LEARNING STRUCTURED DICTIONARY BASED ON INTER-CLASS SIMILARITY AND REPRESENTATIVE MARGINS

Yuyao Zhang[⋆†] *Philip O. Ogunbona*[⋆] *Wanqing Li*[⋆] *Gordon G. Wallace*[†]

[⋆]Advanced Multimedia Research Lab, ICT Research Institute;
[†]Intelligent Polymer Research Institute
ARC Centre of Excellence for Electromaterials Science;
University of Wollongong

ABSTRACT

We consider the problem of learning a structured and discriminative dictionary based on sparse representation for classification task. The structure comprises class-shared and class-specific partitions which allows the separation of common and class-specific information in the data for classification. The resulting optimization problem was a max margin formulation that exploits the hinge loss function property. Comparative evaluation of the proposed classifier against four recent alternatives in a gender classification task indicates a 3-percenatge point improvement.

Index Terms— Sparse Representation, Max Margin, Dictionary learning

1. INTRODUCTION

Data representation, also referred to as feature extraction, plays an important role in successful machine learning algorithms. When it is properly formulated, such representation captures the explanatory factors underlying data variations and constitute an effective input to learning and prediction algorithms [1]. Sparse representation is one of the many representation-learning methods in which the data samples are encoded by coefficient vectors (or sparse codes) having limited number of non-zero elements. The sparse code captures high-level variations underlying the data [2].

Given a dataset $\{\mathbf{Y}|y_1, y_2, \ldots, y_j, \ldots, y_K\}$, the feature vector or sparse code x_j of y_j (resulting from the mapping $\mathbf{D}|x_j \mapsto y_j$ (the dictionary)) is obtained by solving the convex problem,

$$\arg\min_{x_j} \|y_j - \mathbf{D}x_j\|_2^2 + \lambda\|x_j\|_1, \tag{1}$$

where λ is the regularization parameter and the l_1 norm promotes sparsity. A number of methods have been developed to obtain the dictionary $\mathbf{D}$ in Eq. (1) using samples from $\mathbf{Y}$, and good representation performance have been reported [3, 2].

Sparse representation with over-complete dictionaries have been successfully used in classification tasks. When dictionaries are constructed without supervision [3, 2], generally a supervised predictor is required to perform classification while those constructed with supervision [4, 5, 6] do not necessarily require one. Sparse representation-based classification (SRC) [6] is a well known example requiring supervised-dictionary. In SRC, multiple dictionaries are trained to distinctly represent different classes and the classification output is solely based on the minumum reconstruction error relative to the different dictionaries.

The training process could possibly result in dictionaries with sufficient similarity that, given a test sample y_j of class $c_i \in \mathcal{C} = \{c_1, \ldots, c_M\}$, dictionaries, $\mathbf{D}_{c_k}$ $(c_k \neq c_i)$, trained for other classes could also represent it. This phenomenon has been discussed in [7] and described as collaborative representation. The explantion was that two similar samples from different classes could be well represented by a combination of components in a dictionary except for differing reconstruction errors. Hence, there are collabortaively represented components of the two samples. The collaboratively represented components of y_j, hereinafter referred to as the common or class-shared components, are more representative than discriminative. Classification improvement has been reported by eliminating the class-shared components [8, 5].

Previous dictionary learning algorithms generally do not explicitly define and constrain the separated common components. Such separated common components could include parts shared only by a few classes and are still discriminative in differentiating these classes from others [9]. Kong et al. [8] and Zhou et al. [10] did not apply specific constraints in constructing the common dictionary. In general, it is not clear what should be represented by the common dictionary or otherwise by class-specific dictionaries. To be specific, the criteria for decomposing data samples into two separate components that are represented by common and class-specific dictionaries are not clear. The resultant algorithm largely depends on the initialisation of the dictionary. In this paper, we clarify this matter by defining the common components as the parts of data which is shared by more classes. In an effort similar to ours, Shen et al. [11] proposed a multi-level frame-

work with a clear definition of dictionaries and classifiers at each level. Conceptually, their common dictionary is at the top level and captures the basic representation information of all classes; sample classification is achieved by selecting the appropriate branch of the hierachy. We quickly distinguish our definition and purpose from that given in [11].

In this paper, we aim to separately represent the components shared by multiple classes with a common dictionary $\mathbf{D}_{com}$, and discriminative components with class-specific dictionaries $\mathbf{D}_{c_i}$ ($c_i \in \mathcal{C} = \{c_1, \ldots, c_M\}$); the dictionaries, $\mathbf{D}_{c_i}$, are collectively represented by $\mathbf{D}_C$. The discrimination criteria is based on representation error.

