Structured Sparse Coding for Classification via Reweighted $\ell_{2,1}$ Minimization

Yong Xu[1], Yuping Sun[1,2], Yuhui Quan[2], and Yu Luo[1,2]

[1] South China University of Technology, Guangzhou 510006, China,
yxu@scut.edu.cn sun.yp@mail.scut.edu.cn
[2] National University of Singapore, Singapore 119076, Singapore
matquan@nus.edu.sg matluoy@nus.edu.sg

Abstract. In recent years, sparse coding has been used in a wide range of applications including classification and recognition. Different from many other applications, the sparsity pattern of features in many classification tasks are structured and constrained in some feasible domain. In this paper, we proposed a reweighted $\ell_{2,1}$ norm based structured sparse coding method to exploit such structures in the context of classification and recognition. In the proposed method, the dictionary is learned by imposing the class-specific structured sparsity on the sparse codes associated with each category, which can bring noticeable improvement on the discriminability of sparse codes. An alternating iterative algorithm is presented for the proposed sparse coding scheme. We evaluated our method by applying it to several image classification tasks. The experiments showed the improvement of the proposed structured sparse coding method over several existing discriminative sparse coding methods on tested data sets.

Keywords: sparse coding, reweighted $\ell_{2,1}$ minimization, image classification

1 Introduction

In recent years, sparse model has been an important tool with a wide range of applications. Sparse modeling assumes that signals of interest can be succinctly expressed under some suitable system in a linear manner. The elements used for expressing signals are often referred as *atoms* and the collection of all such atoms is called a *dictionary* for sparse modeling. The computational method for sparse modeling is called *sparse coding*, which aims at finding a dictionary, as well as the sparse coefficients, from input signals. This sparse scheme, which rigorously pursues the sparsity of the codes, works quite well in image processing and restoration. However, it's not enough to achieve high discriminability for classification and recognition tasks without exploiting extra structural information existing in signals. Given a data matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K]$ where $\boldsymbol{X}_k$ denotes the data from the K category, the optimal structure of the corresponding sparse coefficient matrix $\boldsymbol{C}$ under an ideal semantic dictionary $\boldsymbol{D}$ for classification is as follows:

$$\boldsymbol{C}^* \triangleq \begin{bmatrix} \boldsymbol{C}_{[1]} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}_{[2]} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{C}_{[K]} \end{bmatrix} . \quad (1)$$

While some recent approaches have attempted to pursue structured sparsity of the form (1) either explicitly or implicitly, the disadvantages of these approaches are obvious. For example, the simultaneous low-rank and sparse constrains used in [35] are implemented with nuclear norm and ℓ_1-norm, which would yield to bias solution as ℓ_1-norm has heavier penalty on larger coefficients. Supervised sparse coding [34] do not explicitly impose structures on sparse codes, which often results in sub-optimal solutions. The explicit structured regularizations on sparse codes, including label consistency [11] and group Lasso [32], require the form of structure to be predefined, which is inflexible and is inaccurate when the size of dictionary is limited. This inspires us to develop an effective structured sparse coding method for classification. Motivated by the effectiveness of the reweighting scheme in compressed sensing [5], we develop a reweighted $\ell_{2,1}$ norm based method for structured sparse coding based classification, which is able to automatically discover the underlying structures of training data and obtain the sparsity patterns of the form (1).

In this paper, a reweighted $\ell_{2,1}$ minimization model is constructed to exploit the class-specific joint structured sparsity patterns existing in labeled data. The weights are determined by the magnitude of sparse codes, which in turn forces the training samples to select active atoms that can span the subspace of the corresponding class and thus encourages the sparse codes to be of the form (1). An alternating iterative algorithm is developed to solve the proposed model. Experimental results on face recognition, gender classification, and scene classification have demonstrated the excellent performance of our method in comparison with several existing representative dictionary learning methods.

