Discriminative Paired Dictionary Learning for Visual Recognition

Hui-Hung Wang
National Chung Cheng University, Taiwan
huihung55@gmail.com

Yi-Ling Chen
National Taiwan University, Taiwan
yilingchenntu@ntu.edu.tw

Chen-Kuo Chiang
National Chung Cheng University, Taiwan
ckchiang@cs.ccu.edu.tw

ABSTRACT

A Paired Discriminative K-SVD (PD-KSVD) dictionary learning method is presented in this paper for visual recognition. To achieve high discrimination and low reconstruction errors simultaneously for sparse coding, we propose to learn class-specific sub-dictionaries from pairs of positive and negative classes to jointly reduce the reconstruction errors of positive classes while keeping the reconstruction errors of negative classes high. Then, multiple sub-dictionaries are concatenated with respect to the same negative class so that the non-zero sparse coefficients can be discriminatively distributed to improve classification accuracy. Compared to the current dictionary learning methods, the proposed PD-KSVD method achieves very competitive performance in a variety of visual recognition tasks on several publicly available datasets.

Keywords

Dictionary learning; sparse coding; visual recognition

1. INTRODUCTION

In recent years, sparse coding is applied to a variety of image processing and computer vision applications, such as image classification [28, 15, 22, 26, 27, 23], image de-noising [3], compression [2, 8], inpainting [3] and other applications. It recovers a sparse linear representation of a query datum with respect to a set of non-parametric basis set, known as *dictionary*. Originally, predefined dictionaries have been used based on various types of wavelets. Lately, learning the dictionary instead of using predefined bases has been shown to improve signal reconstruction significantly [2].

K-SVD [2] aims to learn the optimal dictionary that leads to the lowest reconstruction error with a set of sparse coefficients. The success of K-SVD method triggers the application in image classification tasks. Instead of learning one dictionary for the whole dataset, class-specific sub-dictionaries are learned to improve the discriminative capability of reconstruction residual [17, 20, 6]. Later, researchers attempts to promote discrimination among classes by enforcing the sparse coefficients to be discriminative[28, 15, 22, 26, 23, 27] by means of including the class labels and learning the classifier simultaneously with the dictionary. Recent dictionary learning methods achieve the discrimination by either enforcing structural constraints on the dictionary or imposing class separation criterion in the sparse coding formulation [27, 26, 23]. As a result, previous dictionary learning methods usually has very complex objective functions and require difficult optimization approaches to obtain the solutions.

MM '16, October 15–19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2967184

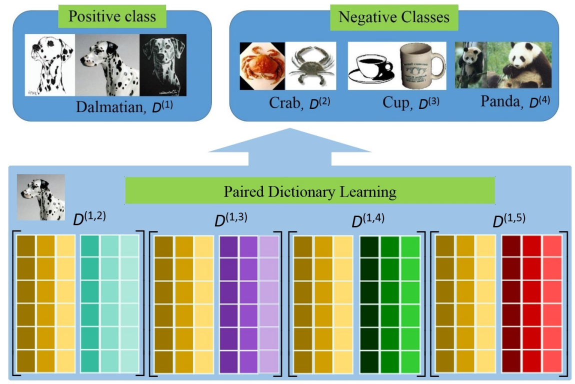

Figure 1: An example of positive and negative classes and the corresponding paired dictionaries of the class "Dalmatian".

To develop a simple yet effective dictionary learning method, we go back to the essence of sparse coding that uses the reconstruction error to achieve the discrimination. Motivated by the current machine learning methods, discriminative classifiers are usually trained by positive and negative data samples. Pair-wise relationships are utilized between data samples or datasets to further improve the discrimination. In [14], highly discriminative variables are selected from the high-dimensional data space according to the statistical analysis of paired samples for feature selection. In [16], a multilinear canonical correlation analysis method is proposed to extract features directly from tensors based on tensor-to-vector projection of paired tensor sets.

