LEARNING DISCRIMINATIVE DICTIONARIES WITH PARTIALLY LABELED DATA

Ashish Shrivastava, Jaishanker K. Pillai, Vishal M. Patel, and Rama Chellappa

Department of Electrical and Computer Engineering
Center for Automation Research, University of Maryland, College Park, MD
{ashish,jsp, pvishalm,rama}@umiacs.umd.edu

ABSTRACT

While recent techniques for discriminative dictionary learning have demonstrated tremendous success in image analysis applications, their performance is often limited by the amount of labeled data available for training. Even though labeling images is difficult, it is relatively easy to collect unlabeled images either by querying the web or from public datasets. In this paper, we propose a discriminative dictionary learning technique which utilizes both labeled and unlabeled data for learning dictionaries. Extensive evaluation on existing datasets demonstrate that the proposed method performs significantly better than state of the art dictionary learning approaches when unlabeled images are available for training.

***Index Terms*—** Semi-supervised dictionary learning, latent variables, classification.

1. INTRODUCTION

Sparse and redundant signal representations have recently gained much interest in image understanding [1]. This is partly due to the fact that signals or images of interest are often sparse in some dictionary. These dictionaries can be either analytic or they can be learned directly from the data. In fact, it has been observed that learning a dictionary directly from data often leads to improved results in many practical applications such as classification and restoration [1].

While these dictionaries are often trained to obtain good reconstruction, training supervised dictionaries with a specific discriminative criterion has also been considered. For instance, linear discriminant analysis (LDA) based basis selection and feature extraction algorithm for classification using wavelet packets was proposed by Etemand and Chellappa [2] in the late nineties. Recently, similar algorithms for simultaneous sparse signal representation and discrimination have also been proposed. See [1], [3] and the references therein for more details.

Dictionary learning methods for unsupervised learning have also been proposed. In [4], a method for simultaneously learning a set of dictionaries that optimally represent each cluster is proposed. To improve the accuracy of sparse coding, this approach was later extended by adding a block incoherence term in their optimization problem [5]. Some of the other sparsity motivated subspace clustering methods include [6], [7].

The performance of a supervised classification algorithm is often dependent on the quality and diversity of training images, which are mainly hand-labeled. However, labeling images is expensive and time consuming due to the significant human effort involved. On the other hand, one can easily obtain large amounts of unlabeled images from public image datasets like Flickr or by querying image search engines like Bing. This has motivated researchers to develop semi-supervised algorithms, which utilize both labeled and unlabeled data for learning classifier models. Such methods have demonstrated improved performance when the amount of labeled data is limited. See [8] for an excellent survey of recent efforts on semi-supervised learning.

This work was partially supported by an ONR grant N00014-12-1-0124.

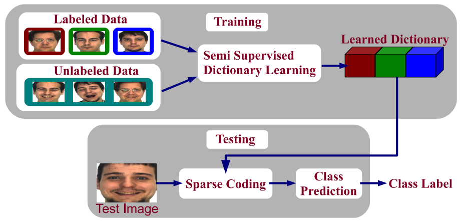


Fig. 1. Block diagram of the proposed Semi-Supervised Dictionary Learning method.

Two of the most popular methods for semi-supervised learning are Co-Training [9] and Semi-Supervised Support Vector Machines (S3VM) [10]. Co-Training assumes the presence of multiple views for each feature and uses the confident samples in one view to update the other. However, in applications such as image classification, one often has just a single feature vector and hence it is difficult to apply Co-Training. S3VM considers the labels of the unlabeled data as additional unknowns and jointly optimizes over the classifier parameters and the unknown labels in the SVM framework [11].

In this paper, we propose a novel method to learn discriminative dictionaries for classification in a semi-supervised manner. Fig. 1 shows the block diagram of the proposed approach which uses both labeled and unlabeled data. While learning a dictionary, we maintain a probability distribution over class labels for each unlabeled data. The discriminative part of the cost is made proportional to the confidence over the assigned label of the participating training sample. This makes the proposed method robust to label assignment errors.

2. PROBLEM FORMULATION

Let $\mathcal{L} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, N_l\}$ be the set of labeled data and $\mathcal{U} = \{\mathbf{x}_i, i = N_l + 1, \ldots, N_l + N_u\}$ be the set of unlabeled data available for learning dictionaries, where N_l and N_u are the number of labeled and unlabeled samples, respectively. Here, $\mathbf{x}_i \in \mathbb{R}^d$ denotes the d dimensional feature vector extracted from the i^{th} sample, $y_i \in \{1, 2, \ldots, K\}$ is the corresponding class label, K is the number of classes and $N = N_u + N_l$ is the total number of available training samples. Let $\mathbf{D}_c \in \mathbb{R}^{d \times n_c}$ be the dictionary with n_c atoms corresponding to class c and $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_K]$ be the concatenation of dictionaries from all the classes. Hence, $\mathbf{D} \in \mathbb{R}^{d \times n}$, where $n = \sum_c n_c$, denotes the total number of atoms in $\mathbf{D}$. We use $\boldsymbol{\alpha}_i \in \mathbb{R}^n$ to denote the sparse coefficients to linearly combine dictionary atoms in order to represent feature vector $\mathbf{x}_i$. Note that $\boldsymbol{\alpha}_i$ can be rewritten as $\boldsymbol{\alpha}_i^T = \left[\boldsymbol{\alpha}_{i1}^T | \boldsymbol{\alpha}_{i2}^T | \dots | \boldsymbol{\alpha}_{iK}^T\right]$, where $\boldsymbol{\alpha}_{ic}$ is the component of the coefficients belonging to the c^{th} class and $(.)^T$ denotes the transposition operation. Finally, let $\boldsymbol{\Lambda} = [\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_N]$ denote the matrix of coefficients corresponding to all the training samples.