The proposed structured sparse coding approach enjoys several advantages. Firstly, using reweighted $\ell_{2,1}$ regularization in the proposed method is able to reduce the bias on large coefficients, while the $\ell_{2,1}$ regularization based methods [21] cannot omit such a bias when dealing with classification tasks. Besides, the reweighting scheme updates the weights according to the magnitude of the sparse codes, which is more flexible compared to the standard weighting strategy proposed in [17, 25]. Secondly, compared to the discriminative sparse coding methods [29, 34] for classification, the proposed method is able to learn dictionaries by which distinct structured sparsity patterns can be enforced on the sparse codes of samples from different classes. Finally, our method could detect the subspace of data from each class spanned by atoms of the dictionary which helps to enhance the performance of classification.

2 Related Work

Group sparsity is a widely-used structured sparsity which assumes that atoms are selected by input signals in a group-wise manner instead of a singleton-wise one, see e.g. [12, 13, 9, 28]. In the group sparsity setting, the coefficients are arranged into a predetermined set of groups, and the sparsity term penalizes the number of active groups. In the past years, various types of group sparsity have been exploited, e.g., overlapping groups [9], tree sparsity [13], and graph sparsity [2]. The group sparse coding has been used in several classification tasks, e.g., real time object recognition system [24], face recognition [10], and image classification [19, 33]. Instead of considering correlation between dictionary atoms, many approaches proposed to seek for the collaborative structured sparsity by encoding the shared high-order information among related samples [21, 4]. For example, in [21], a $\ell_{2,1}$-norm regularization is performed to select features across all related samples with collaborative structured sparsity, i.e., each feature either has small or large values for all data points at the same time. Recently there is a growing interest in exploiting block structured sparsity [7, 6], i.e., sparse groups of features for all related samples are jointly encoded. Elhamifar et.al. [7] explicitly impose block structure on sparse codes for classification. Zhang et.al. [35] implicitly impose block structure on sparse codes by using simultaneous low-rank and sparse constraint.

In the classification case, a natural and simple way is to use sparse coefficients as features to train a classifier, see e.g. [29, 15]. However the obtained sparse code does not have enough discriminative power for classification. Thus many researches proposed to add additional discriminative constrains on the sparse codes during the sparse coding process, e.g. the class separation criterion (e.g. Fisher discrimination criterion [31, 30]), prediction loss (e.g.logistic loss [18] and linear predictive errors [34, 11]). Some approaches [22, 31, 36] partition a dictionary into multiple subdictionaries by associating each atom with certain class labels, and then impose discrimination to sparse codes of each subdictionary. Compared with our method, these approaches need to predefine the block structure of the dictionary.

3 Our Method

In this section, an effective dictionary learning model for structured sparse coding which induces structural sparsity on sparse codes with reweighted $\ell_{2,1}$-norm is proposed. Also an efficient alternating iterative algorithm is developed to solve the proposed model.

3.1 Problem Formulation

Let $\boldsymbol{Y} = [\boldsymbol{Y}_{[1]}, \boldsymbol{Y}_{[2]}, \ldots, \boldsymbol{Y}_{[K]}]$ denote a set of training samples from K categories, where $\boldsymbol{Y}_{[k]}$ denotes the training samples from k-th category. One natural way to obtain class-specific structured sparsity patterns for samples from each category, is to construct a structured sparse coding model as follows,

$$\underset{\boldsymbol{D}\in\mathcal{X},\boldsymbol{C}}{\operatorname{argmin}} \sum_{k=1}^{K} \frac{1}{2}\|\boldsymbol{Y}_{[k]} - \boldsymbol{D}\boldsymbol{C}_{[k]}\|_F^2 + \lambda\|\boldsymbol{C}_{[k]}\|_{2,0}, \tag{2}$$

where

$$\mathcal{X} = \{\boldsymbol{D} \in \mathbb{R}^{n\times m} : \|\boldsymbol{d}_j\|_2 = 1, 1 \le j \le m\}$$

denotes the feasible set of dictionary $\boldsymbol{D}$, which ensures that the atoms are appropriately normalized. And $\boldsymbol{C}_{[k]}$ is a sub-matrix of $\boldsymbol{C}$ collecting the sparse codes of signals from the k-th category (i.e.the sparse coefficients corresponding to $\boldsymbol{Y}_{[k]}$).

