A FLEXIBLE DIRTY MODEL DICTIONARY LEARNING APPROACH FOR CLASSIFICATION

Jiaming Qi[1], *Wei Chen*[1,2]

[1]State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, China
[2]Beijing Engineering Research Center of High-speed Railway Broadband Mobile Communications
Corresponding author: *Wei Chen*
15120125@bjtu.edu.cn, weich@bjtu.edu.cn

ABSTRACT

Various dictionary learning methods have gained tremendous success for signal classification. However, traditional dictionary learning methods for classification assume there is no outlier in the training data, which may not be the case in practical applications. In this paper, we propose a new discriminative dictionary learning framework for classification, which simultaneously learns a discriminative dictionary and detects outliers in the data. We formulate the dictionary learning framework into an optimization problem with designed regularizers to promote both the discrimination and outlier-detection capability. An efficient and effective iterative algorithm based on the alternating direction method of multipliers (ADMM) is provided to solve the proposed optimization problem. We demonstrate the superior performance of the proposed approach in comparison with state-of-the-art methods on some image classification tasks.

Index Terms— Dictionary learning, classification, sparse representation.

I. INTRODUCTION

The theory and algorithms of sparse signal representation have made a rapid development in the past years. They have been playing a central role in signal compression and proved to be very useful in many applications in the filed of signal processing, such as signal acquisition, denoising, and classification [1]–[9]. In the perspective of the sparse signal representation, a signal can be seen as a linear combination of a few atoms selected from a complete or over-complete dictionary, where each atom represents as a column of the dictionary.

The goal of dictionary learning is to learn a basis from a collection of signals so that they can be sparsely represented. By using a dictionary, a signal is transformed into a new representation in a higher dimensional space, where somewhat challenging problems, e.g., classification, may become easier. In contrast to the sparse representation task which concerns the approximation accuracy, the goal of classification is to determine the correct class label for the query signal. Therefore, it would be beneficial to make the learned dictionary have discriminative capability. Existing discriminative dictionary learning approaches in literature can be roughly divided into two categories.

Approaches in the first category learn a class-specific sub-dictionary for each signal class, and these sub-dictionaries together constitute the complete dictionary [10]–[12]. However, the size of the sub-dictionary for each class needs to be predefined, and these approaches involving sub-dictionaries are not scalable with a large number of classes. Approaches in the second category learn a dictionary that is shared by all classes [6], [13], [14]. All of these methods exploit all training samples to learn a dictionary.

Generally, it is more likely to obtain a better discriminative dictionary if more training data are given and the learning algorithm is designed appropriately. In most of the existing work on dictionary learning for classification, it is assumed there is no outlier in the training data, which may not be the case in practical applications. In this paper, we propose a new discriminative dictionary learning framework, which simultaneously learns a discriminative dictionary and detects outliers in the data.

II. DICTIONARY LEARNING FRAMEWORK

Let $\mathbf{x} \in \mathbb{R}^m$ be an m dimensional signal with class label $i \in \{1, \ldots, L\}$, where L denotes the number of classes. The training set with N signals is denoted as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] = [\mathbf{X}_1, \ldots, \mathbf{X}_L]$, where $\mathbf{X}_i$ contains N_i training signals belonging to class i. The dictionary is denoted as $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K] \in \mathbb{R}^{m \times K} (m \leq K < N)$, where $\mathbf{d}_k$ $(k = 1, \ldots, K)$ denotes the kth atom of the dictionary. Columns of the matrix $\mathbf{C} \in \mathbb{R}^{K \times N}$ denote sparse representations of training signals $\mathbf{X}$. In dictionary learning literatures, it is often assumed there is no outlier in the training data, which may not be the case in practical applications. Therefore, it is desired to detect and remove the outliers and only use the remaining training samples to learn a dictionary.

