IMAGE CLASSIFICATION: A HIERARCHICAL DICTIONARY LEARNING APPROACH

Shahin Mahdizadehaghdam, Liyi Dai, Hamid Krim, Erik Skau, Han Wang*

North Carolina State University,
Department of Electrical and Computer Engineering,
Raleigh, NC
*Army Research Office, RTP, Raliegh, NC

ABSTRACT

Hierarchical dictionary learning seeks multiple dictionaries at different image scales to capture complementary coherent characteristics. We propose a method to learn a hierarchy of two overcomplete synthesis dictionaries with an image classification goal. The classification objective in some sense regularizes the joint optimization of the hierarchical dictionaries and injects refinement feedback. The validation of the proposed approach is based on its classification performance using two well-known data sets.

Index Terms— image classification, dictionary learning, sparse coding, dimension reduction, image processing

1. INTRODUCTION

Parsimonious data representation by learned overcomplete dictionaries has recently shown promising results in a variety of problems such as image denoising [1, 2, 3], image restoration [4, 5, 6], audio processing [7, 8], and image classification [9, 10]. This frame-like representation of each data vector as a linear combination of atoms, carries a sparse notion of the associated coefficients. Hence the so-called sparse coding is based on an overcomplete dictionary. Dictionary learning has also been shown to be more flexible in adapting the representation to different tasks.

Over the last decade, many techniques have been developed to perform dictionary learning, and sparse representation. K-SVD [11] is based on a generalized k-means clustering method which efficiently learns an overcomplete dictionary from the training samples. The K-SVD method can be used in conjunction with orthogonal matching pursuit (OMP) [12] to learn the L_0 sparse representations of the data samples. Iterative hard thresholding [13] and coordinate descent algorithm can also be applied to learn an overcomplete dictionary, and the L_1 sparse representations of data. Many other dictionary learning methods such as online dictionary learning [14] have been developed with large training sets in mind.

The above mentioned methods achieve sparse feature vectors with no account for the target task which will exploit the extracted feature vectors. Other works have accounted for the target task, and have yielded higher performances especially in image classification task. Task driven dictionary learning [15] obtains adapted dictionary and sparse representation by minimizing a classification cost function. Label Consistent K-SVD (LC-KSVD) [16] consists of a K-SVD based algorithm to find sparse feature vectors belonging to the same class close to each other, and sparse feature vectors belonging to different classes far from each other.

In this work, we propose to jointly learn a hierarchy of two overcomplete synthesis dictionaries with a minimal classification error as a goal. The resulting representations of images at two different scales, are subsequently used to classify images. Intuitively, the first scale captures the fine low level structures comprising the image vectors used for learning, while the second scale coherently captures more complex structures. The classification is ultimately carried out by assembling the second scale features of an image together and assessing their contribution.

The hierarchical framework learns general structures for the first scale, and the local relationships between them on the second scale for each image class group. Therefore, it, to a large extent, attenuates the subtle differences among the images within the same class. This is by virtue of the scale-based representation, assembling low level features and reconciling those differences and ultimately enhancing the performance. By simultaneously minimizing the error of image classification, we iteratively adapt the dictionaries to help build better feature vectors for the very task.

This paper is organized as follows: We formulate and propose our new approach in Section 2, and present substantiating experimental results in Section 3. We provide some concluding remarks in Section 4.

2. THE PROPOSED METHOD

The classical approach to image classification starts with representing each image as a sparse linear combination of atoms from a synthesis dictionary [14], as shown in Eqn. (1). This is typically followed by classification techniques such as SVM [17], neural networks or linear classifiers operating on the