However, solving the $\ell_{2,0}$ norm related problem is a NP-hard problem. Thus we relax the model (2) to a weighted $\ell_{2,1}$ norm based structural sparse coding method as follows:

$$\underset{\boldsymbol{D}\in\mathcal{X},\boldsymbol{C}}{\operatorname{argmin}} \sum_{k=1}^{K} \frac{1}{2}\|\boldsymbol{Y}_{[k]} - \boldsymbol{D}\boldsymbol{C}_{[k]}\|_F^2 + \lambda\|\boldsymbol{C}_{[k]}\|_{\boldsymbol{w}_k;2,1}, \tag{3}$$

where $\boldsymbol{w}_k$ is a weight vector for k-th category, λ is a scalar controlling the weight of the structured sparsity prior, and $\|\cdot\|_{\boldsymbol{w};2,1}$ is the weighted $\ell_{2,1}$-norm defined as $\|\boldsymbol{X}\|_{\boldsymbol{w};2,1} = \sum_{i=1} w_i \|\boldsymbol{x}^i\|_2$. When all weights are set to have equal magnitude, the minimization (3) turns to be a standard $\ell_{2,1}$ minimization.

The choice of the weight matrix $\boldsymbol{W}$ is essential for classification, as the magnitude of each weight $W_{i,k}$ is able to decide how heavy penalty will be imposed on i-th row of $\boldsymbol{C}_{[k]}$ and thus has great influence on the quality of the resulting structural sparsity pattern. When the optimal dictionary is given, if most samples from the k-th category have significant responses to the i-th atom, the corresponding weight $W_{i,k}$ should be small, vice versa. However the optimal dictionary is unknown, thus a reweighted $\ell_{2,1}$ minimization is presented as follows,

$$\operatorname*{argmin}_{\boldsymbol{D}\in\mathcal{X},\boldsymbol{C},\boldsymbol{W}} \sum_{k=1}^{K} \frac{1}{2}\|\boldsymbol{Y}_{[k]} - \boldsymbol{D}\boldsymbol{C}_{[k]}\|_F^2 + \lambda\|\boldsymbol{C}_{[k]}\|_{\boldsymbol{w}_k;2,1}, \tag{4}$$

where the reweighting scheme is employed to redefine the weights in each iteration of the learning process, as described in Sec. 3.2. The minimization (4) can be viewed as a relaxation of the $\ell_{2,0}$ minimization problem.

The reweighting scheme employed in our method provides a tool to explore the relationships between class-specific structured sparsity patterns and the weighed values, which is able to provide high discriminative information. To be more specific, when $\|\boldsymbol{c}^i_{[k]}\|_2$ is small, it implies that data from k-th category are not likely to fall into the subspace spanned by i-th atom. Setting $W_{i,k}$ large emphasizes the penalty on the corresponding sparse coefficients, which moves the i-th atom away from the favorite list of the samples from k-th category. On the other hand, when $\|\boldsymbol{c}^i_{[k]}\|_2$ is large, it implies that the data from the k-th category are likely to lie in the subspace spanned by i-th atom. Setting $W_{i,k}$ small would provide flexibility for the corresponding sparse coefficients, which improves the adaptivity of i-th atom to the underlying structures of data from the k-th category.

3.2 The Proposed Iterative Algorithm

It is nontrivial to solve the minimization (3). In this section an alternating iterative algorithm is proposed to separate the minimization into several simpler ones. The iteration stops until either of the following stopping criteria is satisfied: *(a)* the change of objective function is small enough; *(b)* the maximum iteration number has been reached. The learned dictionary and the obtained weight matrix are then used to code the test samples and the label prediction for each test sample is based on its corresponding class-specific representation residuals.

Sparse Approximation. Given dictionary $\boldsymbol{D}^{(t)}$ and the weight matrix $\boldsymbol{W}^{(t)}$, the sparse coefficients $\boldsymbol{C} = [\boldsymbol{C}_{[1]}, \ldots, \boldsymbol{C}_{[K]}]$are calculated as follows:

$$\boldsymbol{C}^{(t+1)} = \operatorname*{argmin}_{\boldsymbol{C}} \sum_{k=1}^{K} \frac{1}{2}\|\boldsymbol{Y}_{[k]} - \boldsymbol{D}\boldsymbol{C}_{[k]}\|_F^2 + \sum_{k=1}^{K} \lambda\|\boldsymbol{C}_{[k]}\|_{\boldsymbol{w}_k^{(t)};2,1}, \tag{5}$$