We argue that the discrimination of dictionary can be also achieved through learning from pairs of classes. To learn the class-specific sub-dictionaries, images can be divided into two kinds of classes: *Positive class* and *Negative class*. Positive class is the target class. Negative classes are all the rest of classes. For example, when training the sub-dictionary for class "Dalmatian", the training class "Dalmatian" is the positive class. The rest classes ("Crab", "Cup", "Panda") are negative classes in this case. Figure 1 depicts an example of using positive class and negative classes to learn sub-dictionaries for the class "Dalmatian".

In this paper, we propose a novel Paired Discriminative K-SVD (PD-KSVD) based on K-SVD [2] to learn sub-dictionaries from pairs of positive and negative classes. The idea is to learn such sub-dictionaries that can best reconstruct the positive class while reconstructing the negative class worse. With such capability of reconstruction, multiple sub-dictionaries are concatenated with respect to the same negative class to discriminatively distribute the non-zero coefficients. The sparse coefficients are then used to train multiple SVM classifiers [4].

2. PREVIOUS WORK

K-SVD [2] learns the dictionary by iteratively alternating between sparse coding the input data based on the current dictionary, and updating the dictionary atoms to better fit the training data. It aims to find the optimal dictionary that leads to the lowest reconstruction error with a set of sparse coefficients. Typical reconstructive dictionary learning methods include method of optimal direction (MOD) [9], K-SVD and analysis K-SVD [21].

The success of K-SVD method triggers the application in image classification tasks. Rather than learning one dictionary for all classes, one type of dictionary learning methods learn a sub-dictionary for each class and improving the discriminative capability of reconstruction residual [17, 20, 6]. In the class-specific dictionary learning, each dictionary atom is associated to a single class label. Mairal *et al.* [17] assumed a dictionary associated to the class should reconstruct this class better than the other classes. A penalty term is introduced in the cost function to re-weight the reconstruction error. Ramirez *et al.* [20], introduce a new term into sparse representation to promote the incoherence between dictionaries so that each sub-dictionary can represent the class optimally.

3. PROPOSED METHOD

3.1 Reconstruction Error of Inverted Signals

Following the nature of K-SVD that reduces the reconstruction errors of both positive and negative classes, our idea is to *invert* the negative signals. If the dictionary is learned from the inverted negative samples, the dictionary can reconstruct the the inverted negative samples well. In other words, using such dictionary to construct the original negative samples leads to larger reconstruction errors.

The Caltech101 dataset [10] was used to evaluate our assumption. we first set class 2 as our positive class. Then, sub-dictionary was trained by only using samples from class 2. We create a new class by applying the inversion operation to all samples in class 2. Using the trained sub-dictionary, we calculate the reconstruction errors for all 102 classes (101 objects and one background class) in Caltect101 and the inverted class. The results are presented in Figure 2.

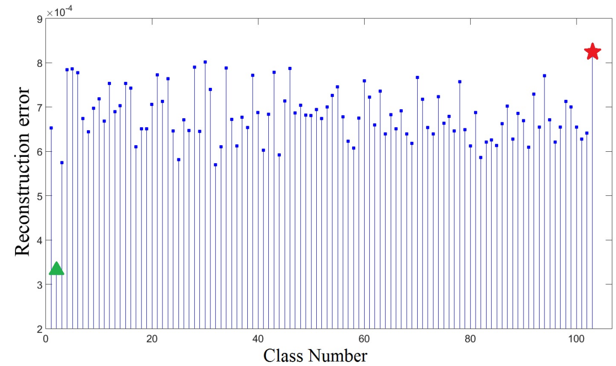

Figure 2: Reconstruction errors of 102 classes of Caltech101 dataset and the inverted class. The dictionary is trained by class 2. The green triangle presents the lowest reconstruction error of positive class 2. The red star sign depicts the highest reconstruction error of the inverted class among all the classes.

One can see that the reconstruction error of class 2 is the lowest (marked by the green triangle sign) since the sub-dictionary was trained from class 2. The inverted class (marked by red star sign) has the highest reconstruction error among all classes. Similar results can be obtained by choosing another class as positive class and repeat the experiment above.