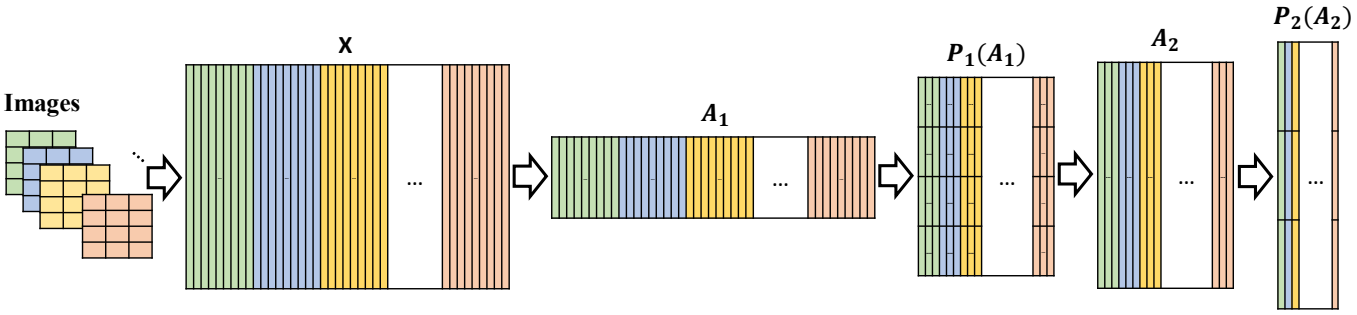


Fig. 1: Sequential steps of a hierarchical dictionary learning.

sparse feature vectors. To that end, one vectorizes all training images into a matrix $\boldsymbol{G}$ and uses K-SVD [11] and OMP [12] to perform dictionary learning and sparse representation,

$$\begin{aligned} &\arg\min_{\boldsymbol{D},\boldsymbol{A}} ||\boldsymbol{G}-\boldsymbol{D}\boldsymbol{A}||_F^2, \\ s.t. \quad &||\boldsymbol{a}_i||_0 < T \ \ \forall i = 1,...,N, \\ &\boldsymbol{D} \in \mathcal{C}, \end{aligned} \tag{1}$$

where $\boldsymbol{G}$ is an $L \times N$ matrix [1]. Each column of $\boldsymbol{G}$ is a vectorized image, and N is the total number of training images. Dictionary $\boldsymbol{D}$ is an $L \times K$ matrix with K atoms, and $\boldsymbol{A}$ is a sparse representation of the $\boldsymbol{G}$ matrix. $\mathcal{C}$ is the convex set of matrices which have columns with unit L_2-norm.

2.1. Problem Formulation

Joint dictionary learning and image classification have been studied in a few recent papers [15, 16, 18]. In this paper, we propose hierarchical dictionaries through learning the first layer dictionary $\boldsymbol{D}^{(1)}$ and the associated feature vectors $\boldsymbol{A}^{(1)}$ on small image patches. The second layer dictionary $\boldsymbol{D}^{(2)}$ is learned on the concatenated feature vectors of a subset of a few adjacent patches from the first layer. The sparse representation of the patches of an entire image are vectorized into a single vector to proceed with classification. The processes of hierarchical dictionary learning and image classification are the result of solving the following functional:

$$\begin{aligned} &\arg\min_{\substack{\boldsymbol{D}^{(1)},\boldsymbol{A}^{(1)},\\ \boldsymbol{D}^{(2)},\boldsymbol{A}^{(2)},\boldsymbol{W}}} \quad ||\boldsymbol{X}-\boldsymbol{D}^{(1)}\boldsymbol{A}^{(1)}||_F^2 + \lambda_1||P_1(\boldsymbol{A}^{(1)}) - \boldsymbol{D}^{(2)}\boldsymbol{A}^{(2)}||_F^2 + \\ &\lambda_2||\boldsymbol{Y}-\Phi(\boldsymbol{W}P_2(\boldsymbol{A}^{(2)}))||_F^2 + \lambda_3||\boldsymbol{A}^{(1)}||_F^2 + \lambda_4||\boldsymbol{W}||_F^2, \\ &s.t. \quad ||\boldsymbol{a}_i^{(2)}||_0 < T \ \ \forall i = 1,...,N_2, \\ &\boldsymbol{D}^{(1)}, \boldsymbol{D}^{(2)} \in \mathcal{C}, \end{aligned} \tag{2}$$