which is separable and can be decomposed into K independent subproblems:

$$\boldsymbol{C}_{[k]}^{(t+1)} = \underset{\boldsymbol{C}}{\operatorname{argmin}} \|\boldsymbol{Y}_{[k]} - \boldsymbol{D}\boldsymbol{C}\|_F^2 + \lambda \|\boldsymbol{C}\|_{\boldsymbol{w}_k^{(t)};2,1}, \quad \forall k. \tag{6}$$

The reweighted $\ell_{2,1}$ minimization (6) is solved via the accelerated proximal gradient algorithm [23].

Dictionary Update. Given the sparse codes $\boldsymbol{C}^{(t+1)}$, the dictionary $\boldsymbol{D}$ is updated as follows:

$$\boldsymbol{D}^{(t+1)} = \underset{\boldsymbol{D}\in\mathcal{X}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{C}^{(t+1)}\|_F^2, \tag{7}$$

where $\boldsymbol{D}^{(t+1)} = [\boldsymbol{d}_1^{(t+1)}, \cdots, \boldsymbol{d}_m^{(t+1)}]$ is updated atom by atom via the projected gradient descent [16]; see [3] for the details.

Weight Refinement. Given the sparse codes $\boldsymbol{C}^{(t+1)}$, based on the discussion in Sec. 3.1, we can calculate the weights as follows,

$$W_{i,k} = \frac{1}{\|\boldsymbol{c}_{[k]}^{i,(t+1)}\|_2 + \epsilon}, \quad \forall i, k, \tag{8}$$

followed by a ℓ_1 normalization on each weight vector corresponding to each category:

$$\boldsymbol{w}_k = \frac{\mu \boldsymbol{w}_k}{\|\boldsymbol{w}_k\|_1}, \quad \forall k, \tag{9}$$

where μ is a constant implemented according to the dictionary size and ϵ is a sufficiently small positive parameter for stability.

3.3 Classification Process

Once the learning process is finished, for each category, we can construct a subset of atoms from the learned dictionary $\boldsymbol{D}$ by measuring the joint sparse representations of related samples. Being associated with class-specific structured sparsity patterns, these subsets of atoms can be used for classifying test samples. For k-th category, the corresponding subset of atoms is defined as $\boldsymbol{D}_{[k]} = \{\boldsymbol{d}_i \mid \|\boldsymbol{c}_{[k]}^i\|_2 > 0, 1 \leq i \leq m\}$. Then for each k, the sparse codes $\boldsymbol{c}_{[k]}$ of a test signal $\boldsymbol{y}$, is obtained by solving the following minimization:

$$\boldsymbol{c}_{[k]} = \underset{\boldsymbol{c}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{D}_{[k]}\boldsymbol{c}\|_2^2 + \alpha\|\boldsymbol{c}\|_1, \tag{10}$$

where α is a parameter that balances the trade-off between sparsity and fidelity of the solution. The problem (10) is also solved by the accelerated proximal gradient algorithm [23]. Then $\boldsymbol{y}$ is classified to be the class with the minimum prediction error:

$$\text{identity}(\boldsymbol{y}) = \underset{k}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{D}_{[k]}\boldsymbol{c}_{[k]}\|_2^2 + \alpha\|\boldsymbol{c}_{[k]}\|_1. \tag{11}$$

4 Experiment

The performance of the proposed method is demonstrated with several classification tasks, including face recognition, gender classification and scene classification. We compared our method with several state-of-the-art dictionary learning approaches, including Discriminative K-SVD (D-KSVD) [34], Label Consistent K-SVD (LC-KSVD) [11], Sparse Representation based Classifier (SRC) [27], Dictionary Learning with Structure Incoherence (DLSI) [22], dictionary learning with COmmonality and PARticularity (COPAR) [14], Joint Dictionary Learning (JDL) [36], Fisher Discrimination Dictionary Learning (FDDL) [31], Latent Dictionary Learning (LDL) [30]. Only the results available in the literature are reported.

