HIERARCHICAL STRUCTURED DICTIONARY LEARNING FOR IMAGE CATEGORIZATION

Tzu-Chan Chuang, Chen-Kuo Chiang, Shang-Hong Lai

National Tsing Hua University, Taiwan, alice20041025@gmail.com
National Chung Cheng University, Taiwan, ckchiang@cs.ccu.edu.tw
National Tsing Hua University, Taiwan, lai@cs.nthu.edu.tw

ABSTRACT

A novel Hierarchical Structured Dictionary Learning (HSDL) algorithm is proposed in this paper. It aims to learn class-specific dictionaries for all classes simultaneously in a hierarchical structure. A discriminative term based on Fisher discrimination criterion is jointly considered for both the class-specific dictionaries in the lower level and the shared dictionaries in the upper level to enhance the discrimination of dictionaries. The experimental results evaluated on the ImageNet database have shown the superior performance of HSDL over the state-of-the-art dictionary learning methods.

Index Terms— Sparse coding, hierarchical, image classification.

1. INTRODUCTION

The rapidly developed sparse coding techniques have led to promising results in image classification problems [16] [5][1][10][11][13][14][2]. Current dictionary learning methods can be roughly categorized into two categories: unsupervised dictionary and supervised dictionary learning. One well-known unsupervised dictionary learning method, K-SVD [6], learns the over-complete dictionary from a training dataset of natural image patches. In supervised dictionary learning [12][1][11][13], each dictionary atom is associated with a single class label. Instead of learning universal dictionary for all classes [12][13][15] which the optimization itself is comparatively complicated, some supervised dictionary learning methods learn multiple class-specific dictionary and improve the discriminative capability of reconstruction residual [1][10].

Many algorithms recently have been proposed to enhance the discrimination of the learned dictionary by either imposing class discrimination criterion or enforcing structural constraints on dictionary [10][11][2]. Zhou *et al*.[17] proposed Joint Dictionary Learning method (JDL) that focuses on object classes that share some common visual properties which are difficult to categorized. JDL learns multiple category-specific dictionaries and a common shared dictionary to better exploit the discrimination embodied in the sparse codes.

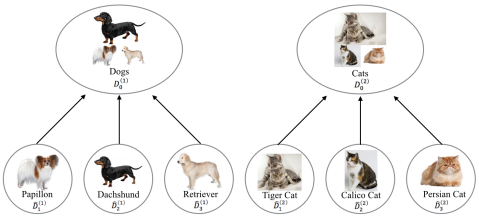

Fig. 1. The first three classes belong to 'Dog'. The common visual properties are characterized by a shared dictionary $D_0^{(1)}$ in the upper level and the class-specific patterns are learned as dictionaries $\hat{D}_{1\cdots3}^{(1)}$ for the lower classes.

However, the previous methods do not consider the problem when there are more than one shared dictionary. In Fig. 1, the lower categories share some common visual properties and belong to *Dog* or *Cat* category in the upper level. Yang *et al*.[11] propose Fisher discrimination dictionary learning that minimizes within-in class scatter and maximizes between-class scatter to learn the shared dictionary. Yang *et al*.[2] exploit a latent vector and a weighted dictionary coherence term in the dictiobary learning to promote the incoherence between dictionary atoms. However, the above methods can only deal with one shared dictionary in the upper level.

In this paper, a novel Hierarchical Structured Dictionary Learning (HSDL) algorithm is proposed to learn class-specific sub-dictionaries and multiple shared dictionaries in hierarchical structure. HSDL discriminates both the lower and the upper classes. Here the lower classes refer to the sub-categories while the upper classes refer to the classes with common visual properties and in the higher level of the structure. In Fig. 1, $D_0^{(1)}$ and $D_0^{(2)}$ are dictionaries of the upper classes. $D_{1\cdots3}^{(1)}$ and $D_{1\cdots3}^{(2)}$ are class-specific dictionaries of the lower classes, which capture the class-specific visual properties. In HSDL, the discrimination term based on within-class scatters and between-class scatters is set for dictionary learning in each level in the hierarchical structure. The main contributions of HSDL are two-fold. Firstly, a novel dictionary learning method is proposed to learn class-

specific dictionaries for all classes and shared classes in the upper level simultaneously for the dataset with hierarchical structure. Secondly, the discrimination term for dictionary learning considers for both the level of upper classes (shared dictionaries) and the level of lower classes (class-specific dictionaries) to enhance the discrimination.