where $\boldsymbol{X}$ is a $p^2 \times N_1$ matrix. Each column of $\boldsymbol{X}$ is a vectorized $p \times p$ patch, and N_1 is the total number of patches over all the training images. The P_1 operator is concatenating n_1 adjacent patches of an image in to a single column. $\boldsymbol{D}^{(1)}$ and $\boldsymbol{D}^{(2)}$ are dictionaries with K_1 and K_2 atoms respectively. The P_2 operator is concatenating all patches of a single image in a single column. The classification is via "one-versus-all" scheme. Hence, $\boldsymbol{Y}$ is a $C \times N$ matrix with C being the number of classes and N being the total number of training images. All entries of column i of $\boldsymbol{Y}$ can be set to -1 with $+1$ in only row c. $\boldsymbol{W}$ as coefficient parameter matrix for the classification model needs to be jointly learned with $\boldsymbol{D}^{(1)}$, $\boldsymbol{A}^{(1)}$, $\boldsymbol{D}^{(2)}$, and $\boldsymbol{A}^{(2)}$. Φ is an activation function. In Fig. (1) we show the required sequence of computational steps with the resulting matrices from Eqn. (2) and corresponding structures.

Feature vectors from the first layer $\boldsymbol{a}_i^{(1)}$ are built from small image patches thus, making the dictionary atoms of $\boldsymbol{D}^{(1)}$ represent primitive characteristics of the images such as vertical, horizontal and diagonal lines. Adjacent representation vectors from the first layer are consolidated in order to build the second layer sparse representations $\boldsymbol{a}_i^{(2)}$. Thus, the dictionary atoms of $\boldsymbol{D}^{(2)}$ are expected to capture relations between neighbouring representations of the images. Building the second layer feature vectors using the first layer feature vectors of small image patches makes the overall representation robust to small differences of images within the same class. Particularly, the hierarchical dictionary learning method yields smaller distances between the feature vectors belonging to the same class.

2.2. Algorithm

As shown next, Algorithm (1) includes initialization and steps for solving the optimization problem in Eqn. (2). Lines 1-3 of the algorithm are finding initial values of the $\boldsymbol{D}^{(1)}$, $\boldsymbol{A}^{(1)}$, $\boldsymbol{D}^{(2)}$, $\boldsymbol{A}^{(2)}$, and $\boldsymbol{W}$ matrices. The optimization problems in the first and the second line may be solved by K-SVD and OMP methods, and the optimization problem in line 3 is a convex problem which can be solved by a gradient descent approach. Lines 5-9 are iteratively solving five relatively simple optimization problems for m iterations to reach the solution of Eqn. (2). m can be chosen based on the performance of the classifier on training data set and cross validation. Regarding the fact that the P_1 and P_2 operators are only reshaping matrices, they are invertible.

[1] In some applications L_0-norms are replaced in practice by L_0-norms.

Algorithm 1

Initialization:

1: $< \boldsymbol{D}_0^{(1)}, \boldsymbol{A}_0^{(1)} >= \underset{\boldsymbol{D}^{(1)},\boldsymbol{A}^{(1)}}{\arg\min} ||\boldsymbol{X} - \boldsymbol{D}^{(1)}\boldsymbol{A}^{(1)}||_F^2 + \lambda_3||\boldsymbol{A}^{(1)}||_F^2,$
$s.t: \quad \boldsymbol{D}^{(1)} \in \mathcal{C}.$

2: $< \boldsymbol{D}_0^{(2)}, \boldsymbol{A}_0^{(2)} >= \underset{\boldsymbol{D}^{(2)},\boldsymbol{A}^{(2)}}{\arg\min} ||P_1(\boldsymbol{A}^{(1)}) - \boldsymbol{D}^{(2)}\boldsymbol{A}^{(2)}||_F^2,$
$s.t: \quad \boldsymbol{D}^{(2)} \in \mathcal{C}, \quad ||\boldsymbol{A}_i^{(2)}||_0 < T \ \ \forall i = 1, .., N_2.$

3: $\boldsymbol{W}_0 = \underset{\boldsymbol{W}}{\arg\min} ||\boldsymbol{Y} - \Phi(\boldsymbol{W}P_2(\boldsymbol{A}^{(2)}))||_F^2 + \lambda_2||\boldsymbol{W}||_F^2.$

Solving by alternating method:

4: **for** $t = 0 : 1 : m$ **do**

5: $\boldsymbol{A}_{t+1}^{(1)} = \underset{\boldsymbol{A}^{(1)}}{\arg\min} ||\boldsymbol{X} - \boldsymbol{D}_t^{(1)}\boldsymbol{A}^{(1)}||_F^2 + \lambda_3||\boldsymbol{A}^{(1)}||_F^2 + \lambda_1||P_1(\boldsymbol{A}^{(1)}) - \boldsymbol{D}_t^{(2)}\boldsymbol{A}_t^{(2)}||_F^2.$

