Discriminative Neighborhood Preserving Dictionary Learning for Image Classification

Shiye Zhang[✉], Zhen Dong, Yuwei Wu, and Mingtao Pei

Beijing Laboratory of Intelligent Information Technology,
School of Computer Science, Beijing Institute of Technology, Beijing 100081, China
{zhangshiye,dongzhen,wuyuwei,peimt}@bit.edu.cn

Abstract. In this paper, a discriminative neighborhood preserving dictionary learning method is proposed. The geometrical structure of the feature space is used to preserve the similarity information of the features, and the features' class information is employed to enhance the discriminative power of the learned dictionary. The Laplacian matrix which expresses the similarity information and the class information of the features is constructed and used in the objective function. Experimental results on four public datasets demonstrate the effectiveness of the proposed method.

Keywords: Dictionary learning · Image classification · Similarity · Discriminative

1 Introduction

Sparse representation has been widely studied due to its promising performance [1–4]. It can be used in image classification [5–10], face recognition [11–14], image retrieval [15], and image restoration [16]. The basic idea is to represent an input signal as a sparse linear combination of the atoms in the dictionary. Since the dictionary quality is a critical factor for the performance of the sparse presentation, lots of approaches focus on learning a good dictionary. Aharon *et al.* [17] presented the K-SVD algorithm, which iteratively updated the sparse codes of samples based on the current dictionary, and optimized the dictionary atoms to better fit the data. The discriminative information resided in the training samples might be ignored in this method. To solve this problem, some approaches [18–24] aim to learn more discriminative dictionaries. Mairal *et al.* [22] added a discriminative reconstruction constraint in the dictionary learning model to gain discrimination ability. Pham *et al.* [23] proposed a joint learning and dictionary construction method with consideration of the linear classifier performance. Yang *et al.* [24] employed the Fisher discrimination criterion to learn a structured dictionary.

However, in these methods, features are used separately while learning the dictionary, which results that the similarity information between the features is lost. Similar features in the same class thus may be encoded as dissimilar codes,

while features in different classes may be encoded as similar codes with the learned dictionary. In order to alleviate this problem, we propose a discriminative neighborhood preserving dictionary learning method that explicitly takes the similarity and class information of features into account. Figure 1 shows the idea of our method. The circle represents the feature x_i's neighborhood which is composed of features close to the x_i. Some of the neighbors are with the same label as x_i, and others are not. Our method encourages the distance between the codes of x_i and its neighbors in the same class as small as possible, at the same time maintains the distance between the codes of x_i and its neighbors in different classes. The learned dictionary can ensure that similar features in the same class could be encoded as similar codes and the features in different classes could be encoded as dissimilar codes.
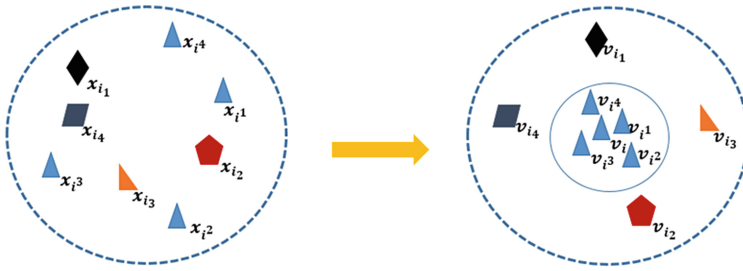
Fig. 1. The basic idea of our method. The left is the neighborhood of feature x_i which contains the features close to x_i. The different colors represent the neighbors of different classes. The blue features in the neighborhood are the neighbors of x_i with the same label. The neighbors with the same label are expected to be encoded close to the code of x_i, while other neighbors are expected to be encoded distant. Therefore, our method is more discriminative for classification.

Inspired by [25,26], we construct a Laplacian matrix which expresses the relationship between the features. The dictionary learned with this Laplacian matrix can well characterize the similarity of the similar features and preserve the consistence in sparse codes of the similar features. Different from [25,26], the class information is taken into account to further enhance the discriminative power of the dictionary in our method. Through introducing the class information, the Laplacian matrix is not only with the similarity information of the features in the same class but also can distinguish features in different classes. By adding the Laplacian term into the dictionary learning objective function, our method is able to learn a more discriminative dictionary. The experimental results demonstrate the encoding step is efficient with the learned discriminative dictionary and the classification performance of our method is improved with the dictionary.