2. HIERARCHICAL STRUCTURED DICTIONARY LEARNING

To learn several-level image categories in dictionary learning, we propose a two-level dictionary learning algorithm. Suppose that in our training data we have M upper classes and C_m lower classes in each upper class, $m = 1, \cdots, M$. Let $X_i^{(m)} \in \mathbb{R}^{d \times N_i^{(m)}}, i = 1, \cdots, C_m$, be a collection of training samples of lower class i in the upper class m. d is the dimension of the training sample and $N_i^{(m)}$ is the number of training samples of the i^{th} lower class in the m^{th} upper class; then we can write $D_i^{(m)} \in \mathbb{R}^{d \times K_i^{(m)}} = [D_0^{(m)}, \hat{D}_i^{(m)}]$ the corresponding dictionary of $X_i^{(m)}$ where $D_0^{(m)} \in \mathbb{R}^{d \times K_0^{(m)}}$ is the dictionary of the m^{th} upper class which is composed of the shared dictionary of all the lower classes under the m^{th} upper class, and is used to describe the common visual properties of those lower classes; here $K^{(m)}$ is the number of atoms of dictionary $D_i^{(m)}$ and $K_0^{(m)}$ is number of atoms of dictionary $D_0^{(m)}$. $\hat{D}_i^{(m)} \in \mathbb{R}^{d \times \left(K_i^{(m)} - K_0^{(m)}\right)}$ is the specific visual properties of the i^{th} lower class in m^{th} upper class. Denote by $A = [A_{1,\cdots,C_1}^{(1)}, A_{1,\cdots,C_2}^{(2)}, \cdots, A_{1,\cdots,C_m}^{(m)}]$ the coefficient matrix of training sample X over dictionary D where $X = [X_{1,\ldots,C_1}^{(1)}, X_{1,\ldots,C_2}^{(2)}, ..., X_{1,\ldots,C_m}^{(m)}]$ and $D = [D_{1,\cdots,C_1}^{(1)}, D_{1,\cdots,C_2}^{(2)}, \cdots, D_{1,\cdots,C_m}^{(m)}]$. We formulate the following HSDL model:

$$\min_{[D_0^{(m)}, \hat{D}_i^{(m)}, A_i^{(m)}]_{i=1 m=1}^{C_m \; M}} \sum_{i=1}^{C_m} \sum_{m=1}^{M} \left\{ \left\| X_i^{(m)} - \left[D_0^{(m)}, \hat{D}_i^{(m)}\right] A_i^{(m)} \right\|_F^2 + \lambda \sum_{j=1}^{K_i^{(m)}} \left\| a_{ij}^{(m)} \right\|_1 \right\} + f(A) \quad (1)$$

where λ is the scalar parameter which relates to the sparsity of the coefficients; $f(A)$ will be described in the next section.

2.1. Discriminative Coefficient term *f(A)*

The term $f(A)$ is designed to make the sparse coefficient be discriminative to the max. Based on the principle of Fisher linear discriminative analysis[3], we minimize the within-class scatter matrix and maximize the between-class scatter simultaneously. We propose two within-class scatters and two between-class scatters instead of one, one for the lower classes and another for the upper classes. The first within-class scatter matrix designed for the lower classes is defined as:

$$S_W = \sum_{j=1}^{C_m} \sum_{a_i \in A_j^{(m)}} (a_i - \mu_j^{(m)})(a_i - \mu_j^{(m)})^T \quad (2)$$

where $\mu_j^{(m)}$ is the mean vector of $A_j^{(m)}$, the sparse coefficient matrix of $X_j^{(m)}$ over $D_j^{(m)} \cdot A_j^{(m)} = [A_{0,j}^{(m)}, \hat{A}_j^{(m)}]$ where $A_{0,j}^{(m)}$ contains the sparse codes over the shared dictionary of the m^{th} upper class $D_0^{(m)}$, and $\hat{A}_j^{(m)}$ is the matrix holding the corresponding coefficients over the class-specific dictionary $\hat{D}_j^{(m)}$ inside the m^{th} upper class. We define the first between-class scatter matrix for the lower classes by excluding the sparse codes over the shared dictionary, given as:

$$S_B = \sum_{j=1}^{C_m} N_j^{(m)} (\hat{\mu}_j^{(m)} - \hat{\mu}^{(m)})(\hat{\mu}_j^{(m)} - \hat{\mu}^{(m)})^T \quad (3)$$

where $\hat{\mu}_j^{(m)}$ is the mean vector of $\hat{A}_j^{(m)}$, and $\hat{\mu}^{(m)}$ is the mean vector of $\hat{A}^{(m)} = [\hat{A}_1^{(m)}, \cdots, \hat{A}_m^{(m)}]$.

The design of within-class and between-class scatter matrices $(S_W^{'}, S_B^{'})$ for the upper classes is the same as that for the lower classes. The dictionary of the m^{th} upper class $D_0^{(m)}$ is the shared dictionary of all the lower classes under the m^{th} upper class. The second within-class scatter matrix is defined as:

$$S_W^{'} = \sum_{m=1}^{M} \sum_{a_i \in A_0^{(m)}} (a_i - \mu_0^{(m)})(a_i - \mu_0^{(m)})^T \quad (4)$$

where $\mu_0^{(m)}$ is the mean vector of $A_0^{(m)}$, the sparse coefficient of the m^{th} upper class. $A_0^{(m)} = [A_{0,1}^{(m)}, A_{0,2}^{(m)}, \cdots, A_{0,C_m}^{(m)}]$. And the second between-class scatter matrix is defined as:

$$S_B^{'} = \sum_{m=1}^{M} N_0^{(m)} (\mu_0^{(m)} - \mu)(\mu_0^{(m)} - \mu)^T \quad (5)$$

where μ is the mean vector of $A_0 = [A_0^{(1)}, \cdots, A_0^{(m)}]$ and $N_0^{(m)}$ is the number of training samples of the m^{th} upper class. According to the equations above, the discriminative coefficient term is therefore defined as:

$$f(A) = \eta_1 \left(t_r(S_W) - t_r(S_B)\right) + \eta_2 \left(t_r(S_W^{'}) - tr(S_B^{'})\right) \quad (6)$$

where parameter $\eta_1, \eta_2 \geq 0$ controls the trade-off between reconstruction and discrimination.

2.2. The HSDL model

Incorporating Eq. 6 into Eq. 1, we have the HSDL model given as:

$$\min_{[D_0^{(m)},\hat{D}_i^{(m)},A_i^{(m)}]_{i=1m=1}^{C_m\ M}} \sum_{i=1}^{C_m}\sum_{m=1}^{M} \left\{ \left\| X_i^{(m)} - \left[D_0^{(m)}, \hat{D}_i^{(m)}\right] A_i^{(m)} \right\|_F^2 + \lambda \sum_{j=1}^{K_i^{(m)}} \left\| a_{ij}^{(m)} \right\|_1 \right\} + \eta_1 \left(t_r(S_W) - t_r(S_B)\right) + \eta_2 \left(t_r(S_W^{'}) - tr(S_B^{'})\right) \quad (7)$$

By learning discriminative coefficients, both the upper and lower layers can be more discriminative from one another.

2.3. Optimization of HSDL

The optimization procedure of the HSDL is to go through three sub-procedures under each upper class: 1) updating the sparse coefficients by fixing both class-specific and shared dictionary, 2) updating the class-specific dictionaries by fixing the coefficients and the shared dictionary, 3) updating the shared dictionary by fixing the coefficients and the class-specific dictionaries and go through 1),2) and 3) until all the upper classes are computed.