To identify outliers, one could use either empirical domain knowledge or some model to distinguish outliers and valuable data. However, empirical domain knowledge is not

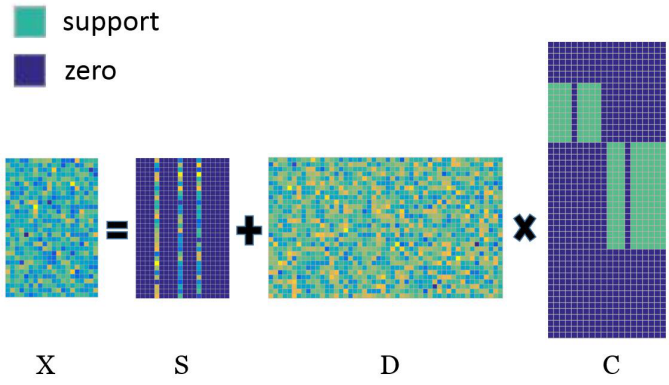

Fig. 1. An illustration of the dictionary learning model for classification. Assume that training data $\mathbf{X}$ belongs to two classes, and dictionary $\mathbf{D}$ is shared by all classes. We extract the outliers, i.e. $\mathbf{S}$, and train the dictionary with the rest training samples. The constructed sparse representation matrix for training data is $\mathbf{C}$ and the dictionary is $\mathbf{D}$.

easy to acquire in some applications, and an outlier could be close to a valuable data sample in terms of the Euclidean distance. To deal with this problem, we propose to detect outliers by using a dictionary where valuable data have sparse representations, while outliers cannot be represented by the dictionary with a sparse vector. With this principle for determining outliers, we could formulate the problem for simultaneously discriminative dictionary learning and outliers detection. A visual illustration of this construction is shown in Fig. 1, where the nonzero columns in $\mathbf{S}$ represent outliers. Now we can formulate the dictionary learning problem as

$$\min_{\mathbf{D},\mathbf{C},\mathbf{S}} \|\mathbf{X}-\mathbf{S}-\mathbf{D}\mathbf{C}\|_F^2+\lambda_1 f(\mathbf{C})+\lambda_2 h(\mathbf{C})+\lambda_3 \left\|\mathbf{S}^T\right\|_{row}. \tag{1}$$

For the sparse representation matrix $\mathbf{C}$, the support of signals of the same class should be as close as possible, while to enhance discrimination, the supports of signals of different classes should have as less overlap as possible. In (1), $f(\mathbf{C})$ and $h(\mathbf{C})$ denote regularizers that capture intra-class similarity and inter-class discrimination, respectively. $\left\|\mathbf{S}^T\right\|_{row}$ penalizes the transpose of $\mathbf{S}$ to have a row-sparse structure, and the non-zero columns of $\mathbf{S}$ correspond to the detected outliers. λ_1, λ_2 and λ_3 are weights used to balance different terms in (1).

To capture similarity of signals from a same class, the sparse representation matrix of the same class, i.e., $\mathbf{C}_i$ $(i = 1,\ldots,L)$, is modeled with a row-sparse structure, where elements in each row of $\mathbf{C}_i$ are either all zero or mostly non-zero. The non-zero rows correspond to dictionary atoms among signals in the same class. However, directly minimizing the number of non-zero rows in a matrix leads to an NP hard problem. To facilitate algorithm derivation, we use the convex relaxed ℓ_2/ℓ_1 norm instead as a measure of row-sparsity. Therefore, we design the regularization term $f(\mathbf{C})$ as

$$f(\mathbf{C}) = \sum_{i=1}^{L} \|\mathbf{C}_i\|_{2,1} = \sum_{i=1}^{L}\sum_{k=1}^{K} \left\|\mathbf{c}_i^k\right\|_2, \tag{2}$$

where $\mathbf{C}_i \in \mathbb{R}^{K\times N_i}$ is a sub-matrix of $\mathbf{C}$, denoting the sparse representation matrix for the ith class, and $\mathbf{c}_i^k$ denotes the kth row of the sparse representation matrix $\mathbf{C}_i$. We would like to allow different classes to have partly overlapped supports as in [15]. However, the approach in [15] is computationally expensive, and thus requires relatively high computational resource. Besides, the impact of outliers is not considered in [15].

Now we provide the second regularizer $h(\mathbf{C})$ in (1), which leads to the design of our computational efficient algorithm. To enhance the discrimination capability for the signal representations, the regularizer $h(\mathbf{C})$ is used to make the size of the overlapped support between $\mathbf{C}_i$ and $\mathbf{C}_j$ $(i \neq j)$ as small as possible. For example, considering two signal representations from distinct classes, the number of non-zero of their element-wise product is just the size of the overlapped support. Therefore, the discrimination regularization term can be modeled as

$$h(\mathbf{C}) = \sum_{i=1}^{L}\sum_{p=1}^{N_i} \left\|\mathbf{W}_{/i}\mathbf{c}_{i,p}\right\|_2^2, \tag{3}$$

where $\mathbf{c}_{i,p}$ denotes the pth column of $\mathbf{C}_i$, and $\mathbf{W}_{/i}$ is a diagonal matrix with the kth diagonal element is

$$w_{/i}^k = \left\|\mathbf{c}_{/i}^k\right\|_2, \quad (k = 1,\ldots,K), \tag{4}$$

where $\mathbf{c}_{/i}^k$ denotes the kth row of the matrix $\mathbf{C}_{/i}$, and $\mathbf{C}_{/i} \in \mathbb{R}^{K\times(N-N_i)}$ is a sub-matrix of $\mathbf{C}$ generated by removing the columns in $\mathbf{C}_i$.