The rest of this paper is organized as follows. In Sect. 2, we provide a brief description of the sparse presentation problem and introduce our discriminative neighborhood preserving dictionary learning method. In Sect. 3, the optimization

scheme of our method is presented, including learning sparse codes and learning the dictionary. The experimental results and discussions are displayed in Sect. 4. Finally, we conclude the paper in Sect. 5.

2   Discriminative Neighborhood Preserving Dictionary Learning Method

2.1   Sparse Representation Problem

We briefly review sparse representation. Given a data matrix $X = [x_1, \cdots, x_n] \in R^{d \times n}$, dictionary matrix $D = [d_1, \cdots, d_k] \in R^{d \times k}$, where each d_i represents a basis vector in the dictionary, coefficient matrix $V = [v_1, \cdots, v_n] \in R^{k \times n}$, where each column is a sparse representation for a data point. Each data point x_i can be represented as a sparse linear combination of those basis vectors in the dictionary. The objective function of sparse presentation can be formulated as follows:

$$\min \sum_{j=1}^{n} \|v_i\|_0 \qquad s.t. X = DV \tag{1}$$

$\|v_i\|_0$ is the number of nonzero entries of v_i, representing the sparseness of v_i. However, the minimization problem for this sparse representation with l_0 norm is shown to be an NP-hard problem [27]. The most widely used approach is to replace the l_0 norm with its l_1 norm. With the loss function, the objective function then becomes

$$\min_{D,V} \|X - DV\|_F^2 + \lambda \sum_{i=1}^{n} \|v_i\|_1 \qquad s.t. \|d_i\|^2 \le c, \quad i = 1, \ldots, k \tag{2}$$

The first term in Eq. (2) represents the reconstruction error, λ is the parameter used to balance the sparsity and the reconstruction error.

2.2   Formulation of Discriminative Neighborhood Preserving Dictionary Learning

In most current methods, the features are used separately while learning the dictionary. The similarity information among the features is lost which lead to the similar features can be encoded as totally different codes. In order to alleviate this problem, we propose a discriminative neighborhood preserving dictionary learning method. The dictionary learned by our method can well represent the intrinsic geometrical structure of the features to better characterize the relationship between the features and get more discriminative power through the features' class information.

Given the training features set $X = \{x_1, x_2, \ldots, x_n\}$ and the label of the training features. For each feature x_i, we choose l-nearest neighbors of x_i in the

same class to form $\{x_{i^1}, x_{i^2}, \ldots, x_{i^l}\}$ and choose m-nearest neighbors of x_i in different classes to form $\{x_{i_1}, x_{i_2}, \ldots, x_{i_m}\}$. All of these neighbors make up a local neighborhood of x_i which can be represented as $X_i = \{x_{i^1}, x_{i^2}, \ldots, x_{i^l}, x_{i_1}, x_{i_2}, \ldots, x_{i_m}\}$. $V_i = \{v_{i^1}, v_{i^2}, \ldots, v_{i^l}, v_{i_1}, v_{i_2}, \ldots, v_{i_m}\}$ is the codes of X_i about the dictionary. As shown in Fig. 1, the purpose of our method is to learn a discriminative dictionary which make the distance between v_i and its neighbors in the same class as small as possible and the distance between v_i and its neighbors in different classes as large as possible

$$\min \sum_{i=1}^{n} (\sum_{j=1}^{l} \|v_i - v_{ij}\|^2 - \beta \sum_{p=1}^{m} \|v_i - v_{i_p}\|^2) \tag{3}$$

β is the metric factor. We define W as the similarity matrix corresponding to the features, whose entry W_{ij} measures the similarity between x_i and x_j. If x_i is among   the l-nearest neighbors in the same class of x_j or x_j is among the l-nearest neighbors in the same class of x_i, $W_{ij} = 1$. If x_i is among the m-nearest neighbors in different classes of x_j or x_j is among the m-nearest neighbors in different classes of x_i, $W_{ij} = -\beta$, otherwise, $W_{ij} = 0$. Through the similarity matrix, the Eq. (3) can be represented as

$$\min \sum_{i=1}^{n} (\sum_{j=1}^{l} \|v_i - v_{ij}\|^2 - \beta \sum_{p=1}^{m} \|v_i - v_{i_p}\|^2) = \min \sum_{i=1}^{n} \sum_{j=1}^{l} \|v_i - v_j\|^2 W_{ij} \tag{4}$$