2. MODELLING THE CLASS-SHARED COMPONENTS

The common components of the samples in a dataset can be shared by a few or all the classes. If shared by a few group of classes, they are also discriminative in the classification process [9]. It is expected that the separated class-shared components will be shared globally by as many classes as possible. This may be achieved by enforcing common components to be shared uniformly by most of the samples in the dataset. Essentially, the common dictionary should be a low-rank approximation of the samples in the dataset. We proceed as follows.

Let the components of the data samples in $\mathbf{Y}$ represented by the common dictionary $\mathbf{D}_{com}$ be denoted by $\mathbf{A} = \mathbf{D}_{com}\mathbf{X}^{\mathbf{D}_{com}}$; $\mathbf{X}^{\mathbf{D}_{com}}$ are the correspondng sparse codes. The dissimilarity among the elements of $\mathbf{A}$ can be expressed as

$$\theta_{com} = \|\mathbf{A} - UU^T\mathbf{A}\|_F^2, \tag{2}$$

where $U \in R^{m\times p\ (p<m)}$ is a matrix of the eigenvectors corresponding to the p largest eigenvalues of the covariance matrix $\mathbf{Y}\cdot\mathbf{Y}^T$. By minimizing θ_{com}, we force $\mathbf{A}$ to be a solution of the homogeneous equation,

$$(\mathbf{I} - UU^T)\mathbf{A} = 0, \tag{3}$$

which makes $rank(A) \approx p$ and embeds it around the same subspaces of rank p across different classes. Fig. 1 geometrically illustrates an example with $p = 2$ where the images of the unit sphere S under transformation by matrices $\mathbf{Y}$ and $\mathbf{A}$ are two ellipses marked YS and AS respectively. In this example, $U = [u_1,\ u_2]$ consists of the two left singular vectors of $\mathbf{Y}$ and $\Sigma = [\sigma_1,\ \sigma_2]$, the corresponding singular values. Further, $[u_1u_1^TA,\ u_2u_2^TA]$ are two semi-axes of AS along the directions of U. Minimising θ_{com} essentially encourages $\mathbf{A}$ to be a matrix which only stretches S along U. The resultant $\mathbf{A}$ represents a transformed copy of $U^T\Sigma$ which is the roughly estimated common components using all samples across different classes.

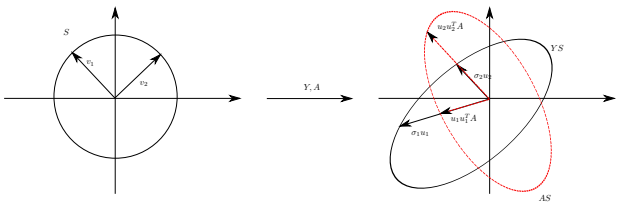


Fig. 1. Toy example on θ_{com} with $p = 2$ in 2D

With the foregoing discussion, the sparse coding problem can be written in terms of the common dictionary, $\mathbf{D}_{com}$, and the combined specific dictionaries $\mathbf{D}_C$ (assumed known):

$$\begin{aligned}\arg\min_{\mathbf{X},\mathbf{D}_{com}} &\|\mathbf{Y} - \mathbf{D}_C\mathbf{X}^{\mathbf{D}_C} - \mathbf{D}_{com}\mathbf{X}^{\mathbf{D}_{com}}\|_F^2 \\ &+ \lambda\sum_{i=1}^{K}\|x_i\|_1 + \beta\theta_{com},\end{aligned} \tag{4}$$

where λ and β are control parameters.

3. REPRESENTATION OF CLASS-SPECIFIC COMPONENTS

The class-specific components are defined based on the discriminative criterion that the common dictionary and a class-specific dictionary, $\mathbf{D}_{(com,c_i)} = [\mathbf{D}_{com}\mathbf{D}_{c_i}]$, can represent samples from class c_i better than other combinations of common and specific dictionaries. Using inner product as a measure of similarity among different vectors (assuming an inner product space), we formulate the discriminative constraints on specific dictionary c_i and sample y_j as the following error measure,

$$\begin{aligned} r_j^{c_i} &= y_j \cdot \mathbf{D}_{(com,c_i)} x_j^{\mathbf{D}_{(com,c_i)}} - \left| \sum_{c_k \neq c_i}^{M,com} y_j \cdot \mathbf{D}_{(c_k)} x_j^{\mathbf{D}_{(c_k)}} \right| \\ &= y_j \cdot \mathbf{D}_{c_i} x_j^{\mathbf{D}_{c_i}} - \left| \sum_{c_k \neq c_i}^{M} y_j \cdot \mathbf{D}_{c_k} x_j^{\mathbf{D}_{c_k}} \right|. \end{aligned} \tag{5}$$