Therefore, the discrimination regularization term $h(\mathbf{C})$ can be expressed as

$$h(\mathbf{C}) = \sum_{i=1}^{L} \left\|\mathbf{W}_{/i}\mathbf{C}_i\right\|_F^2. \tag{5}$$

Lastly, we provide the approximation of the third regularization term in (1). Because $\mathbf{S}^T$ is desired to have be row-sparse, we also employ the convex relaxed ℓ_2/ℓ_1 norm instead as a measure of row-sparsity.

$$\left\|\mathbf{S}^T\right\|_{row} = \left\|\mathbf{S}^T\right\|_{2,1}. \tag{6}$$

By substituting the designed regularization terms (2), (5) and (6) into the objective function of (1), the proposed dictionary learning framework can be described as the following

optimization problem

$$\min_{\mathbf{D},\mathbf{C},\mathbf{S}} \|\mathbf{X}-\mathbf{S}-\mathbf{D}\mathbf{C}\|_F^2+\lambda_1\sum_{i=1}^{L}\|\mathbf{C}_i\|_{2,1} +\lambda_2\sum_{i=1}^{L}\left\|\mathbf{W}_{/i}\mathbf{C}_i\right\|_F^2+\lambda_3\left\|\mathbf{S}^T\right\|_{2,1}. \tag{7}$$

The objective function of (7) involves three variables, where are not jointly convex for the optimization problem. Therefore, we exploit the coordinate decent method that solves the optimization problem by alternatively optimizing one variable with all the others fixed.

III. OPTIMIZATION ALGORITHM

Solving the optimization problem in (7) by coordinate decent method involves three sub-problems that are formulated by updating $\mathbf{C}$ with fixing $\mathbf{D}$ and $\mathbf{S}$; updating $\mathbf{S}$ with fixing $\mathbf{D}$ and $\mathbf{C}$; and updating $\mathbf{D}$ with fixing $\mathbf{C}$ and $\mathbf{S}$. These sub-problems are iteratively solved until the discriminative dictionary $\mathbf{D}$ and the sparse representations $\{\mathbf{C},\ \mathbf{S}\}$ are converged.

Firstly, given $\mathbf{D}$ and $\mathbf{S}$ fixed, the original optimization problem (7) turns into a sparse coding problem with the variable $\mathbf{C}=[\mathbf{C}_1,\ldots,\mathbf{C}_L]$. We update the sub-matrix $\mathbf{C}_i$ $(i=1,\ldots,L)$ class by class, with all the other $\mathbf{C}_j$ $(j\neq i)$ fixed. By removing unrelated terms, the optimization problem for each class is further reduced to

$$\min_{\mathbf{C}_i}\|\mathbf{X'}_i-\mathbf{D}\mathbf{C}_i\|_F^2+\lambda_1\|\mathbf{C}_i\|_{2,1}+\lambda_2\left\|\mathbf{W}_{/i}\mathbf{C}_i\right\|_F^2, \tag{8}$$

where $\mathbf{X'}_i=\mathbf{X}_i-\mathbf{S}_i$. To solve the optimization problem in (8), we employ the ADMM owning to its success in solving various sparse approximation related problems [16]–[18]. By introducing one auxiliary variable $\mathbf{Z}_i\in\mathbb{R}^{K\times N_i}$, the optimization problem in (8) can be reformulated as

$$\begin{aligned}\min_{\mathbf{C}_i,\mathbf{Z}_i}\quad & \|\mathbf{X'}_i-\mathbf{D}\mathbf{C}_i\|_F^2+\lambda_1\|\mathbf{Z}_i\|_{2,1}+\lambda_2\left\|\mathbf{W}_{/i}\mathbf{C}_i\right\|_F^2\\ \text{s.t.}\quad & \mathbf{C}_i=\mathbf{Z}_i.\end{aligned} \tag{9}$$