We define the degree of x_i as $d_i = \sum_{j=1}^{n} W_{ij}$, and $D = diag(d_1, \ldots, d_n)$. The Eq. (4) can be converted as [28]

$$\frac{1}{2} \min \sum_{i=1}^{n} \sum_{j=1}^{l} \|v_i - v_j\|^2 W_{ij} = \min Tr(VLV^T) \tag{5}$$

where $L = D - W$ is the Laplacian matrix. By adding this Laplacian term into the sparse presentation, we get the objective function of our method:

$$\min_{D,V} \|X - DV\|_F^2 + \lambda \sum_{i=1}^{n} \|v_i\|_1 + \alpha Tr(VLV^T) \qquad s.t. \|d_i\|^2 \le c, \quad i = 1, \ldots, k \tag{6}$$

Due to the Laplacian term, both the similarity among the features and the class information are considered during the process of dictionary learning and the similarity of codes among the similar features can be maximally preserved.

The Eq. (6) is not convex for D and V simultaneously, but it is convex for D when V is fixed and it is also convex for V when D is fixed. Motivated by the work in [29], we propose the following two-stage strategy to solve the Eq. (6): learning the codes V while fixing the dictionary D, and learning dictionary D while fixing the codes V.

3   Optimization

3.1   Learning Codes V

When fixing the dictionary D, Eq. (6) becomes the following optimization problem:

$$\min_V \|X - DV\|_F^2 + \lambda \sum_{i=1}^{n} \|v_i\|_1 + \alpha Tr(VLV^T) \tag{7}$$

Equation (7) is an L1-regularized least squares problem. This problem can be solved by several approaches [30,31]. Instead of optimizing the whole codes matrix V, we optimize each v_i one by one until the whole V converges following [26,32]. The vector form of Eq. (7) can be written as

$$\min \sum_{i=1}^{n} \|x_i - Dv_i\|^2 + \lambda \sum_{i=1}^{n} \|v_i\|_1 + \alpha \sum_{i,j=1}^{n} L_{ij} v_i^T v_j \tag{8}$$

When updating v_i, the other codes $v_j (j \neq i)$ are fixed. We rewrite the optimization with respect to v_i as follow:

$$\min_{v_i} f(v_i)\|x_i - Dv_i\|^2 + \lambda \sum_{j=1}^{k} |v_i^{(j)}| + \alpha L_{ii} v_i^T h_i \tag{9}$$

where $h_i = 2\alpha(\sum_{j \neq i} L_{ij} v_j)$, $v_i^{(j)}$ is the j-th coefficient of v_i. We use the feature-sign search algorithm in [29] to solve this problem. Define $h(v_i) = \|x_i - Dv_i\|^2 + \alpha L_{ii} v_i^T v_i + v_i^T h_i$, then $f(v_i) = h(v_i) + \lambda \sum_{j=1}^{k} |v_i^{(j)}|$. If we know the signs (positive, zero, or negative) of the $v_i^{(j)}$ at the optimal value, we can use either $v_i^{(j)}$ (if $v_i^{(j)} > 0$), $-v_i^{(j)}$ (if $v_i^{(j)} < 0$), or 0 (if $v_i^{(j)} = 0$) to replace each of the terms $|v_i^{(j)}|$. Considering only nonzero coefficients, the Eq. (9) is reduced to a standard, unconstrained quadratic optimization problem, which can be solved analytically and efficiently. When we update each v_i in the algorithm, maintaining an active set of potentially nonzero coefficients and their corresponding signs (all other coefficients must be zero). Our purpose is to search for the optimal active set and coefficient signs which minimize the objective function. The algorithm proceeds in a series of feature-sign steps: on each step, it is given the active set and the signs of current target, then it computes the analytical solution about the Eq. (9) and updates the solution, the active set and the signs using an efficient discrete line search between the current solution and the analytical solution. The detailed steps of the algorithm are stated in Algorithm 1.