The data sample y_j is assumed to be sample-wisely normalised to 1 using norm-2. Intuitively, in Eq. 5, the first term is unity if all representation is due to the corresponding class-specific dictionary, implying that the second term will be nearly zero. For classification, we have

$$\begin{cases} \text{if } y_j \text{ is in class } c_i, & r_j^{c_i} > 0 \\ \text{else} & r_j^{c_i} < 0 \end{cases}. \tag{6}$$

A geometric interpretation of $r_j^{c_i}$ is shown in Fig. 2 where the

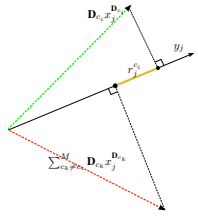


Fig. 2. Visualization of error measure constraint for $r_j^{c_i} > 0$

emboldened yellow segment $r_j^{c_i}$ indicates whether $\mathbf{D}_{c_i}$ represents y_j better than other class-specific dictionaries $\mathbf{D}_{c_k}$ ($c_k \neq c_i$). The sign of $r_j^{c_i}$ is an indicator of whether y_j will be predicted as a sample from class c_i.

Note that this constraint is applied to class-specific dictionaries only, thus leaving the sparse coding problem in learning and testing stage as in the unsupervised problem. A careful inspection of the discriminative constraints on class-specific dictionaries indicates that they are similar to those in max-margin learning where the purpose is to separate data with margins. We thus incorporate the constraints described in Eq. (5) into the dictionary updating stage similarly to the constraints in the least square SVM [12] with a linear kernel. Assume the training set, $\mathbf{Y}_{c_i}$, for class c_i has N_{c_i} samples. Training the class-specific dictionary $\mathbf{D}_{c_i}$ for class c_i can be formulated as,

$$\begin{aligned}\arg\min_{\mathbf{D}_{c_i}} \sum_{j=1}^{N_{c_i}} \{\|y_j - \mathbf{D}_{c_i} x_j^{\mathbf{D}_{c_i}} - \sum_{c_k \neq c_i}^{M,com} \mathbf{D}_{c_k} x_j^{\mathbf{D}_{c_k}}\|_2^2 + \gamma e_j^2\} \\ \text{s.t.} \quad y_j^T \mathbf{D}_{c_i} x_j^{\mathbf{D}_{c_i}} + b_j^{\mathbf{D}_{c_i}} = 1 - e_j, \quad \forall j \in [1, N_{c_i}], \end{aligned} \tag{7}$$

where e_j is a γ-controlled tolerance for violating the constraints and

$$b_j^{\mathbf{D}_{c_i}} = -\left| \sum_{c_k \neq c_i}^{M} y_j^T \mathbf{D}_{c_k} x_j^{\mathbf{D}_{c_k}} \right|.$$

In the objective function (7), the class-specific dictionary $\mathbf{D}_{c_i}$ is forced to have the smallest representation error over $\mathbf{Y}_{c_i}$, compared to all the other class-specific dictionaries. The usage of e_j essentially allows some samples y_j to be represented by the common dictionary, resulting in more compact and discriminative dictionaries.

Improved convergence property is obtained by simultaneously updating $\mathbf{x}_j^{\mathbf{D}_{c_i}}$ and $\mathbf{D}_{c_i}$ as in KSVD [3]. We denote by $w_j^{\mathbf{D}_{c_i}}$ the estimated specific parts captured by $\mathbf{D}_{c_i}$ in y_j and by $W^{\mathbf{D}_{c_i}} = \mathbf{D}_{c_i} \mathbf{X}_{c_i}^{\mathbf{D}_{c_i}} = [w_1^{\mathbf{D}_{c_i}}, \ldots, w_{N_{c_i}}^{\mathbf{D}_{c_i}}]$ the estimated specific components represented by $\mathbf{D}_{c_i}$ in $\mathbf{Y}_{c_i}$. Solving problem (7) for an optimal $\mathbf{D}_{c_i}$ is the same as solving,

