol{S}_{kn}\left\{\sum_{l=1,l\neq k}^{K}[\boldsymbol{B}^T\boldsymbol{B}]_{kl}\boldsymbol{S}_{ln} - [\boldsymbol{X}^T\boldsymbol{B}]_{nk} + \alpha\right\} \end{aligned} \tag{12}$$

where $pu - pv + pw = \frac{C}{N^2}\{\beta(N_0-1)-2\gamma\}$, in this paper, we set $\beta = \gamma$, so $pu - pv + pw > 0$. Thus $f(\boldsymbol{S_{kn}})$ is a parabola which opens up. Based on the convexity and monotonic property of parabolic function, it is not difficult to know that $f(\boldsymbol{S_{kn}})$ reaches the minimum at the unique point.

$$\boldsymbol{S}_{kn} = \max\{H_{kn}-\alpha, 0\}/(1+pu-pv+pw) \tag{13}$$

where $H_{kn} = -\sum\limits_{l=1,l\neq k}^{K}[\boldsymbol{B}^T\boldsymbol{B}]_{kl}\boldsymbol{S}_{ln} + [\boldsymbol{X}^T\boldsymbol{B}]_{nk} - \sum\limits_{m=1,m\neq n}^{N}\boldsymbol{G}_{nm}\boldsymbol{S}_{km}$.

4.2. Learning bases

Without the regularization term in (9) and additional constraints in (10), the solution to $\boldsymbol{S}$ and $\boldsymbol{B}$ are dual in objective function $\|\boldsymbol{X}-\boldsymbol{B}\boldsymbol{S}\|_F^2$. Hence, $\forall d \in \{1, 2, \ldots, D\}, k \in \{1, 2, \ldots, K\}$, with $\{\boldsymbol{B}_{pq,p=1,2,\ldots,D,q=1,2,\ldots,K}\}/\boldsymbol{B}_{dk}$ and $\boldsymbol{S}$ fixed, the constrained single variable minimization problem of (10) has the closed-form solution

$$\begin{aligned} \boldsymbol{B}_{dk} \;&=\; \frac{[\boldsymbol{X}\boldsymbol{S}^T]_{dk} - \sum\limits_{l=1,l\neq k}^{K}\boldsymbol{B}_{dl}[\boldsymbol{S}\boldsymbol{S}^T]_{lk}}{[\boldsymbol{S}\boldsymbol{S}^T]_{kk}}, \\ \boldsymbol{B}_{\bullet k} \;&=\; \frac{\boldsymbol{B}_{\bullet k}}{\|\boldsymbol{B}_{\bullet k}\|_2} \end{aligned} \tag{14}$$

4.3. Overall algorithm

Discriminant sparse coding algorithm for learning dictionary is shown in Algorithm 1. $\boldsymbol{1} \in \mathbb{R}^{K\times K}$ is a square matrix with all elements 1, $\boldsymbol{I} \in \mathbb{R}^{K\times K}$ is the identity matrix, and $\odot$ indicates element dot product.