3.2   Learning Dictionary D

In this section, we present a method for learning the dictionary D while fixing the coefficients matrix V. Equation (6) reduces to the following problem:

$$\min_D \|X - DV\|_F^2 \qquad s.t.\|d_i\|^2 \leq c, i = 1, ..., k \tag{12}$$

Algorithm 1. Feature-sign Search Algorithm for Solving Eq. (9)

Input: the training data set X, the dictionary D, the Laplacian matrix L, the
    parameters α, λ, the initial codes V.
Output: the optimal codes matrix V.
For $1 \leq n \leq n$ do
Step 1. Initialization
If not exist(V)
$v_i = \overrightarrow{0}, \theta = \overrightarrow{0}$, active set is empty.
else
$\theta_j \in \{-1, 0, 1\}$ denotes sign $(v_i^{(j)})$, add $j(v_i^{(j)} \neq 0)$ to the active set.
Step 2. Activate
We define $\nabla_i^{(j)} h(v_i)$ as the subgradient value of the $h(v_i)$. From zero coefficients
of v_i, select $j = argmax_j |\nabla_i^{(j)} h(v_i)|$, active $v_i^{(j)}$ only if it locally improves the
equation (9), namely:
if $\nabla_i^{(j)} h(v_i) > \lambda$, then set $\theta_i = -1$, active set $= \{j\} \bigcup$ active set.
if $\nabla_i^{(j)} h(v_i) < -\lambda$, then set $\theta_i = 1$, active set $= \{j\} \bigcup$ active set.
Step 3. Feature-sign
Let $\widehat{D}$ be a submatrix of D that contains only the columns corresponding to the
active set. Let $\widehat{v_i}, \widehat{h_i}, \widehat{\theta}$ be the subvectors of v_i, h_i, θ corresponding to the active
set. Compute the analytical solution to the resulting unconstrained QP:

$$\min g(\widehat{v_i}) = \|x_i - \widehat{D}\widehat{v_i}\|^2 + \alpha L_{ii}\widehat{v_i}^T \widehat{v_i} + \widehat{v_i}^T \widehat{h_i} + \lambda \widehat{\theta}^T \widehat{v_i} \qquad (10)$$

Let $(\partial g(\widehat{v_i})/\partial \widehat{v_i}) = 0$,we can get the optimal value of v_i under the current
active set:

$$\widehat{v_i}^{new} = (\widehat{D}^T D + \alpha L_{ii} I)^{-1}(\widehat{D}^T x_i - (\lambda\widehat{\theta} + \widehat{h_i})/2) \qquad (11)$$

Perform a discrete line search on the closed line segment from $\widehat{v_i}$ to $\widehat{v_i}^{new}$, check
the objective value at $\widehat{v_i}^{new}$ and all points where any coefficient changes value,
update $\widehat{v_i}$ (and the corresponding entries in v_i) to the point with the lowest
objective value, remove zero coefficients from the active set and update
$\theta = sign(v_i)$.
Step 4. Check the optimality conditions
Condition (a): Optimality condition for nonzero coefficients:
$\nabla_i^{(j)} h(v_i) + \lambda sign(v_i^{(j)}) = 0, \forall v_i^{(j)} \neq 0$. If condition (a) is not satisfied, go to
Step 3(without any new activation); Else check condition(b).
Condition (b): Optimality condition for zero coefficients:
$\nabla_i^{(j)} h(v_i) \leq \lambda, \forall v_i^{(j)} = 0$. If condition (b) is not satisfied, go to Step 3; otherwise
return v_i as the solution, and update the V with current v_i.
End for

Equation (12) is a least squares problem with quadratic constraints. It can be
efficiently solved by a Lagrange dual method [29].

Let $\lambda = [\lambda_1, ..., \lambda_k]$, and λ_i is the Lagrange multiplier associated with the
i-th inequality constraint $\|d_i\|^2 - c \leq 0$, we obtain the Lagrange dual function:

$$\min_{D} L(D, \lambda) = Tr((X - DV)^T(X - DV)) + \sum_{j=1}^{n} \lambda_j (\sum_{i=1}^{k} d_{ij}^2 - c) \qquad (13)$$

Define $\Lambda = diag(\lambda)$, Eq. (13) can be written as

$$\min_{D} L(D, \lambda) = Tr(X^T X - XV^T(VV^T + \Lambda)^{-1}(XV^T)^T - c\Lambda) \qquad (14)$$

The optimal solution is obtained by letting the first-order derivative of Eq. (14) equal to zero

$$D^* = XV^T(VV^T + \Lambda)^{-1} \qquad (15)$$

Substituting Eqs. (15) into (14), the Lagrange dual function becomes:

$$\min_{\Lambda} Tr(XV^T(VV^T + \Lambda)^{-1}VX^T) + cTr(\Lambda) \qquad (16)$$

We optimize the Lagrange dual Eq. (16) using the conjugate gradient. After obtaining the optimal solution Λ^*, the optimal dictionary D can be represented by $D^* = XV^T(VV^T + \Lambda^*)^{-1}$.