$$\begin{aligned}\arg\min_{\mathbf{D}_{c_i}} \|\mathbf{R}^{\mathbf{D}_{c_i}} - W^{\mathbf{D}_{c_i}}\|_F^2 + \gamma \sum_{j=1}^{K} e_j^2 \\ \text{s.t.} \quad y_j^T w_j^{\mathbf{D}_{c_i}} + b_j^{\mathbf{D}_{c_i}} = 1 - e_j, \end{aligned} \tag{8}$$

where $\mathbf{R}^{\mathbf{D}_{c_i}} = \mathbf{Y}_{c_i} - \sum_{c_k \neq c_i}^{M,com} \mathbf{D}_{c_k} \mathbf{X}_{c_i}^{\mathbf{D}_{c_k}}$. Fortunately, the problem (8) is convex with simple and computationally efficient analytical solution. Each specific dictionary is updated against the estimated $W^{\mathbf{D}_{c_i}}$ using KSVD and will be described in details in Section 4.

In problem (8) and analogous to SVM theory, $w_j^{\mathbf{D}_{c_i}}$ is the normal vector of the hyperplane $G(w_j^{\mathbf{D}_{c_i}}, b_j^{\mathbf{D}_{c_i}})$. The constraint encourages the error indicator $r_j^{c_i}$ to be $(1 - e_j)$ with a minimum e_j. Going through all class-specific dictionaries, we essentially make $\mathbf{D}_{c_i}$ better at representing samples from class c_i and poorer at those from other classes.

Compared to the softmax loss function used by [4], our formulation is 1) simple and convex and does not require the local linear approximation of the softmax loss function, 2) more discriminative because we force the reconstruction error of one class-specific dictionary to be smaller than those of the remaining class-specific dictionaries by incorporating common dictionary and thus reduce collaborative representation, 3) linear and provides exact linear penalty even when samples are very close to the margin.

4. SOLVING THE DICTIONARY LEARNING PROBLEM

In solving the proposed optimisation problem (4), the process consists of alternately encoding training samples and updating atoms in the dictionaries. We fix $\mathbf{D}$ and update $\mathbf{X}$, and vice versa. The regularisation term θ_{com} can be written as,

$$\theta_{com} = \|(\mathbf{I} - UU^T)\mathbf{DHX}\|_F^2, \tag{9}$$

where $\mathbf{I}$ is an $m \times m$ identity matrix, $\mathbf{H}$ is a diagonal matrix structured as

$$diag(\mathbf{H}) = \left[\overbrace{1, 1, \ldots, 1}^{N_{com}}, \underbrace{0, 0, \ldots, 0}_{N_1}, \ldots, \overbrace{0, 0, \ldots, 0}^{N_c} \right], \tag{10}$$

which makes $\mathbf{A} = \mathbf{DHX}$. The objective function (4) is then rewritten as

$$\begin{aligned}\arg\min_{\mathbf{X}} \left\| \begin{pmatrix} \mathbf{Y} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbf{D} \\ (\sqrt{\beta}(\mathbf{I} - UU^T)\mathbf{DH} \end{pmatrix} \mathbf{X} \right\|_F^2 \\ + \lambda \sum_{i=1}^{K} \|x_i\|_1, \end{aligned} \tag{11}$$

where $\mathbf{0}$ is a zero matrix with the same size as $\mathbf{Y}$. The objective function (11) is the standard sparse coding problem with various efficient solvers such as the orthogonal matching pursuit (OMP) [13] and feature sign search algorithm [2]. We adopted the feature sign search algorithm.

By fixing $\mathbf{D}_{com}$, we apply the method of Lagrange multipliers to the objective function (8) and define the Lagrangian as [12, 14],

$$\begin{aligned}L(W_{c_i}, \mathbf{E}, \Lambda) = \|\mathbf{R}^{\mathbf{D}_{c_i}} - W^{\mathbf{D}_{c_i}}\|_F^2 + \gamma \|\mathbf{E}\|_F^2 \\ -trace\{\Lambda[S_{c_i}(\mathbf{Y}^T W^{\mathbf{D}_{c_i}} + \mathbf{B}^{\mathbf{D}_{c_i}}) - \mathbf{I} + \mathbf{E}^{\mathbf{D}_{c_i}}]\}, \end{aligned} \tag{12}$$

where λ_j is the real-valued Lagrangian multiplier for each training sample, $\Lambda = diag([\lambda_1, \ldots, \lambda_K])$, $\mathbf{I} \in R^{K \times K}$ is an identity matrix and

$$\begin{aligned}S_{c_i} = diag([s_1^{c_i}, \ldots, s_K^{c_i}]),\ \mathbf{B}^{\mathbf{D}_{c_i}} = diag([b_1^{\mathbf{D}_{c_i}}, \ldots, b_K^{\mathbf{D}_{c_i}}]), \\ \mathbf{E}^{\mathbf{D}_{c_i}} = diag([e_1^{\mathbf{D}_{c_i}}, \ldots, e_K^{\mathbf{D}_{c_i}}]).\end{aligned}$$