Considering that under the m^{th} upper class, dictionaries $D_0^{(m)}, \left\{\hat{D}_i^{(m)}\right\}_{i=1}^{C_m}$ are fixed, then Eq. 7 is reduced to a sparse coding problem. Here we update $\left\{A_i^{(m)}\right\}_{i=1}^{C_m}$ class by class. When update $A_i^{(m)}$, all $X_j, j \neq i$, are fixed so the objective function is reduced to:

$$F\left(A_i^{(m)}\right) = \left\| X_i^{(m)} - \left[D_0^{(m)}, \hat{D}_i^{(m)}\right] \right\| + \lambda \left\| A_i^{(m)} \right\| + f\left(A_i^{(m)}\right) \quad (8)$$

where $F\left(A_i^{(m)}\right)$ is the discrimination constraint derived from $f\left(A_1^{(m)}, \cdots, A_{C_m}^{(m)}\right)$, given as:

$$F\left(A_i^{(m)}\right) = \eta\left(\left\|A_i^{(m)} - M_i^{(m)}\right\|_F^2 - \sum_{j=1}^{C_m}\left\|\hat{M}_j^{(m)} - \hat{M}^{(m)}\right\|_F^2\right) + \eta_2\left(\left\|A_0^{(M)} - M_0^{(m)}\right\|_F^2 - \sum_{k=1}^{M}\left\|M_0^{(k)} - M\right\|_F^2\right) \quad (9)$$

where $M_i^{(m)}$ is the mean vector matrices by taking $N_i^{(m)}$ copies of mean vector $\mu_i^{(m)}$ as its columns, $\hat{M}_j^{(m)}$ and $\hat{M}^{(m)}$ are produced by $N_j^{(m)}$ copies of $\hat{\mu}_j^{(m)}$ and $\hat{\mu}^{(m)}$ as their column vectors, $M_0^{(m)}$ consist of $N_0^{(m)}$ copies of the mean vector $\mu_0^{(m)}$ as its columns, finally, $M_0^{(k)}$ and M contains $N_0^{(k)}$ copies of the mean vector μ as its column vectors, $k \neq m$. We can see that except for the l_i penalty term, the other two terms in Eq. 8 are differentiable. There are several l_i-Minimization algorithms to solve it[4], and here we adopt TwIST (two-step iterative shrinkage/thresholding).

Considering under the m^{th} upper class the coefficients are fixed, there are two steps for updating dictionaries. First we update the class-specific dictionary $\left\{\hat{D}_i^{(m)}\right\}_{i=1}^{C_m}$ class by class and then the shared dictionary of the m^{th} upper class $D_0^{(m)}$. When $\left\{A_i^{(m)}\right\}_{i=1}^{C_m}$ and $D_0^{(m)}$ are fixed, the objective function is reduced to:

Algorithm 1 Hierarchical Structured Dictionary Learning

1: **Input:** training data $\left\{X_{1\cdots C_m}^{(m)}\right\}^M$, the scalar parameter λ, discrimination parameter η_1, η_2
2: **Initialization:**
3: Initialize $\left\{D_{1\cdots C_m}^{(m)}\right\}_{m=1}^{M}$ using K-SVD[6]
4: Initialize $\left\{D_0^{(m)}\right\}_{m=1}^{M}$ by stacking the atoms in $D_{1\cdots C_m}^{(m)}$ whose inner products are larger than the threshold ξ columns by columns. Form the initial $\left\{\hat{D}_{1\cdots C_m}^{(m)}\right\}_{m=1}^{M} s.t.\ D_i^{(m)} = \left[D_0^{(m)}, \hat{D}_i^{(m)}\right]$
5: Initialize $\left\{A_{1\cdots C_m}^{(m)}\right\}_{m=1}^{M}$ using CVX[7]
6: Repeat until convergence
7: **for** each upper class $m = 1 \cdots M$
8: **for** each lower class $i = 1 \cdots C_m$
9: update $A_i^{(m)}$ by solving Eq. 8
10: **for** each lower class $i = 1 \cdots C_m$
11: update $D_i^{(m)}$ by solving Eq. 10
12: **for** each lower class $i = 1 \cdots C_m$
13: update $D_0^{(m)}$ by solving Eq. 11
14: **Output:** $\left\{D_{1\cdots C_m}^{(m)}\right\}_{m=1}^{M}, \left\{D_0^{(m)}\right\}_{m=1}^{M}$

$$\min_{\hat{D}_i^{(m)}} \left\| X_i^{(m)} - D_0^{(m)} A_{0,i}^{(m)} - \hat{D}_i^{(m)} \hat{A}_i^{(m)} \right\|_F^2 \quad (10)$$
$$s.t. \left\| \hat{d}_j^{(m)} \right\|_2^2 \leq 1, \forall j = 1, \cdots, K_i^{(m)}$$