4 Experiments

In this section, we evaluate our method on four public datasets for image classification: Scene 15, UIUC-Sport, Caltech-101, and Caltech-256. For each experiment, we describe the information of datasets and detailed settings. The effectiveness of our method is validated by comparisons with popular methods.

4.1 Parameters Setting

In the experiment, we first extract SIFT descriptors from 16×16 patches which are densely sampled using a grid with a step size of 8 pixels to fairly compare with others. Then we extract the spatial pyramid feature based on the extracted SIFT features with three grids of size 1×1, 2×2 and 4×4. In each spatial sub-region of the spatial pyramid, the codes are pooled together by max pooling method to form a pooled feature. These pooled features from each sub-region are concatenated and normalized by L2 normalization as the final spatial pyramid features of the images. The dictionary in the experiment is learned by these spatial pyramid features.

In our method, the weight of the Laplacian term α, the sparsity of the coding λ, and the constraints of the neighborhood in different classes β play more important roles in dictionary learning. According to our observation, the performance is good when β is fixed at 0.2 for Scene 15 and UIUC-Sport. For Caltech-101 and Caltech-256, 0.1 is much better for β. For Scene 15, the value of α is 0.2 and the value of λ is 0.4. For UIUC-Sport, Caltech-101, and Caltech-256, the value of α is 0.1 and the value of λ is 0.3.

4.2   Scene 15 Dataset

Scene 15 dataset contains 15 categories. Each category contains 200 to 400 images and the total image number is 4485. In order to compare with other work, we use the same setting to choose the training images. We randomly choose 100 images per category and test on the rest. This process is repeated for ten times to obtain reliable results.

Table 1 gives the performance comparison of our method and several other methods on the Scene 15 dataset. We can see that our method can achieve high performance on scene classification. It outperforms ScSPM by nearly 11 % by considering the geometrical structure of the feature space based on sparse representation and outperforms LScSPM by nearly 2 % by adding the class information. Both of them demonstrate the effectiveness of our method. Our discriminative neighborhood preserving dictionary learning method can not only make use of the geometrical structure of the feature space to preserve more similarity information, but also make the final dictionary more discriminative by considering the class information which can improve the image classification performance.

Table 1. Performance comparison on the Scene-15 dataset

Methods	Accuracy (%)
KSPM [33]	81.40 ± 0.5
KCSPM [35]	76.70 ± 0.40
ScSPM [6]	80.28 ± 0.93
HIK+OCSVM [34]	84.00 ± 0.46
LScSPM [25]	89.75 ± 0.50
LR-Sc^+SPM [7]	90.03 ± 0.70
DSC [9]	84.21 ± 0.44
DLMM [10]	83.67 ± 0.49
Our method	91.23 ± 0.84

4.3   UIUC-Sport Dataset

UIUC-Sport dataset contains 8 categories for image-based event classification and 1792 images in all. These 8 categories are badminton, bocce, croquet, polo, rock climbing, rowing, sailing and snow boarding. The size of each category ranges from 137 to 250. Following the standard setting for this dataset, we randomly choose 70 images from each class for training and test on the rest images. We repeat this process for ten times for fair comparison.

Table 2 gives the performance comparison of our method and several other methods on the UIUC-Sport dataset. We can see that our method outperforms ScSPM by nearly 5 % and outperforms LScSPM by nearly 2 %. This demonstrates the effectiveness of our proposed method.