We should note that the *signals* here are referred to as the image features. Each feature can be considered as a vector of signal. For example, one SIFT feature is a 128-dimensional vector of signal. Figure 3 depicts one possible operation of signal inversion. Signals can be inverted according to the data mean of minimum and maximum values in this class. The *max* value of class c can be calculated by: $\max^{(c)} = \max(\max(y_1^{(c)}, ..., y_n^{(c)}))$ which calculates a maximum vector from all feature vectors $y_1^{(c)}, ..., y_n^{(c)}$. Then, applying another *max* operation to extract the maximum value of class c. The signal mean can be defined by $mean^{(c)} = (1/2)(\max^{(c)} + \min^{(c)})$. Denote $y_{i,j}^{(c)}$ the j-th element of $y_i^{(c)}$, the inverted signal $\hat{y}_{i,j}^{(c)}$ can be obtained by $\hat{y}_{i,j}^{(c)} = 2 * mean^{(c)} - y_{i,j}^{(c)}$. Under this operation, signals are symmetry to class mean. The maximum value becomes the minimum and vice versa. The operation of inversion can be also conducted by finding the orthogonal vector of signal $y_i^{(c)}$, which means to find a $\hat{y}_i^{(c)}$ such that $y_i^{(c)} \cdot \hat{y}_i^{(c)} = 0$. In our experiments, the difference of classification accuracy is around 0.1%.

3.2 Paired Discriminative K-SVD

Denote $Y^{(c)}, c = 1, ..., C$ the positive training samples of class c. The inverted signals can be defined by $\hat{Y}^{(c)}$. Our goal of dictionary learning is to learn sub-dictionaries that can simultaneously increase the reconstruct error of negative classes while reducing the reconstruction error of the positive class. To this end, our dictionary learning is performed on a pair of one positive class and one inverted negative class. If the learned dictionary achieves low reconstruction errors of both the positive class and the inverted negative class in a K-SVD fashion, it introduces high reconstruction errors to the original negative class. This boosts the discrimination capability between the positive class and the negative class.

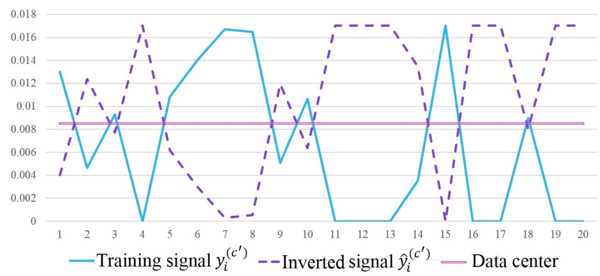

Figure 3: An example of one inverted signal.

To obtain the sub-dictionary of the positive class c, we choose one negative class c' with the minimal reconstruction error. Then we form the training set by joining $Y^{(c)}$ and the inverted negative class $\hat{Y}_i^{(c')}$. The objective function of paired discriminative dictionary learning can be written as:

$$min_{D^{(c,c')},X^{(c,c')}} ||[Y^{(c)}\hat{Y}^{(c')}] - D^{(c,c')}X^{(c,c')}||_F^2, \quad s.t.\forall i, ||x_i^{(c,c')}||_1 \leq T_0 \tag{1}$$

where $D^{(c,c')}$ is the dictionary learned from the paired positive class c and the inverted negative class c'. $X^{(c,c')}$ is the sparse coefficients. Then, the reconstruction residual can be formulated as:

$$\begin{aligned} &||[Y^{(c)}\hat{Y}^{(c')}] - D^{(c,c')}X^{(c,c')}||_F^2 \\ &= ||[Y^{(c)}\hat{Y}^{(c')}] - \sum_{j=1}^{K} d_j^{(c,c')} x_T^{j(c,c')}||_F^2 \\ &= ||[Y^{(c)}\hat{Y}^{(c')}] - \sum_{j\neq k} d_j^{(c,c')} x_T^{j(c,c')} - d_k^{(c,c')} x_T^{k(c,c')}||_F^2 \\ &= ||e_k^{(c,c')} - d_k^{(c,c')} x_T^{k(c,c')}||_F^2 \end{aligned} \tag{2}$$