The augmented Lagrangian function can be formed as

$$\begin{aligned}L_\rho(\mathbf{C}_i,\mathbf{Z}_i,\mathbf{G})=&\|\mathbf{X'}_i-\mathbf{D}\mathbf{C}_i\|_F^2+\lambda_1\|\mathbf{Z}_i\|_{2,1}\\ &+\lambda_2\left\|\mathbf{W}_{/i}\mathbf{C}_i\right\|_F^2+\mathrm{tr}\left(\mathbf{G}^T(\mathbf{C}_i-\mathbf{Z}_i)\right)\\ &+\frac{\rho}{2}\|\mathbf{C}_i-\mathbf{Z}_i\|_F^2,\end{aligned} \tag{10}$$

where $\mathbf{G}\in\mathbb{R}^{K\times N_i}$ is the Lagrangian multiplier for the equation constraint in (10), and $\rho>0$ is a preselected penalty parameter. The augmented Lagrangian function (10) can be minimised over $\mathbf{C}_i$, $\mathbf{Z}_i$ and $\mathbf{G}$ iteratively by updating one variable at a time and fixing the others. The resulting algorithm is summarized in Algorithm 1. The Shrink function in Algorithm 1 updates $\mathbf{Z}_i$ by using row-wise shrinkage,

Algorithm 1 Shared supports sparse coding via ADMM

Input: Training data $\mathbf{X'}$, dictionary $\mathbf{D}$, number of classes L, regulariser parameters λ_1, λ_2, penalty parameter ρ.
Output: Sparse code $\mathbf{C}$.
Initialization: $\mathbf{C}^0=0$, $\mathbf{G}^0=0$, iteration index $t=0$.
For $i=1:L$
Do

1) Set the diagonal matrix $\mathbf{W}_{/i}$ by:
$$(w^k_{/i})^{t+1}=\left\|(\mathbf{c}^k_{/i})^t\right\|_2.$$
2) Fix $\mathbf{Z}_i$ $\mathbf{G}$, and update $\mathbf{C}_i$ by:
$$\begin{aligned}\mathbf{C}_i^{t+1}=&\arg\min_{\mathbf{C}_i}L\rho(\mathbf{C}_i,\mathbf{Z}_i^t,\mathbf{G}^t)\\ =&(\mathbf{D}^T\mathbf{D}+\lambda_2\mathbf{W}_{/i}^{(t+1)T}\mathbf{W}_{/i}^{t+1}+\rho\mathbf{I})^{-1}\\ &(\mathbf{D}^T\mathbf{X'}_i+\rho\mathbf{Z}_i^k-\frac{1}{2}\mathbf{G}^t)\end{aligned}$$
3) Fix $\mathbf{C}_i$ $\mathbf{G}$, and update $\mathbf{Z}_i$ by row-wise shrinkage:
$$\mathbf{Z}_i^{t+1}=\mathrm{Shrink}(\mathbf{C}_i^{t+1}+\frac{1}{\rho}\mathbf{G}^t,\frac{\lambda_1}{\rho})$$
4) Fix $\mathbf{C}_i$ $\mathbf{Z}_i$, and update the Lagrange multiplier $\mathbf{G}$:
$$\mathbf{G}^{t+1}=\mathbf{G}^t+\rho(\mathbf{C}_i^{t+1}-\mathbf{Z}_i^{t+1})$$

Increment t.
until convergence
end for

which can be represented as:

$$\mathbf{z}^k=\frac{\max\left\{\|\mathbf{r}^k\|_2-\frac{\lambda_1}{\rho},0\right\}}{\|\mathbf{r}^k\|_2}\mathbf{r}^k, \tag{11}$$

where $\mathbf{z}^k$, $\mathbf{r}^k$ denotes the kth row of the matrix $\mathbf{Z}_i$ and $\mathbf{R}=\mathbf{C}_i+\frac{1}{\rho}\mathbf{G}$, respectively.