Table 2. Performance comparison on the UIUC-Sport dataset

Methods	Accuracy (%)
ScSPM [6]	82.74 ± 1.46
HIK+OCSVM [34]	83.54 ± 1.13
LR-Sc^+SPM [7]	86.69 ± 1.66
LScSPM [25]	85.31 ± 0.51
DSC [9]	83.72 ± 1.68
DLMM [10]	86.93 ± 0.99
Our method	87.13 ± 1.02

4.4   Caltech-101 Dataset

The Caltech-101 dataset contains 9144 images in 101 classes with high intra-class appearance shape variability. The number of images per category varies from 31 to 800. We follow the common experimental setup and randomly choose 30 images per category for training and the rest for testing. This process is repeated for ten times.

The average classification rates of our method and several other methods on Caltech-101 dataset are reported in Table 3. From these results, we see that our method performs better than most existing methods. As compared to the LLC, our method makes a $2.4\,\%$ improvement. It demonstrates the effectiveness of our proposed method.

Table 3. Performance comparison on the Caltech-101 dataset

Methods	Accuracy (%)
KSPM [33]	64.40 ± 0.80
KCSPM [35]	64.14 ± 1.18
ScSPM [6]	73.20 ± 0.54
LLC [36]	73.44
LR-Sc^+SPM [7]	75.68 ± 0.89
DSC [9]	71.96 ± 0.83
DLMM [10]	74.87 ± 0.67
Our method	75.86 ± 0.78

4.5   Caltech-256 Dataset

Caltech-256 dataset contains 256 categories and a background class in which none of the image belongs to those 256 categories. The number of images is 29780 with much higher intra-class variability and higher object location variability as compared to Caltech-101. Therefore Caltech-256 is a very challenging dataset so far for object recognition and classification. The number of images per category is no less than 80. We randomly choose 30 images per category for training and repeat this process for ten times.

Table 4. Performance comparison on the Caltech-256 dataset

Methods	Accuracy (%)
KSPM [33]	34.10
KCSPM [35]	27.17 ± 0.46
ScSPM [6]	34.02 ± 0.35
LScSPM [25]	35.74 ± 0.10
DLMM [10]	36.22 ± 0.33
Our method	37.81 ± 0.21

The average classification rates of our method and several other methods on Caltech-256 dataset are reported in Table 4. We can see that our method can achieve the state-of-the-art performances on this dataset.

5   Conclusion

In this paper, we propose a discriminative neighborhood preserving dictionary learning method for image classification. We consider the geometrical structure of the feature space in the process of dictionary learning to preserve the similarity information of the features. By introducing the class information, the discriminative power of the learned dictionary is enhanced. The learned dictionary can ensure that the similar features in the same class are encoded as similar codes and the features in different classes are encoded as dissimilar codes. Experimental results on four public datasets demonstrate the effectiveness of our method.

References

1. Chen, X., Zou, D., Li, J., Cao, X., Zhao, Q., Zhang, H.: Sparse dictionary learning for edit propagation of high-resolution images. In: 27th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2854–2861. IEEE Press, Columbus (2014)
2. Lan, X., Ma, A.J., Yuen, P.C.: Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation. In: 27th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1194–1201. IEEE Press, Columbus (2014)
3. Cherian, A.: Nearest neighbors using compact sparse codes. In: 31st International Conference on Machine Learning, pp. 1053–1061. ACM Press, Beijing (2014)
4. Eldar, Y.C., Mishali, M.: Robust recovery of signals from a structured union of subspaces. IEEE Trans. Inf. Theory **55**(11), 5302–5316 (2009)

5. Liu, B.-D., Wang, Y.-X., Shen, B., Zhang, Y.-J., Hebert, M.: Self-explanatory sparse representation for image classification. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part II. LNCS, vol. 8690, pp. 600–616. Springer, Heidelberg (2014)

6. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: 22th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1794–1801. IEEE Press, Miami (2009)

7. Zhang, C., Liu, J., Tian, Q., Xu, C., Lu, H., Ma, S.: Image classification by non-negative sparse coding, low-rank and sparse decomposition. In: 24th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1673–1680. IEEE Press, Colorado (2011)

8. Yuan, X.T., Liu, X., Yan, S.: Visual classification with multitask joint sparse representation. IEEE Trans. Image Process. **21**(10), 4349–4360 (2012)

9. Liu, B.D., Wang, Y.X., Zhang, Y.J., Zheng, Y.: Discriminant sparse coding for image classification. In: 37th IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2193–2196. IEEE Press, Kyoto (2012)