Algorithm 1 Fast algorithm for sparse coding

```
Require: Data matrix $\boldsymbol{X} \in \mathbb{R}^{D\times N}$, label matrix $\boldsymbol{U} \in \mathbb{R}^{N\times(N_0\times C)}$, $\boldsymbol{V} \in \mathbb{R}^{N\times C}$, $\boldsymbol{W} \in \mathbb{R}^{N\times 1}$, parameter $\alpha$, $\beta$, $\gamma$ and $K$
$\boldsymbol{B} \leftarrow rand(D,K), \boldsymbol{B}_{\bullet k} = \frac{\boldsymbol{B}_{\bullet k}}{\|\boldsymbol{B}_{\bullet k}\|_2} \forall k, \boldsymbol{S} \leftarrow zeros(K,N)$, $pu = \frac{\beta}{MN}$, $pv = \frac{(\beta+2\gamma)C-\beta}{N(N-N_0M)}$, $pw = \frac{2\gamma}{N(N-N_0M)}$
$iteration = 0$
while $(f(iteration) - f(iteration+1))/f(iteration) > 1e{-}6$ do
    $iteration \leftarrow iteration + 1$
    Update $\boldsymbol{S}$:
    Compute $\boldsymbol{A} = (\boldsymbol{B}^T\boldsymbol{B}) \odot (\boldsymbol{1} - \boldsymbol{I})$ and $\boldsymbol{E} = \boldsymbol{B}^T\boldsymbol{X}$
    for $k = 1; k \leq K; k{+}{+}$ do
        $\boldsymbol{P}_1 \leftarrow \mathrm{reshape}(\mathrm{repmat}(pu \times \mathrm{sum}(\mathrm{reshape}(\boldsymbol{S}_{k\bullet}, \mathrm{M}, \mathrm{N}_0 \times \mathrm{C})), \mathrm{M}, 1), 1, \mathrm{N})$
        $\boldsymbol{P}_2 \leftarrow \mathrm{reshape}(\mathrm{repmat}(pv \times \mathrm{sum}(\mathrm{reshape}(\boldsymbol{S}_{k\bullet}, \mathrm{M} \times \mathrm{N}_0, \mathrm{C})), \mathrm{M} \times \mathrm{N}_0, 1), 1, \mathrm{N})$
        $\boldsymbol{P}_3 \leftarrow pw \times \mathrm{sum}(\boldsymbol{S}_{k\bullet})$
        $\boldsymbol{P}_4 \leftarrow (pu - pv + pw) \times \boldsymbol{S}_{k\bullet}$
        $\boldsymbol{S}_{k\bullet} = \max\{\boldsymbol{E}_{k\bullet} - \boldsymbol{A}_{k\bullet}\boldsymbol{S} - \boldsymbol{P}_1 + \boldsymbol{P}_2 - \boldsymbol{P}_3 + \boldsymbol{P}_4 - \alpha, 0\}/(1 + pu - pv + pw)$
    end for
    Update $\boldsymbol{B}$:
    Compute $\boldsymbol{G} = (\boldsymbol{S}\boldsymbol{S}^T) \odot (\boldsymbol{1} - \boldsymbol{I}), \boldsymbol{W} = \boldsymbol{X}\boldsymbol{S}^T$
    for $k = 1; k \leq K; k{+}{+}$ do
        $\boldsymbol{B}_{\bullet k} = \frac{\boldsymbol{W}_{\bullet k} - \boldsymbol{B}\mathbf{G}_{\bullet k}}{\|\boldsymbol{W}_{\bullet k} - \boldsymbol{B}\mathbf{G}_{\bullet k}\|_2}$
    end for
    Update the objective function:
    $f = \|\boldsymbol{X}-\boldsymbol{B}\boldsymbol{S}\|_F^2 + 2\alpha\|\boldsymbol{S}\|_1 + pu \times \|\boldsymbol{S}\boldsymbol{U}\|_F^2 - pv \times \|\boldsymbol{S}\boldsymbol{V}\|_F^2 + pw \times \|\boldsymbol{S}\boldsymbol{W}\|_F^2$
end while
return $\boldsymbol{B}$, and $\boldsymbol{S}$
```

5. EXPERIMENTAL RESULTS

In this section, our discriminant sparse coding (DSC) algorithm is evaluated on three benchmark datasets, including 8 sports event dataset [10], 15 natural scene dataset [2, 11, 12], and Caltech101 dataset [13]. For each dataset, the data are randomly split into training set and testing set based on published protocols. To make the results more convincing, the experimental process is repeated 8 times, the mean and standard deviation of the classification accuracy are recorded. The images are resized with a maximum side 300 pixels. As for the image features, the image patches are densely sampled from each image with step size 8 pixels and side length 16 pixels, and SIFT descriptors are adopted with grid size 4 × 4 to form 128 dimensional feature vectors. The features used for learning dictionary are randomly sampled from all training images and the amount of them is about 120,000. The dictionary size is 1024. Spatial pyramid matching kernel is embedded in the coding step (The image is split into three layers, each of which has 1, 4,

Table 1. Image classification results on three datasets

DataSets	event8(%)	Caltech101(%)	scene(%)
Li[10]	73.4		
Wu[1]	81.87(1.14)	61.00(0.90)	82.02(0.54)
Lazebnik[2]		64.40(0.80)	81.40(0.50)
van Gemert[3]		64.10(1.50)	76.70(0.40)
Yang[4]		66.68(1.66)	73.92(1.03)
Boureau[5]		70.30(1.3)	83.20(0.40)
DSC	**83.72(1.68)**	**71.96(0.83)**	**84.21(0.44)**

and 16 segments, respectively). The pooling strategy is average. Histogram Intersection Kernel SVM classifier and one against all multi-classification strategy are adopted, and LIBSVM [14] package is used.