After updating $\left\{\hat{D}_i^{(m)}\right\}_{i=1}^{C_m}$, we update $D_0^{(m)}$. The objective function is given as:

$$\min_{\hat{D}_0^{(m)}} \left\| X_0^{(m)} - D_0^{(m)} A_0^{(m)} \right\|_F^2 \quad (11)$$
$$s.t. \left\| d_j^{(m)} \right\|_2^2 \leq 1, \forall j = 1, \cdots, K_0^{(m)}$$

where

$$A_0^{(m)} \triangleq \left[A_{0,1}^{(m)}, \cdots, A_{0,C_m}^{(m)}\right] \quad (12)$$

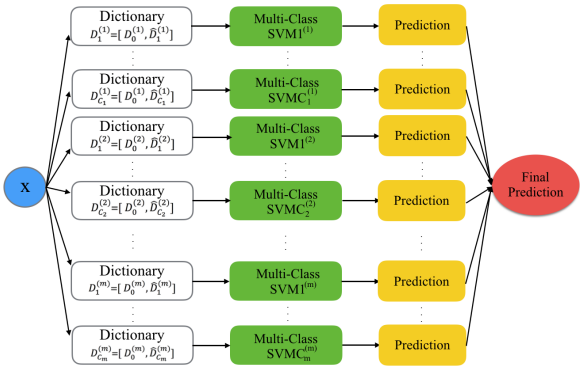

Fig. 2. Classification scheme of multiple learned dictionaries.

$$X_0^{(m)} \triangleq \left[X_1^{(m)} - \hat{D}_1^{(m)}\hat{A}_1^{(m)}, \cdots, X_{C_m}^{(m)} - \hat{D}_{C_m}^{(m)}\hat{A}_{C_m}^{(m)}\right] \tag{13}$$

Eq. 10 and 11 are both quadratic programming problems which can be solved by using the Lagrange duals [5].

2.4. Classification

In [18], multiple linear SVMs were trained by taking the sparse representation over dictionaries as features to produce the final prediction via equal voting scheme by combining the outputs of the classifiers. Here we modified the classification approach proposed in [18]. In each category we have the corresponding shared dictionary, and each dictionary which is used to train SVM contains learned shared dictionary and learned class-specific dictionary ($D_i^{(m)} = \left[D_0^{(m)}, \hat{D}_i^{(m)}\right]$). We use both shared and class-specific dictionaries to train SVM. After training multiple linear SVMs, we also use equal voting scheme to produce final prediction. The illustration of the classification scheme is shown in Fig. 2.

3. EXPERIMENTAL RESULTS

We applied HSDL to the tasks of object recognition and the experiment was evaluated on the ImageNet database. We compare the performance of the proposed HSDL algorithm with some previous methods, including the state-of-the-art learning methods and those using only one shared dictionary, as shown in Table 1. Our image classification method first obtains salient features from an input image using the SIFT[16] interest point detector. We extract 100 interest points in each image. After that, our method clusters these descriptors into N centroids by using the standard K-means unsupervised learning algorithm. The extracted descriptors are used to compute the Bag-Of-Words (BoW) histogram vector for the image. We represent each image by its histogram obtained by the hard assignment of each local feature to BoW clusters. The histogram is normalized by the size of the BoW. Finally, we let SVM to classify the images based on their BoW. The dictionary sizes were set to be equal across all categories. The scalar parameter λ is set to 0.1, both discrimination parameter η_1, η_2 are set to 0.1 and the similarity threshold ξ between every two columns of lower classes dictionaries be 0.94, 0.925 respectively in the two experiments.