The reconstruction error when taking out the k^{th} atom can be written as:

$$e_k^{(c,c')} = [Y^{(c)}\hat{Y}^{(c')}] - \sum_{j\neq k} d_j^{(c,c')} x_T^{j(c,c')} \tag{3}$$

To learn dictionary column $d_k^{(c,c')}$, SVD is applied to decompose $e_k^{(c,c')}$ to three matrices: $e_k^{(c,c')} = U\triangle V^T$. The first column of U is used to update dictionary column $d_k^{(c,c')}$ and first column of V multiplied by $\triangle$ $(1,1)$ is exploited to update the row $x_T^{k(c,c')}$ of the coefficient matrix. In PD-KSVD, a sub-dictionary is learned from the positive class c and one of its negative class c'.

3.3 Classification

To efficiently exploit all sub-dictionaries for classification, all sub-dictionaries with respect to the same negative class are combined. Denote the combined dictionary with respect to negative class c', $D_{com}^{(c')} = [D^{(1,c')}, D^{(2,c')}, ..., D^{(i,c')}], i = 1, ..., C, i \neq c'$. Therefore, $D_{com}^{(c')}$ can not reconstruct samples of class c' well. On the contrary, it can well construct class $c, c = 1, ..., C, c \neq c'$.

The reasons to combine such sub-dictionaries are twofold. Firstly, for a test sample y_1 of class 1, using $D_{com}^{(2)} = [D^{(1,2)}, D^{(3,2)}, ..., D^{(6,2)}]$ to reconstruct y_1 makes the nonzero coefficients more concentrated on dictionary atoms of $D^{(1,2)}$, since $D^{(1,2)}$ is trained by positive samples of class 1 and inverted negative samples of class 2. To better reconstruct y_1, more dictionary atoms in $D^{(1,2)}$ which are trained from class 1 should be selected for reconstruction instead of using dictionary atoms trained from inverted negative samples of class 2. This makes the non-zero coefficients more concentrated to the dictionary atoms of the correct class. Secondly, after combining all sub-dictionaries with respect to the same negative class, it reduces the number of dictionaries which could have practical significance in terms of computational efficiency in real-world applications.

In the classification process, training signals are sparsely coded by dictionary $D = [D_{com}^{(1)}, D_{com}^{(2)}, ..., D_{com}^{(C)}]$. The corresponding sparse coefficients $SP_{train}^{(1)}$, $SP_{train}^{(2)}$,...,$SP_{train}^{(C)}$ are used to train multiple multi-class SVMs: $SVM^{(1)}$, $SVM^{(2)}$,...,$SVM^{(C)}$. After obtaining all the decisions from SVMs, a major voting mechanism is exploited to predict the class label. The process of classification is depicted in Figure 4.

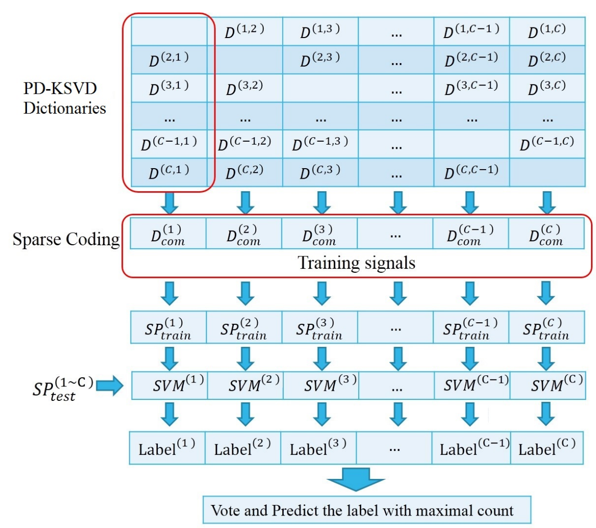

Figure 4: The classification process using combined sub-dictionaries.