After obtaining the coding matrix $\mathbf{C}$, we update the matrix $\mathbf{S}$ with $\mathbf{D}$ and $\mathbf{C}$ fixed. Here, the optimization problem can be formulated as

$$\min_{\mathbf{S}}\|\mathbf{X''}-\mathbf{S}\|_F^2+\lambda_3\left\|\mathbf{S}^T\right\|_{2,1}, \tag{12}$$

where $\mathbf{X''}=\mathbf{X}-\mathbf{D}\mathbf{C}$. Here we employ the column-wise shrinkage operation to solve the problem in (12). The solution is given by

$$\mathbf{s}_k=\frac{\max\left\{\|\mathbf{x''}_k\|_2-\frac{\lambda_3}{2},0\right\}}{\|\mathbf{x''}_k\|_2}\mathbf{x''}_k, \tag{13}$$

where $\mathbf{s}_k$ and $\mathbf{x''}_k$ denote the kth column of $\mathbf{S}$ and $\mathbf{X''}$, respectively.

We have described the process of updating $\mathbf{C}$ and $\mathbf{S}$ above, and now we update the dictionary $\mathbf{D}$ with $\mathbf{C}$ and $\mathbf{S}$ fixed. The objective function is reduced to

$$\min_{\mathbf{D}}\quad\|\mathbf{X}-\mathbf{S}-\mathbf{D}\mathbf{C}\|_F^2. \tag{14}$$

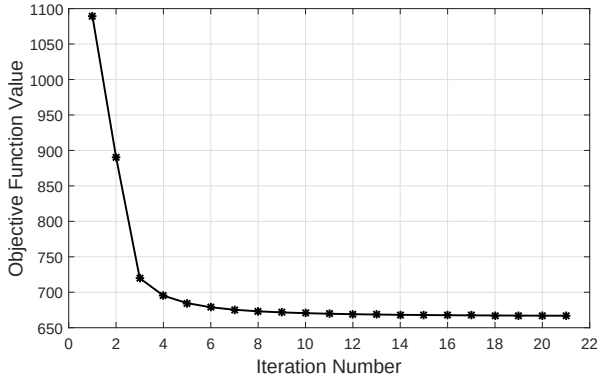

Fig. 2. The convergence curve on 15 scene categories dataset, where the number of training images per category is 100, and the dictionary size is K = 450.

The solution of the above least square estimation problem is directly given by

$$\mathbf{D} = (\mathbf{X} - \mathbf{S})\mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1}. \tag{15}$$

After learning the discriminative dictionary $\mathbf{D}$, any test signal can be classified based on its sparse representation over $\mathbf{D}$. Although various classifiers can be applied, we simply choose a linear classifier, as it is the most widely used approach in related literature [6], [15]. The algorithm is halted when the maximum number of iterations is reached or the value of the objective function (7) in adjacent iterations are sufficiently close. Fig. 2 demonstrates the convergence behaviour of the proposed approach by using the Fifteen Scene dataset [19].

IV. EXPERIMENT VALIDATION

In this section, we compare the proposed classification approach with Sparse Representation Classification (SRC) [10], K-SVD [20], Label-Consistent K-SVD (LC-KSVD) [6], the Fisher Discrimination Dictionary Learning algorithm (FDDL) [13], Low-Rand Shared Dictionary Learning (LRSDL) [21], and the Support Discrimination Dictionary Learning (SDDL) [15] using the Extended Yale B dataset [22], the AR face dataset [23] and the Fifteen scene dataset [19].

The Extended Yale B dataset contains 2414 frontal images of 38 people, the images are captured under different conditions. The images were cropped to 192×168 pixels, normalized and projected to a dimension of 504 using a random Gaussian matrix. We randomly select half of the images for training and cross validation and the rest for testing. The AR face dataset contains over 4000 color images of 100 subjects, we randomly choose 20 images per subject for training and the other 6 images for testing. Each face image is cropped to the dimension of 165×120, normalized and projected to a 540 dimension vector using a random Gaussian matrix. The fifteen scene dataset [19] contains 15 natural scene categories. The average image size is about 300 $\times$ 250 pixels. We use the Spatial Pyramid Features of the images as the input signal, each feature descriptor is a vector of dimension 3000, and we randomly select 100 images per category for training and the rest for testing.

Table I. Comparison of classification accuracy for various dictionary based methods.

Method	Extended Yale	AR	15 Scene
SRC	80.54	66.57	91.80
K-SVD	93.40	86.50	93.60
LC-KSVD	95.00	93.70	97.01
FDDL	94.92	94.10	97.92
LRSDL	98.00	97.33	98.12
SDDL	97.08	98.00	98.02
Proposed	**99.08**	**98.67**	**98.83**

The initialised dictionary for each experiment is generated by randomly selecting samples from the training data. The dictionary size of the Extended Yale B dataset, the AR dataset and the 15 scene dataset are 570, 500 and 450, respectively. For the proposed approach and all the compared methods, we employ the cross validation [24] to tune the parameters for the best performance if the optimal parameters are not reported in literature. The experimental results are summarized in Table I.