Table 1. Recognition accuracy(%) on ImageNet dataset

Dictionary Size of Each Category	100	1000
SVM [18]	0.389	0.490
K-SVD [6]	0.378	0.488
LC-KSVD [14]	0.317	0.274
DPL [19]	0.339	0.477
SVGDL [20]	0.378	0.484
JDDLDR [21]	**0.394**	0.466
FDDL [11]	0.392	0.420
HSDL	**0.394**	**0.492**

3.1. Evaluation on ImageNet database

We choose two groups of animals which are visually similar as our training data. In each group we have three different categories, so there are totally six categories. In each category we use 1300 images in the first experiment and 130 images in the second. Sample images are shown in Figure 1. The ratio of training data and testing data was both set to 10:3.The experimental results are shown in Table 1. The ratio of number of atoms of $\hat{D}^{(m)}$ to number of atoms of $D_0^{(m)}$ is found to be approximately 4:1 in both experiments of both groups, and for the simplicity of experiment we randomly pick some atoms of $\hat{D}^{(m)}$ and $D_0^{(m)}$ to set their ratio to be 4:1 and to make the dictionary sizes of both groups in each experiment to be equal. It can be seen in Table 1 that the performance of some methods was affected by different dictionary sizes. However, HSDL outperforms all the other methods for both the experiments with smaller and larger dictionary sizes.

4. CONCLUSION

We proposed a novel sparse coding based Hierarchical Structured Dictionary Learning (HSDL) algorithm to exploit the visual correlation within multiple object categories. HSDL learns multiple class-specific dictionaries and shared dictionaries. A discriminative term that is based on Fisher discrimination criterion is developed for both the level of class-specific dictionaries and shared dictionaries. Our experimental results revealed that the proposed HSDL algorithm is superior to the previous dictionary learning methods for the problems of classifying visually similar objects. Our future work is to combine the projection matrix into the inter-related dictionary to learn a compact and hierarchical representation.

5. REFERENCES

[1] J. Mairal, F. R. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA, 2008.

[2] M. Yang, D. Dai, L. Shen, and L. Van Gool. Latent dictionary learning for sparse representation based classification. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 4138-4145, June 2014.

[3] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification (2nd Edition). Wiley-Interscience, 2 edition, 2001.

[4] A. Y. Yang, A. Ganesh, Z. Zhou, S. Shankar Sastry, and Y. Ma. A Review of Fast l1-Minimization Algorithms for Robust Face Recognition. ArXiv e-prints, July 2010.

[5] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In NIPS, pages 801-808, 2006.

[6] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. Signal Processing, IEEE Transactions on, 54(11):4311-4322, November 2006.

[7] S.Becker, E.J.Candes and M.Grant. Templates for Convex Cone Problems with Applications to Sparse Signal Recovery. Stanford University Technical Report, September 2010.

[8] Q. Zhang and B.X. Li. Discriminative K-SVD for dictionary learning in face recognition. In CVPR, 2010.

[9] D. Pham and S. Venkatesh. Joint learning and dictionary construction for pattern recognition. In CVPR, 2008.

[10] l. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In CVPR, pages 3501-3508, 2010.

[11] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In ICCV, 2011.

[12] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In NIPS, pages 1033-1040, 2008.

[13] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In CVPR, pages 2691 -2698, june 2010.

[14] Z. Jiang, Z. Lin, and L. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In CVPR, pages 1697-1704, june 2011.

[15] J. Yang, K. Yu, and T. S. Huang. Supervised translation-invariant sparse coding. In CVPR, pages 3517-3524, 2010.

[16] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In CVPR, pages 2559-2566, 2010.

[17] Ning Zhou, Yi Shen, Jinye Peng and Jianping Fan. Learning inter-related visual dictionary for object recognition. In CVPR, page 3490-349, 2012

[18] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1-27:27, 2011.

[19] S. Gu, L. Zhang, W. Zuo, and X. Feng. Projective dictionary pair learning for pattern classification. In Advances in Neural Information Processing Systems, pages 793-801, 2014.

[20] S. Cai, W. Zuo, L. Zhang, X. Feng, and P. Wang. Support vector guided dictionary learning. Proc. 13th Eur. Conf. Comput. Vis. (ECCV), pp. 624-639, 2014.

[21] Z. Feng, M. Yang, L. Zhang, Y. Liu, and D. Zhang. Joint discriminative dimensionality reduction and dictionary learning for face recognition. Pattern Recognition, 46(8):2134-2143, 2013.