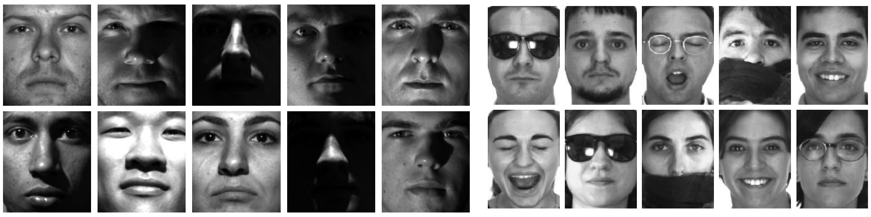
To verify the effectiveness of the proposed reweighting scheme, a baseline method (denoted by Baseline) is implemented for comparison, which is based on the standard $\ell_{2,1}$ minimization. The training stage of the baseline method runs similarly to that of the proposed approach except all the weight are set to be the same constant. The classification stage of the baseline method is the same as which described in Section 3.3. The parameters of the baseline method are set to be the same as ours.

4.1 Implementation Details

Parameter setting. There are five parameters in our approach, i.e., the dictionary size m, the regularization parameters λ and α, the reweighting parameters μ and ϵ. In all the experiments, if no specific instructions mentioned, a five-fold cross validation scheme is used to find λ and α. To have a fair comparison, if no specific instructions mentioned, the dictionary sizes of all the compared methods mentioned above are set to be the same as [11, 30]. Besides, the parameter μ is set equal to m for simplicity and ϵ is set to be a small positive real number (for example 10^{-6}).
Initialization. The initial dictionary $\boldsymbol{D}^{(0)}$ is generated by sampling from training data. More precisely, we randomly select a certain number of samples from each category as the dictionary atoms. For the initialization of the weight matrix $\boldsymbol{W}$, we set $W_{i,k}$ equal to 0.5 if the i-th initial atom is taken from the k-th category and 1 otherwise.
Computational time. To demonstrate the scalability of our method, the proposed method is tested on two datasets of different sizes, the average training time and test time for Ext.YaleB of 504 dimension is 4s and 43s respectively, and for Scene-15 of dimension 3000 is 126s and 111s.

4.2 Face Recognition

We demonstrate the effectiveness of our method in face recognition using the Extended YaleB dataset [8], which contains $2,414$ images of 38 human frontal faces under 64 illumination conditions and expressions. The original images were cropped to 192×168 pixels. As done in [34], each face image is projected into a 504-dimensional feature vector using a random matrix of zero-mean normal distribution. The dataset is randomly split into two halves. One half is used for training and the other half is used for testing. See Figure 1 for some examples.



Samples of Ext. YaleB dataset

Samples of AR face dataset

Fig. 1: Some sample images from dataset YaleB (left) and AR face (right)

The parameters used in this experiment are set as follows: $m = 570$, $\lambda = 0.01$ and $\alpha = 0.002$. We repeated the training and testing processes 10 times with different random splits of the training and testing samples to calculate the average recognition accuracies. The experimental results of all competing methods are summarized in Table 1. Note that the dictionary size of SRC is the same as the number of training samples, so we compare SRC with the same dictionary size as ours(denoted SRC*). It can be seen that our approach is competitive among all the compared methods. The proposed outperformed the baseline method and many state-of-the-art approaches except SRC. But note that the performance of the SRC method degrades dramatically when using dictionaries of the same size as ours.

Table 1: Recognition accuracies (%) of the compared methods on the Ext. YaleB dataset.

KSVD [1]	SRC [27]	D-KSVD [34]	LC-KSVD [11]	LLC [26]	SRC*	Baseline	Our method
93.10	97.20	94.10	95.00	90.70	80.50	84.63	94.52

4.3 Gender Classification

We conducted gender classification on the AR face database [20] with the same experimental setting as [31]. We first chose a non-occluded subset (14 samples per subject) from the AR face database, which consists of 50 males and 50 females, to conduct the experiments. Some sample images are shown in Figure 1. Images of the first 25 males and 25 females were used for testing. Each image is reduced to a 300-dimensional feature vector by PCA.