10. Liu, B.D., Wang, Y.X., Zhang, Y.J., Shen, B.: Learning dictionary on manifolds for image classification. Pattern Recogn. **46**(7), 1879–1890 (2013)

11. Yang, M., Dai, D., Shen, L., Gool, L.V.: Latent dictionary learning for sparse representation based classification. In: 27th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 4138–4145. IEEE Press, Columbus (2014)

12. Yang, M., Zhang, L., Yang, J., Zhang, D.: Metaface learning for sparse representation based face recognition. In: 17th International Conference on Image Processing, pp. 1601–1604. IEEE Press, Hong Kong (2010)

13. Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mobahi, H., Ma, Y.: Toward a practical face recognition system: robust alignment and illumination by sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **34**(2), 372–386 (2012)

14. Yang, M., Zhang, D., Yang, J.: Robust sparse coding for face recognition. In: 24th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 625–632. IEEE Press, Colorado (2011)

15. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: 20th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE Press, Minneapolis (2007)

16. Mairal, J., Elad, M., Sapiro, G.: Sparse representation for color image restoration. IEEE Trans. Image Process. **17**(1), 53–69 (2008)

17. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing over-complete dictionaries for sparse representation. IEEE Trans. Sig. Process. **54**(11), 4311–4322 (2006)

18. Cao, X., Ren, W., Zuo, W., Guo, X., Foroosh, H.: Scene text deblurring using text-specific multiscale dictionaries. IEEE Trans. Image Process. **24**(4), 1302–1314 (2015)

19. Zhou, N., Fan, J.: Jointly learning visually correlated dictionaries for large-scale visual recognition applications. IEEE Trans. Pattern Anal. Mach. Intell. **36**(4), 715–730 (2014)

20. Gao, S., Tsang, I.H., Ma, Y.: Learning category-specific dictionary and shared dictionary for fine-grained image categorization. IEEE Trans. Image Process. **23**(2), 623–634 (2014)

21. Bao, C., Quan, Y., Ji, H.: A convergent incoherent dictionary learning algorithm for sparse coding. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VI. LNCS, vol. 8694, pp. 302–316. Springer, Heidelberg (2014)
22. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Discriminative learned dictionaries for local image analysis. In: 21th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE Press, Anchorage (2008)
23. Pham, D.S., Venkatesh, S.: Joint learning and dictionary construction for pattern recognition. In: 21th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE Press, Anchorage (2008)
24. Yang, M., Zhang, D., Feng, X.: Fisher discrimination dictionary learning for sparse representation. In: 13th IEEE International Conference on Computer Vision, pp. 543–550. IEEE Press, Barcelona (2011)
25. Gao, S., Tsang, I.W., Chia, L.T., Zhao, P.: Local features are not lonely Laplacian sparse coding for image classification. In: 23th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3555–3561. IEEE Press, San Francisco (2010)
26. Zheng, M., Bu, J., Chen, C., Wang, C., Zhang, L., Qiu, G., Cai, D.: Graph regularized sparse coding for image representation. IEEE Trans. Image Process. **20**(5), 1327–1336 (2011)
27. Donoho, D.L.: For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution. Commun. Pure Appl. Math. **59**(6), 797–829 (2006)
28. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In: 15th Annual Conference on Neural Information Processing Systems, vol. 14, pp. 585–591. MIT Press, Vancouver (2001)
29. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: 20th Annual Conference on Neural Information Processing Systems, pp. 801–808. MIT Press, Vancouver (2006)
30. Koh, K., Kim, S.J., Boyd, S.P.: An interior-point method for large-scale L1-regularized logistic regression. J. Mach. Learn. Res. **8**(8), 1519–1555 (2007)
31. Andrew, G., Gao, J.: Scalable training of L1-regularized log-linear models. In: 24st International Conference on Machine Learning, pp. 33–40. ACM Press, Oregon (2007)
32. Gao, S., Tsang, I.H., Chia, L.T.: Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 92–104 (2013)
33. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: 19th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178. IEEE Press, New York (2006)
34. Wu, J., Rehg, J.M.: Beyond the euclidean distance: creating effective visual codebooks using the histogram intersection kernel. In: 12th IEEE International Conference on Computer Vision, pp. 630–637. IEEE Press, Kyoto (2009)
35. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: 21th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE Press, Anchorage (2008)
36. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: 23th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3360–3367. IEEE Press, San Francisco (2010)