There are three parameters α, β, γ in the objective function. The parameter α is used for adjusting the sparsity of the codes; the bigger α is, the sparser the codes are. The best performance in [4] is achieved when α is set to 0.15. We follow the same setting of 0.15. The parameters β and γ are used for adjusting the within-class scatter and the between-class scatter, respectively. In our experiment, we make $\beta = \gamma = 0.3 \times N$. The extra multiplication by the number of features N is introduced to keep the fitting and discriminant terms within the same order of magnitude.

For 8 sports event dataset, there are 8 sports event classes with totally 1579 images. 70 images per class are randomly selected as the training data, and 60 images per class for testing. For 15 natural scene dataset, there are 15 classes of indoor or outdoor scene with totally 4485 images. The number of images in each class varies from 200 to 400, and the image size is about 300×300. 100 images per class are randomly selected as the training data, and the rest for testing. For Caltech101 dataset, there are 102 classes, one of which is the background. After removing the background class, the rest 101 classes are used for our classification. These 101 classes contain 8677 images, and the number of images in each class varies from 31 to 800. 30 images per class are randomly selected as the training data, and the rest for testing (with a maximum of 50 images per class).

Table 1 lists the comparisons of DSC with existig work for image classification currently. Our proposed algorithm achieves superior accuracy under similar conditions (such as the same features, the same size of dictionary, the same pooling strategy, and the like). It is worth to note that the conditions (dictionary with 1024 bases + SPM + average pooling + HIK-kernel SVM) used by Boureau [5] are almost the same as ours, but our DSC algorithm outperform Boureau's algorithm 1.66% and 1.01% on Caltech101 and 15 natural scene, respectively. The probable reason for more accuracy improvement on Caltech101 is that for Caltech101 dataset, the within-class difference of image patches is higher than 15 natural scene dataset, after transforming to corresponding sparse codes, this phenomenon still exists for traditional sparse coding algorithm.

6. CONCLUSION

In this paper, we have proposed a novel sparse coding algorithm for image classification. The algorithm incorporates linear discriminant analysis information into sparse coding algorithm so as to reduce the within-class scatter and increase the between-class scatter. Based on the convexity and monotonic property of parabolic function, the algorithm directly obtains the closed-form solution to the separable optimization subproblems. The enhanced discriminative ability makes it more congruent with the classification task. Experimental results on three benchmark datasets show that our algorithm is superior to other image classification algorithms.

In the future, the evaluation of proposed algorithm on larger dataset will be carried out. We also plan to shift to the max pooling strategy and design corresponding discriminant sparse coding algorithm.

7. REFERENCES

[1] J. Wu and J.M. Rehg, "Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," in *Proceedings of the 12th ICCV*. IEEE, 2009, pp. 630–637.

[2] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the 19th CVPR*. IEEE, 2006, vol. 2, pp. 2169–2178.

[3] J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders, and J.M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[4] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proceedings of the 22nd CVPR*. IEEE, 2009, pp. 1794–1801.

[5] Y.L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proceedings of the 23rd CVPR*. IEEE, 2010, pp. 2559–2566.

[6] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proceedings of the 23rd CVPR*. IEEE, 2010, pp. 3360–3367.

[7] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proceedings of the 10th ICCV*. IEEE, 2005, pp. 1458–1465.

[8] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 7, pp. 1294–1309, 2008.

[9] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Proceedings of NIPS*, 2008, pp. 1033–1040.

[10] L.J. Li and F.F. Li, "What, where and who? classifying events by scene and object recognition," in *Proceedings of the 11th ICCV*. IEEE, 2007, pp. 1–8.

[11] F.F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of the 18th CVPR*. IEEE, 2005, vol. 2, pp. 524–531.

[12] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[13] F.F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *Workshop of the 17th CVPR*. IEEE, 2004, p. 178.

[14] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.