6: $\boldsymbol{D}_{t+1}^{(1)} = \underset{\boldsymbol{D}^{(1)}}{\arg\min} ||\boldsymbol{X} - \boldsymbol{D}^{(1)}\boldsymbol{A}_t^{(1)}||_F^2, \quad s.t: \quad \boldsymbol{D}^{(1)} \in \mathcal{C}.$

7: $\boldsymbol{A}_{t+1}^{(2)} = \underset{\boldsymbol{A}^{(2)}}{\arg\min} ||P_1(\boldsymbol{A}_t^{(1)}) - \boldsymbol{D}_t^{(2)}\boldsymbol{A}^{(2)}||_F^2 + \frac{\lambda_2}{\lambda_1}||\boldsymbol{Y} - \Phi(\boldsymbol{W}_t P_2(\boldsymbol{A}^{(2)}))||_F^2, \quad s.t: \ ||\boldsymbol{A}_i^{(2)}||_0 < T \ \forall i = 1, .., N_2.$

8: $\boldsymbol{D}_{t+1}^{(2)} = \underset{\boldsymbol{D}^{(2)}}{\arg\min} ||P_1(\boldsymbol{A}_t^{(1)}) - \boldsymbol{D}^{(2)}\boldsymbol{A}_t^{(2)}||_F^2, \ s.t: \ \boldsymbol{D}^{(2)} \in \mathcal{C}.$

9: $\boldsymbol{W}_t = \underset{\boldsymbol{W}}{\arg\min} ||\boldsymbol{Y} - \Phi(\boldsymbol{W}P_2(\boldsymbol{A}_t^{(2)}))||_F^2 + \frac{\lambda_4}{\lambda_2}||\boldsymbol{W}||_F^2.$

10: **end for**

2.3. Adaptivity by Gradient

The procedure of finding the gradient of a term, such as $||\boldsymbol{Y} - \Phi(\boldsymbol{W}_t P_2(\boldsymbol{A}^{(2)}))||_F^2$, on the right-hand side of equation at line 7 with respect to $\boldsymbol{A}^{(2)}$, are as follows; Let $\xi(\boldsymbol{A}) = ||\boldsymbol{Y} - \Phi(\boldsymbol{W}P_2(\boldsymbol{A}))||_F^2$, then $\xi(\boldsymbol{A})$ can be written as:

$$\xi(\boldsymbol{A}) = \sum_{j=1}^{N} ||\boldsymbol{y}_j - \Phi(\boldsymbol{W}P_2(\boldsymbol{a}_j))||_2^2 = \sum_{j=1}^{N} \boldsymbol{e}_j^T \boldsymbol{e}_j, \qquad (3)$$

where $\boldsymbol{Y} = [\boldsymbol{y}_1, .., \boldsymbol{y}_N]$, $\boldsymbol{A} = [\boldsymbol{a}_1, .., \boldsymbol{a}_{N_2}]$, and $\boldsymbol{e}_j = \boldsymbol{y}_j - \Phi(\boldsymbol{W}P_2(\boldsymbol{a}_j))$. Let $\boldsymbol{v}_j = WP_2(\boldsymbol{a}_j)$, and $\hat{\boldsymbol{y}}_j = \Phi(\boldsymbol{v}_j)$. We can then, write the gradient of $\xi(\boldsymbol{A})$ with respect to $\boldsymbol{A}$ as follows:

$$\begin{aligned} \frac{\partial \xi(\boldsymbol{A})}{\partial \boldsymbol{A}} &= P_2^{-1}(\frac{\partial \xi(\boldsymbol{A})}{\partial P_2}), \\ \frac{\partial \xi(\boldsymbol{A})}{\partial P_2} &= [\frac{\partial \xi(\boldsymbol{A})}{\partial P_{2,1}}, .., \frac{\partial \xi(\boldsymbol{A})}{\partial P_{2,N}}], \\ \frac{\partial \xi(\boldsymbol{A})}{\partial P_{2,j}} &= \frac{\partial \boldsymbol{v}_j}{\partial P_{2,j}} \frac{\partial \hat{\boldsymbol{y}}_j}{\partial \boldsymbol{v}_j} \frac{\partial \boldsymbol{e}_j}{\partial \hat{\boldsymbol{y}}_j} \frac{\partial \xi(\boldsymbol{A})}{\partial \boldsymbol{e}_j}, \\ \frac{\partial \xi(\boldsymbol{A})}{\partial P_{2,j}} &= \boldsymbol{W}^T \times diag(\Phi'(\boldsymbol{v}_j)) \times -1\boldsymbol{I} \times 2\boldsymbol{e}_j, \end{aligned} \qquad (4)$$

where for writing simplicity, we used $P_2(\boldsymbol{A}^{(2)})$ and $P_2(\boldsymbol{a}_j^{(2)})$ for P_2 and $P_{2,j}$ respectively. $diag(\Phi'(\boldsymbol{v}_j))$ is a diagonal matrix of vector $\Phi'(\boldsymbol{v}_j)$, and P_2^{-1} is the inverse function of P_2 which is reshaping a matrix to its original form.

Fig. 2: Subset of images in Extended YaleB dataset.

3. EXPERIMENTS

Our evaluation of the proposed methodology, was carried out on the Extended YaleB database [19], and STL-10 dataset [20].

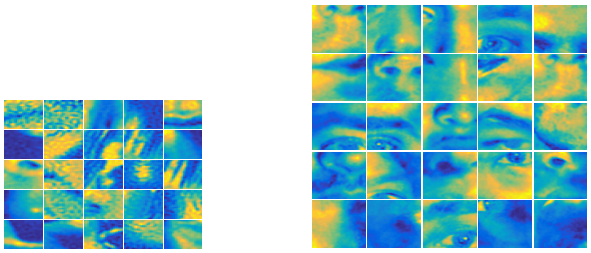
3.1. Extended YaleB Dataset

This database contains 2,414 face images from 38 individuals [19]. Each individual has about 64 images, and the size of each image is 192 × 168 pixels. Half of the images per individual are chosen randomly for training and the other half is used for testing. A subset of the images in this dataset are shown in Fig. (2). Because of varying illumination conditions and face expressions, this dataset is a challenging dataset for classification. In our approach, the images are partitioned into non-overlapping patches of size 24 × 24 pixels with 200 atoms in the first layer dictionary. Four adjacent patches are concatenated to learn the sparse representations as well as the dictionary in the second layer. The second layer dictionary has 1000 atoms, and the sparse representations have at most 300 non-zero entries. We compare the performance of our method with the state of some art methods in Table (1).

Table 1: Recognition results using random-faces features on the Extended YaleB dataset and comparing with Locality-constrained Linear Coding (LLC), Sparse Representation-based Classification (SRC), and Label Consistent K-SVD (LC-KSVD) methods

Method	Accuracy (%)
SRC [21]	97.2
LLC [22]	90.7
LC-KSVD [16]	96.7
Our approach	**98.5**

Even though the LC-KSVD [23] approach is learning discriminative dictionaries via joint classification and dictionary learning but, as maybe seen from Table (1), our approach registers the highest accuracy. This is due to using the hierarchical approach. On the first layer the elementary details of the images are learned, and the higher level characteristics of the images are learned via the second layer dictionary. Fig



(a) (b)

Fig. 3: Sample dictionary atoms from first and second layer dictionaries.

(3.a) shows a subset of dictionary atoms from the first layer dictionary. We can recognize a few basic characteristics of a face in this picture, and we expect the second layer dictionary to assemble these lower scale features and show higher scale features of faces. Fig (3.b) shows a subset of dictionary atoms from the second layer dictionary $(\boldsymbol{I} \otimes \boldsymbol{D}^{(1)})\boldsymbol{D}^{(2)}$. As expected, the atoms of the second layer dictionary are showing more distinct features of faces such as the shape of the noses, distance between the eyes and the eyebrows.

3.2. STL-10 Dataset

This database contains 10 image classes of airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck. The images are 96×96 pixels, color images. There are 100 training images and 800 test images per class. A subset of the images in this dataset are shown in Fig. (4).