As there are only two classes and each class has enough training samples, we set the dictionary size relatively small ($m = 50$). The parameters λ and α are set to be 0.06 and 0.003 respectively. As shown in table 2 that our approach outperformed all the tested methods excepted LDL [30]. Although the recognition accuracy of LDL is slightly better than ours, our approach use a dictionary with smaller size than LDL.

Moreover, our method does not involve any discrimination term explicitly and thus has lower computational complexity than LDL.

Table 2: Classification accuracies (%) of the compared methods on the AR database.

DLSI [22]	COPAR [14]	JDL [36]	FDDL [31]	LDL [30]	Baseline	Ours
93.70	93.00	91.00	93.70	95.00	93.13	94.00

4.4 Scene Classification

Our method was applied to scene classification and evaluated on the Scene-15 dataset [15]. The Scene-15 dataset contains 15 scene categories, the number of images per category varies from 210 to 410, and the resolution of each image is about 250×300. See Figure 2 for the sample images per category from the dataset.

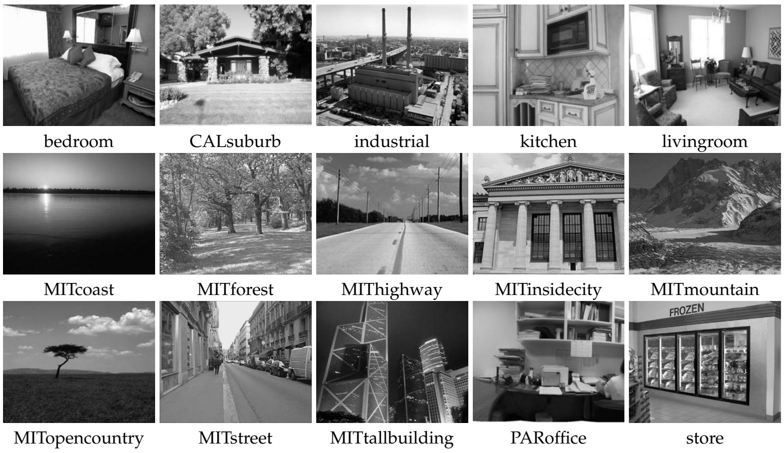


Fig. 2: Example images from the Scene-15 dataset.

The 3000-dimensional SIFT-based spatial pyramid features [15] extracted from the images are used as the input of all the compared methods. Same as the standard experimental protocol used in [15], for each category 100 images are randomly picked up for training and the rest for testing. The parameters m, λ and α are set to be 450, 0.05 and 0.003 respectively. Considering the randomness in the training and testing processes, we run all the experiments 10 times and report the average prediction accuracies. As shown in Table 3, our approach outperformed all the tested methods except FDDL [31], however the training and test time of FDDL is 20 times slower than ours.

Table 3: Classification accuracies (%) on the Scene-15 dataset.

LLC [26]	SRC [27]	KSVD [1]	D-KSVD [34]	LC-KSVD [11]	FDDL [31]	Baseline	Ours
89.20	91.80	86.70	89.10	92.90	98.35	96.62	97.94

5 Summary

A novel dictionary learning approach for structured sparse coding is presented in this paper. Different from existing supervised dictionary learning methods, we proposed a reweighted $\ell_{2,1}$ minimization algorithm to exploit class-specific structured sparsity patterns for signals from each category, which brings benefits to dealing with multi-class classification problem. In the learning process, the dictionary is well adapted to the subspace of training data with a reweighting scheme, which strengthens its discriminability. An efficient alternating iterative scheme is developed to solve the proposed model. We applied our method to several classification tasks. The experimental results demonstrated the competitive performance of our method in comparison with some latest dictionary learning methods. In future, we would like to develop an effective algorithm to solve $\ell_{2,0}$ minimization problem together with the convergence analysis.