Applying the optimal conditions and properties of trace derivatives, we can obtain a set of linear equations [12, 14],

$$\begin{cases} \frac{\partial L}{\partial W^{\mathbf{D}_{c_i}}} = 0 \longrightarrow & 2W^{\mathbf{D}_{c_i}} + \mathbf{0} - \mathbf{Y}\Lambda S_{c_i} = 2\mathbf{R}^{\mathbf{D}_{c_i}} \\ \frac{\partial L}{\partial \mathbf{E}} = 0 \longrightarrow & \mathbf{0} + 2\gamma \mathbf{E}^{\mathbf{D}_{c_i}} - \Lambda = \mathbf{0} \\ \frac{\partial L}{\partial \Lambda} = 0 \longrightarrow & S_{c_i} \mathbf{Y}^T W^{\mathbf{D}_{c_i}} + \mathbf{E}^{\mathbf{D}_{c_i}} - \mathbf{0} = \mathbf{I} - \mathbf{B}^{\mathbf{D}_{c_i}} \end{cases} \tag{13}$$

The linear system (13) can be solved by firstly solving for Λ and then, $W^{\mathbf{D}_{c_i}}$ and $\mathbf{E}^{\mathbf{D}_{c_i}}$ can be solved. Note that since Λ is diagonal in our formulation, formal matrix inversion is obviated and $W^{\mathbf{D}_{c_i}}$ is easily obtained. The class-specific dictionary $\mathbf{D}_{c_i}$ is then updated to minimise $W^{\mathbf{D}_{c_i}}$ using the standard KSVD algorithm [3]. This process is repeated for all M class-specific dictionaries.

When given the sparse codes $\mathbf{X}$ for $\mathbf{Y}$ and all $\mathbf{D}_C$ are fixed, following the procedures from Eq. (9) to Eq. (11), we could formulate the common dictionary update problem as,

$$\arg\min_{\mathbf{D}_{com}} \left\| \begin{pmatrix} \mathbf{Y}^{\mathbf{D}_{com}} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbf{D}_{com} \\ (\sqrt{\beta}(\mathbf{I} - UU^T)\mathbf{D}_{com} \end{pmatrix} \mathbf{X}^{\mathbf{D}_{com}} \right\|_F^2 \tag{14}$$

The objective function (14) can be solved efficiently using KSVD mechanism to update atoms sequentially.

5. EXPERIMENTAL DESIGN AND RESULTS

The proposed algorithm has been validated on the AR Face dataset [15] in a gender classification task and comparative evaluation performed relative to other four recent learning algorithms. Experimental results are either taken as reported in the respective papers or regenerated.

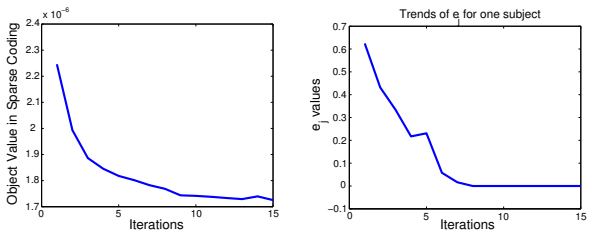
Images are classified as in [16] both globally and locally according to the representation errors. Global coding is achieved by concatenating all common and class-cspecific dictionaries as the mapping $\mathbf{D}$ while local coding encodes each sample with $\mathbf{D}_{(com,c_i)}$. Reconstruction errors are calculated for each class-specific dictionary for classification.