Due to the small training set size, this dataset is recognized as a very challenging dataset for classification. In our approach, the images are partitioned in non-overlapping patches of size 12×12 pixels with 200 atoms in the first layer dictionary. Four adjacent patches are concatenated to learn the sparse representations as well as the dictionary in the second layer. The second layer dictionary has 1000 atoms and the sparse representations have at most 300 non-zero entries. We compare the performance of our method with the state of some art methods in Table (2).

As maybe seen from Table (2), our approach registers the highest accuracy. Similarly to the previous dataset, the basic structure of the images are learned in the first layer, and more complex characteristics of the images are learned in the second layer dictionary.

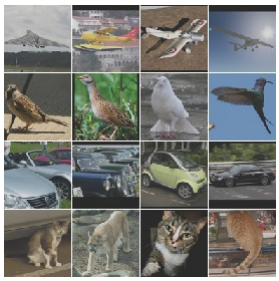


Fig. 4: Subset of images in STL-10 dataset.

Table 2: Classification results using features on the STL-10 dataset

Method	Accuracy (%)
K-SVD [11]	40.6
Coates [20]	51.5
LC-KSVD [23]	41.3
Our approach	**53.8**

4. CONCLUSIONS

In this paper, we have used two image datasets to evaluate the classification performance of the proposed hierarchical dictionary learning method. We have demonstrated the importance of representing images by learning image characteristics at multiple scales via hierarchical dictionaries. We have also shown that refining the dictionary learning and feature selection by accounting for the target task improves the performance of the algorithm. The evaluation results show the merit of the proposed method for classifying the images. The idea can be generalized to arbitrary number of layers in different datasets.

5. REFERENCES

[1] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[2] L. Shao, R. Yan, X. Li, and Y. Liu, "From heuristic optimization to dictionary learning: a review and comprehensive comparison of image denoising algorithms," *IEEE Transactions on Cybernetics*, vol. 44, no. 7, pp. 1001–1013, 2014.

[3] J. Dong, W. Wang, and J. Chambers, "Removing speckle noise by analysis dictionary learning," in *Sensor Signal Processing for Defence (SSPD), 2015*. IEEE, 2015, pp. 1–5.

[4] M. Xu, X. Jia, M. Pickering, and A. J. Plaza, "Cloud removal based on sparse representation via multitemporal dictionary learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 5, pp. 2998–3006, 2016.

[5] W. Dong, G. Shi, Y. Ma, and X. Li, "Image restoration via simultaneous sparse coding: Where

structured sparsity meets gaussian scale mixture," *International Journal of Computer Vision*, vol. 114, no. 2, pp. 217–232, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11263-015-0808-y

[6] S. Xiang, G. Meng, Y. Wang, C. Pan, and C. Zhang, "Image deblurring with coupled dictionary learning," *International Journal of Computer Vision*, vol. 114, no. 2, pp. 248–271, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11263-014-0755-z

[7] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural computation*, vol. 13, no. 4, pp. 863–882, 2001.

[8] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-invariance sparse coding for audio classification," *arXiv preprint arXiv:1206.5241*, 2012.

[9] L. Shen, G. Sun, Q. Huang, S. Wang, Z. Lin, and E. Wu, "Multi-level discriminative dictionary learning with application to large scale image classification," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3109–3123, 2015.

[10] D. Zhang, P. Liu, K. Zhang, H. Zhang, Q. Wang, and X. Jing, "Class relatedness oriented-discriminative dictionary learning for multiclass image classification," *Pattern Recognition*, vol. 59, pp. 168 – 175, 2016, compositional Models and Structured Learning for Visual Recognition.

[11] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[12] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *International Journal of control*, vol. 50, no. 5, pp. 1873–1896, 1989.

[13] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[14] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 689–696.

[15] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, 2012.

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

[17] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[18] Y. Suo, M. Dao, U. Srinivas, V. Monga, and T. D. Tran, "Structured dictionary learning for classification," *arXiv preprint arXiv:1406.1943*, 2014.

[19] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[20] A. Coates, H. Lee, and A. Y. Ng, "An analysis of single-layer networks in unsupervised feature learning," *Ann Arbor*, vol. 1001, no. 48109, p. 2, 2010.

[21] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[22] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

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