References

1. M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.*, 54(11):4311–4322, 2006.
2. F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Structured sparsity through convex optimization. *STAT SCI*, 27(4):450–468, 2012.
3. C. Bao, H. Ji, Y. Quan, and Z. Shen. L0 norm based dictionary learning by proximal methods with global convergence. In *CVPR*, pages 3858–3865. IEEE, 2014.
4. X. Cai, F. Nie, and H. Huang. Exact top-k feature selection via $\ell_{2,0}$-norm constraint. In *IJCAI*, pages 1240–1246. AAAI Press, 2013.
5. E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *J Fourier Anal. Appl.*, 14(5-6):877–905, 2008.
6. E. Elhamifar and R. Vidal. Robust classification using structured sparse representation. In *CVPR*, pages 1873–1879. IEEE, 2011.
7. E. Elhamifar and R. Vidal. Block-sparse recovery via convex optimization. *IEEE Trans. Signal Process.*, 60(8):4094–4107, 2012.
8. A. S. Georghiades, P. N. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):643–660, 2001.
9. L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *ICML*, pages 433–440. ACM, 2009.
10. R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. *arXiv preprint arXiv:0909.1440*, 2009.
11. Z. Jiang, Z. Lin, and L. Davis. Label consistent K-SVD: Learning a discriminative dictionary for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2651–2664, 2013.

12. K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun. Learning invariant features through topographic filter maps. In *CVPR*, pages 1605–1612. IEEE, 2009.
13. S. Kim and E. P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML*, pages 543–550, 2010.
14. S. Kong and D. Wang. A dictionary learning approach for classification: separating the particularity and the commonality. In *ECCV*, pages 186–199. Springer, 2012.
15. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178. IEEE, 2006.
16. C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.
17. C.-Y. Lu, H. Min, J. Gui, L. Zhu, and Y.-K. Lei. Face recognition via weighted sparse representation. *Journal of Visual Communication and Image Representation*, 24(2):111–116, 2013.
18. J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):791–804, 2012.
19. A. Majumdar and R. K. Ward. Classification via group sparsity promoting regularization. In *ICASSP*, pages 861–864. IEEE, 2009.
20. A. M. Martinez. The AR face database. *CVC Technical Report*, 24, 1998.
21. F. Nie, H. Huang, X. Cai, and C. H. Ding. Efficient and robust feature selection via joint l2,1-norms minimization. In *NIPS*, pages 1813–1821, 2010.
22. I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *CVPR*, pages 3501–3508. IEEE, 2010.
23. Z. Shen, K.-C. Toh, and S. Yun. An accelerated proximal gradient algorithm for frame-based image restoration via the balanced approach. *SIAM J Imaging Sci*, 4(2):573–596, 2011.
24. A. Szlam, K. Gregor, and Y. LeCun. Fast approximations to structured sparse coding and applications to object classification. In *ECCV*, pages 200–213. Springer, 2012.
25. X. Tang, G. Feng, and J. Cai. Weighted group sparse representation for undersampled face recognition. *Neurocomputing*, 145:402–415, 2014.
26. J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367. IEEE, 2010.
27. J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, 2009.
28. Y. Xu, Y. Sun, Y. Quan, and B. Zheng. Discriminative structured dictionary learning with hierarchical group sparsity. *Comput. Vis. Image Underst.*, 136:59–68, 2015.
29. J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, pages 1794–1801. IEEE, 2009.
30. M. Yang, D. Dai, L. Shen, and L. V. Gool. Latent dictionary learning for sparse representation based classification. In *CVPR*, pages 4138–4145. IEEE, 2014.
31. M. Yang, D. Zhang, and X. Feng. Fisher discrimination dictionary learning for sparse representation. In *ICCV*, pages 543–550. IEEE, 2011.
32. M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B Stat Methodol*, 68(1):49–67, 2006.
33. L. Zelnik-Manor, K. Rosenblum, and Y. C. Eldar. Dictionary optimization for block-sparse representations. *IEEE Trans. Signal Process.*, 60(5):2386–2395, 2012.
34. Q. Zhang and B. Li. Discriminative K-SVD for dictionary learning in face recognition. In *CVPR*, pages 2691–2698. IEEE, 2010.
35. Y. Zhang, Z. Jiang, and L. S. Davis. Learning structured low-rank representations for image classification. In *CVPR*, pages 676–683. IEEE, 2013.
36. N. Zhou, Y. Shen, J. Peng, and J. Fan. Learning inter-related visual dictionary for object recognition. In *CVPR*, pages 3490–3497. IEEE, 2012.