Following the selection criterion in [16], we selected a subset of non-occluded face images from the AR face dataset [15]. The subset contains 50 males and 50 females with each person having 14 images. For a fair comparison with results reported in [16], we use the same set of training and testing samples. The dictionary is trained using the first 25 males and 25 females with the remaining images used as test set. The size (number of atoms in the dictionary) of each class-specific dictionary was varied from 250 to 25; the common dictionary was fixed at 50. The classification results for both cases are shown in columns 2 and 3 of Table 1. In general, the classification results of our method in terms of both the local and global coding mechanisms outperform all the other methods. Our results are over 3% points better than those in LDL [16]. The local coding generally works better than global coding. In Table. 1 the worst accuracy of our algorithm with smaller class-specific dictionaries and global coding is the same as LDL-LC using a 10 times larger class-specific dictionary and local coding. This can be attributed to our use of a common dictionary which encodes the general information for representation. A small class-specific dictionary is still sufficient to represent the discriminative components among classes. However, other discriminative dictionary learning methods (e.g. COPAR) do not clearly define the shared and class-specific components and this makes the common dictionary powerful enough for representation when the class-specific dictionary becomes small. Pairwise t-tests between our algorithm and others (see Table 1) at $\alpha = 0.05$ was each found to be statistically significant indicating that our algorithm outperform them. Overall $p < 0.021$.

Table 1. Gender classification accucary on AR Face Dataset for various dictionary sizes

Algorithm	250	100	50	25
DLSI [5]	94.0	97.0	95.4	93.7
COPAR [8]	93.4	95.3	94.1	93.0
FDDL-	94.3	96.1	93.7	93.7
LC(GC) [17]	(94.3)	(92.9)	(94.4)	(92.1)
LDL-	95.3	93.3	93.0	95.0
LC(GC) [16]	(94.8)	(93.0)	(92.3)	(92.4)
Ours-LC(GC)	**98.6**	**99.0**	**98.6**	**98.3**
	(97.9)	**(98.3)**	**(97.0)**	**(96.0)**

We further analyse the convergence of our algorithm by investigating the value of the objective as well as the value of the error tolerance $e_j^{\mathbf{D}_{c_i}}$. Fig. 3 (a) illustrates the fast convergence of the objective function in sparse coding stage. The initial values of e_j are around 0.7 because of the initialisation using KSVD and quickly goes down to around zero as shown in Fig. 3 (b). As shown in Fig. 3 (b), the values of e_j during the optimisation process dropped continuously towards zero, clearly indicating the gain from the property of hinge loss function.



(a) (b)

Fig. 3. (a) Object values in the sparse coding stage. (b) The values of e_j for one subject over the training process.

6. CONCLUSION

A discriminative structured dictionary learning algorithm based on representation error discrimination rule was proposed and its performance verified. Class-specific dictionary representation error was formulated as a criterion that enforces class-specific representation while making common dictionary contribute to general sample representation. The SVM-like formulation suggests possible extension to a nonlinear high-dimensional case by exploiting the kernel trick.

7. REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug 2013.

[2] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng, "Efficient sparse coding algorithms," in *Proc. Neural Information and Processing Systems*. 2007, pp. 801–808, NIPS.

[3] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[4] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[5] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2010, pp. 3501–3508.

[6] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb 2009.

[7] D. Zhang, Meng Yang, and Xiangchu Feng, "Sparse representation or collaborative representation: Which helps face recognition?," in *Proc. IEEE Intl Conf. Computer Vision*, Nov 2011, pp. 471–478.

[8] Shu Kong and Donghui Wang, "A dictionary learning approach for classification: Separating the particularity and the commonality," in *Proc. European Conf. Computer Vision*, vol. 7572 of *Lecture Notes in Computer Science*, pp. 186–199. 2012.

[9] A. Torralba, K.P. Murphy, and W.T. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 854–869, May 2007.

[10] Ning Zhou, Yi Shen, Jinye Peng, and Jianping Fan, "Learning inter-related visual dictionary for object recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2012, pp. 3490–3497.

[11] Li Shen, Shuhui Wang, Gang Sun, Shuqiang Jiang, and Qingming Huang, "Multi-level discriminative dictionary learning towards hierarchical visual categorization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2013, pp. 383–390.

[12] J.A.K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[13] J.A. Tropp and A.C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Information Theory*, vol. 53, no. 12, pp. 4655–4666, Dec 2007.

[14] R. Fletcher, *Practical Methods of Optimization*, Wiley-Interscience, New York, NY, USA, 1987.

[15] A.M. Martinez and R. Benavente, "The AR face database," Tech. Rep., 1998.

[16] Meng Yang, Dengxin Dai, Linlin Shen, and L. Van Gool, "Latent dictionary learning for sparse representation based classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2014, pp. 4138–4145.

[17] Meng Yang, D. Zhang, Xiangchu Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Nov 2011, pp. 543–550.