4. EXPERIMENTAL RESULTS

We applied PD-KSVD to the tasks of object recognition, handwritten digits recognition and scene recognition in our experiments. Our method was evaluated on the following datasets: Caltech101 database [10], Caltech256 database [12], USPS database [1] and MIT Indoor Scene Database [19]. The parameter setting of PD-KSVD is easy. Our method is nearly parameter-free except for the sparsity factor when solving the sparse coefficients. Following the experimental settings in [15], the sparsity factor was set to 30 in our experiments and the maximum number of iterations was set to 80 following [2].

We compared PD-KSVD to SVM [7], K-SVD [2], Label Consistent K-SVD (LC-KSVD) [15], Dictionary Pair Learn-

Table 1: **Recognition accuracy (%) on Caltech256 Dataset with dictionary size 7710 and 15420, respectively.**

Dictionary Size	7710	15420
SVM [7]	18.5	24.5
K-SVD [2]	15.8	20.2
DPL [13]	15.9	21.8
SVGDL [5]	19.9	23.2
JDDLDR [11]	18.3	21.3
PD-KSVD	**22.2**	**29.3**

ing (DPL)[13], Support Vector Guided Dictionary Learning(SVGDL) [5], Joint Discriminative Dimensionality Reduction and Dictionary Learning(JDDLDR) [11]. We used OMP [18] for solving the optimization problem of sparse coding. SVM is used as the baseline method in our comparison. K-SVD and another K-SVD based method LC-KSVD were also included in our comparison. DLP and SVGDL which also explore the pair-wise relationship in dictionary learning achieve the state-of-the-art performance.

We use bag-of-words (BOW) of 200 vocabulary and spatial pyramid matching (SPM) as feature representation of each image. SIFT descriptors are extracted on three level spatial pyramid of sizes 1×1, 2×2 and 4×4 regions. Then, we obtain 4200-dimension features for image representation. We also extract LLC [24] feature with 1204 vocabulary in BOW model with the SPM representation of sizes 1×1, 2×2 and 4×4. Finally, the 21,504 dimensional feature is extracted.

The reference codes of the methods listed above are all available on authors' websites. We ran their reference codes with features extracted as aforementioned to compare the experimental results. Therefore, the accuracy might be different from those reported in the original papers.

4.1 Caltech 256 Object Dataset

Caltech 256 [12] contains 30607 images in 256 object categories and one background class. It is a very challenging dataset because the number of object categories is very high and contains more object variations in each category. Following the custom setup, 30 and 60 images per class are randomly selected for training, and all the rest for testing. SPM feature is used for performance evaluation.

The experiments are repeated three times and the average accuracy are reported. The results are summarized in Table 1. On this benchmark, the proposed PD-KSVD outperforms all these five methods by using 7710 and 15420 dictionary atoms. It achieves nearly 5% higher than the second best recognition accuracy with 15420 dictionary atoms.

4.2 USPS Handwritten Digit Dataset

We then perform handwritten digit recognition on the USPS dataset [1]. It contains images with contents from ”0” to ”9” in 8-bit grayscale format. The dataset is composed of 10 classes with 1100 images for each class. The images in the dataset are resized to 16×16. Then, features are extracted by the 256-dimensional histogram of the image intensity in grayscale.

In this experiment, we use 100 and 500 images per class for training, and the rest 1000 and 600 images for testing. The results are summarized in Table 2. From the table,

Table 2: Recognition accuracy (%) on USPS Dataset.

Dictionary Size	1000	5000
SVM [7]	85.3	87.8
SRC [25]	86.2	95.0
K-SVD [2]	84.7	88.8
LC-KSVD [15]	83.7	89.7
DPL [13]	**89.8**	**95.2**
SVGDL [5]	88.7	91.8
JDDLDR [11]	86.0	87.8
PD-KSVD	87.7	92.5

Table 3: Recognition accuracy (%) on MIT Indoor Scene Dataset with dictionary size 5360.

Method	Accuracy
SVM [7]	15.3
K-SVD [2]	14.8
LC-KSVD [15]	12.6
DPL [13]	13.3
SVGDL [5]	20.9
JDDLDR [11]	16.9
PD-KSVD	**21.4**

DPL [13] outperforms other methods. Our PD-KSVD also provide comparable recognition accuracy for handwritten digit recognition. It is worth pointing out that the recognition can achieve 93.5 when PD-KSVD trains dictionary of size 2500.