It can be seen that dictionary learning based methods perform better than SRC, which shows that better performance can be achieved by learning a discriminative dictionary. The proposed method achieves the highest classification accuracy and outperforms all the other competing approaches.

V. CONCLUSIONS

In this paper, we propose a new discriminative dictionary learning framework for classification. The proposed approach simultaneously learns a discriminative dictionary and detects outliers in the data. The proposed approach is evaluated by using five different datasets involving human faces, object images, hyperspectral images and scene images. The conducted experiments consistently demonstrate that the proposed approach yields good classification results, and outperforms the existing state-of-the-art dictionary learning based approaches for classification.

VI. REFERENCES

[1] Xuefeng Chen, Zhaohui Du, Jimeng Li, Xiang Li, and Han Zhang, "Compressed sensing based on dictionary learning for extracting impulse components," *Signal Processing*, vol. 96, pp. 94–109, 2014.

[2] W. Chen, I. J. Wassell, and M. R. D. Rodrigues, "Dictionary design for distributed compressive sensing," *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 95–99, Jan 2015.

[3] X. Ding, W. Chen, and I. J. Wassell, "Joint sensing matrix and sparsifying dictionary optimization for tensor compressive sensing," *IEEE Transactions on Signal Processing*, vol. 65, no. 14, pp. 3632–3646, July 2017.

[4] W. Chen and M. R. D. Rodrigues, "Dictionary learning with optimized projection design for compressive sensing applications," *IEEE Signal Processing Letters*, vol. 20, no. 10, pp. 992–995, Oct 2013.

[5] Ron Rubinstein, Michael Zibulevsky, and Michael Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1553–1564, 2010.

[6] Zhuolin Jiang, Zhe Lin, and Larry S Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2651–2664, 2013.

[7] W. Chen, "Simultaneous sparse bayesian learning with partially shared supports," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1641–1645, 2017.

[8] W. Chen, D. Wipf, Y. Wang, Y. Liu, and I. J. Wassell, "Simultaneous bayesian sparse approximation with structured sparse models," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6145–6159, Dec 2016.

[9] W. Chen, M. R. D. Rodrigues, and I. J. Wassell, "Projection design for statistical compressive sensing: A tight frame based approach," *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 2016–2029, April 2013.

[10] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[11] Jianchao Yang, Kai Yu, and Thomas Huang, "Supervised translation-invariant sparse coding," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3517–3524.

[12] Yuanming Suo, Minh Dao, Trac Tran, Hojjat Mousavi, Umamahesh Srinivas, and Vishal Monga, "Group structured dirty dictionary learning for classification," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 150–154.

[13] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *International Journal of Computer Vision*, vol. 109, no. 3, pp. 209–232, 2014.

[14] Sijia Cai, Wangmeng Zuo, Lei Zhang, Xiangchu Feng, and Ping Wang, "Support vector guided dictionary learning," in *European Conference on Computer Vision*. Springer, Cham, 2014, pp. 624–639.

[15] Yang Liu, Wei Chen, Qingchao Chen, and Ian Wassell, "Support discrimination dictionary learning for image classification," in *European Conference on Computer Vision*. Springer, 2016, pp. 375–390.

[16] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[17] Wei Deng, Wotao Yin, and Yin Zhang, "Group sparse optimization by alternating direction method," Tech. Rep., Rice Univ Houston Tx Dept of Computational and Applied Mathematics, 2012.

[18] Junfeng Yang and Yin Zhang, "Alternating direction algorithms for ℓ_1-problems in compressive sensing," *SIAM Journal on Scientific Computing*, vol. 33, no. 1, pp. 250–278, 2011.

[19] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2006, vol. 2, pp. 2169–2178.

[20] Michal Aharon, Michael Elad, and Alfred Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[21] Tiep Huu Vu and Vishal Monga, "Fast low-rank shared dictionary learning for image classification," *IEEE Transactions on Image Processing*, 2017.

[22] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[23] Aleix M Martinez, "The ar face database," *CVC Technical Report*, 1998.

[24] Ron Kohavi et al., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artificial Intelligence*. Stanford, CA, 1995, vol. 14, pp. 1137–1145.