4.3 MIT Indoor Scene Dataset

MIT Indoor Scene dataset contains 67 Indoor categories. It is a very challenging dataset because of the large within-class variety and significant between-class confusion. Each category has at least 100 images. The dataset has 15620 images in total. We use 80 images per class for training, and all the rest for testing. SPM feature is used for performance evaluation.

Table 3 lists the recognition results of PD-KSVD and the competing methods. It can be seen that accuracy of PD-KSVD is superior to all the other methods.

5. CONCLUSIONS

In this paper, we propose PD-KSVD algorithm to learn class-specific sub-dictionaries by using pairs of positive and negative classes. The data samples from the negative class are inverted to train the sub-dictionary, making the reconstruction of samples from the original negative class worse by the learned dictionary. Overall, PD-KSVD is simple yet effective. It also holds a good property that better recognition results can be achieved by reducing the dictionary size. This makes it efficient for real applications. The experimental results show that the accuracy of PD-KSVD is very competitive to others methods in all recognition tasks. In the future, we would like to analyze the importance of each sub-dictionary. It will be interesting to weighted combine these sub-dictionaries based on the analysis of their importance to improve the discrimination of dictionary learning.

6. REFERENCES

[1] Usps handwritten digit database. http://www-i6.informatik.rwth-aachen.de/ keysers/usps.html.

[2] 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, Nov 2006.

[3] C. Bao, J.-F. Cai, and H. Ji. Fast sparsity-based orthogonal dictionary learning for image restoration. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3384–3391, Dec 2013.

[4] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, 1998.

[5] S. Cai, W. Zuo, L. Zhang, X. Feng, and P. Wang. Support vector guided dictionary learning. In *Computer Vision–ECCV 2014*, pages 624–639. Springer, 2014.

[6] A. Castrodad and G. Sapiro. Sparse modeling of human actions from motion imagery. *International Journal of Computer Vision*, 100(1):1–15, 2012.

[7] 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. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[8] B. Deka, K. Gorain, N. Kalita, and B. Das. Single image super-resolution using compressive sensing with learned overcomplete dictionary. In *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on*, pages 1–5, Dec 2013.

[9] K. Engan, S. O. Aase, and J. H. Husøy. Frame based signal compression using method of optimal directions (MOD). In *International Symposium on Circuits and Systems (ISCAS 1999), May 30 - June 2, 1999, Orlando, Florida, USA*, pages 1–4, 1999.

[10] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[11] 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.

[12] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[13] 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.

[14] S. Heng Toh, P. Prathipati, E. Motakis, K. Chee Keong, P. Yenamandra, and V. Kuznetsov. A robust tool for discriminative analysis and feature selection in paired samples impacts the identification of the genes essential for reprogramming lung tissue to adenocarcinoma. *BMC Genomics*, 12(3), 2011.

[15] Z. Jiang, Z. Lin, and L. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. pages 1697–1704, June 2011.

[16] H. Lu. Learning canonical correlations of paired tensor sets via tensor-to-vector projection. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, 2013.

[17] 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.

[18] Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. pages 40–44 vol.1, Nov 1993.

[19] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 413–420, 2009.

[20] I. Ramírez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3501–3508, 2010.

[21] R. Rubinstein, T. Peleg, and M. Elad. Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model. *IEEE Transactions on Signal Processing*, 61(3):661–677, 2013.

[22] L. Shen, S. Wang, G. Sun, S. Jiang, and Q. Huang. Multi-level discriminative dictionary learning towards hierarchical visual categorization. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 383–390, June 2013.

[23] Y. Sun, Q. Liu, J. Tang, and D. Tao. Learning discriminative dictionary for group sparse representation. *Image Processing, IEEE Transactions on*, 23(9):3816–3828, Sept 2014.

[24] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3360–3367, 2010.

[25] 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.

[26] 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.

[27] M. Yang, D. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 543–550, Nov 2011.

[28] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. pages 2691–2698, June 2010.