We are interested in developing discriminative techniques for dictionary learning using both labeled data $\mathcal{L}$ and unlabeled data $\mathcal{U}$. This can be done by learning coarse initial dictionaries using the labeled data alone and roughly inferring the membership of unlabeled samples based on how well they are represented by dictionaries of different classes. Furthermore, confident unlabeled members could be used to further refine the dictionaries. To this end, for each feature vector $\mathbf{x}_i$ and for each class j, we define a latent variable P_{ij}, which represents the confidence of $\mathbf{x}_i$ belonging to the j^{th} class. Hence, by definition

$$\begin{aligned} P_{ij} = 1 \;\; &\text{if } \mathbf{x}_i \text{ is labeled and } j = y_i. \\ P_{ij} = 0 \;\; &\text{if } \mathbf{x}_i \text{ is labeled and } j \neq y_i. \\ 0 \le P_{ij} \le 1 \;\; &\text{if } \mathbf{x}_i \text{ is unlabeled.} \end{aligned} \tag{1}$$

To compactly represent the confidence values of all the training samples, we introduce a confidence matrix $\mathbf{P} \in \mathbb{R}^{N \times K}$ whose $(i,j)^{\text{th}}$ entry corresponds to the latent variable P_{ij} defined above.

We now formalize the constraints to be satisfied by the learned dictionaries. Intuitively, we want the learned dictionary to represent the data samples well. In other words, for each sample $\mathbf{x}_i$, the representation error given by $\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2$ should be small, where $\|.\|_2$ is the L_2 norm. Furthermore, if a data sample $\mathbf{x}_i$ belongs the the c^{th} class, it should be well represented by the corresponding dictionary, $\mathbf{D}_c$. Also to add the discrimination capability to the learned dictionary, we enforce that the feature vectors should be poorly represented by wrong classes. Hence, if a data sample $\mathbf{x}_i$ belongs to the c^{th} class, then $\sum_{j \neq c} \mathbf{D}_j \boldsymbol{\alpha}_{ij}$ should be small $\forall i, c$ [12].

One can easily enforce the class confidence into these constraints as follows. If a data sample has high confidence of belonging to a particular class, then it should be well represented by the corresponding dictionary. Similarly, if a sample has low confidence of belonging to a particular class, it should be poorly represented by that class. Thus, the discriminative fidelity cost J_i of each data sample $\mathbf{x}_i$ is given by

$$\begin{aligned} J_i(\mathbf{D}, \mathbf{P}_i, \boldsymbol{\alpha}_i) = \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \sum_{c=1}^{K} \|(\mathbf{x}_i - \mathbf{D}_c \boldsymbol{\alpha}_{ic}) P_{ic}\|_2^2 \\ + \sum_{c=1}^{K} \|(\mathbf{D}_c \boldsymbol{\alpha}_{ic})(1 - P_{ic})\|_2^2, \end{aligned}$$

where $\mathbf{P}_i$ is the i^{th} row of the confidence matrix $\mathbf{P}$ which represents the confidence of i^{th} sample over all the classes.

In order to further enhance the discriminability of a learned dictionary, we introduce a discriminative cost on the coefficients $\boldsymbol{\Lambda}$ [2], [12]. Define the within-class scatter matrix as $S_W(\boldsymbol{\Lambda}) = \sum_{c=1}^{K} \sum_{\{i: y_i = c\}} (\boldsymbol{\alpha}_i - \mathbf{m}_c)(\boldsymbol{\alpha}_i - \mathbf{m}_c)^T$ and the between-class scatter matrix as $S_B(\boldsymbol{\Lambda}) = \sum_{c=1}^{K} (\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^T$, where $\mathbf{m}_c$ is the mean for the coefficients corresponding to the c^{th} class training sample and $\mathbf{m}$ is the mean of all the coefficients.

We achieve good separability for classification by having large between-class scatter and small within-class scatter simultaneously. This is done by introducing the following cost [12] $F(\boldsymbol{\Lambda}) = tr(S_W(\boldsymbol{\Lambda})) - tr(S_B(\boldsymbol{\Lambda}))$, where $tr(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$. Finally, with the sparsity constraint on the coefficients, we propose the following optimization problem for semi-supervised discriminative dictionary learning

$$\mathbf{D}^*, \boldsymbol{\Lambda}^*, \mathbf{P}^* = \arg \min_{\mathbf{D}, \mathbf{P}, \boldsymbol{\Lambda}} \left\{ \sum_{i=1}^{N} (J_i + \gamma \|\boldsymbol{\alpha}_i\|_1) + \lambda F \right\}, \tag{2}$$

where λ controls the discrimination capability of the coefficients of training samples of different classes, and γ controls the sparsity of the coefficients.

3. SEMI-SUPERVISED DISCRIMINATIVE DICTIONARY (S2D2) LEARNING

Since the above objective function in (2) is jointly non-convex in $\mathbf{D}, \mathbf{P}, \boldsymbol{\Lambda}$, we solve it using an iterative alternating algorithm. At each iteration, three steps are performed namely coefficient update, dictionary update and confidence update. In the following, we use the notation $\mathbf{D}^{(t)}, \mathbf{P}^{(t)}$ and $\boldsymbol{\Lambda}^{(t)} = [\boldsymbol{\alpha}_1^{(t)}, \dots \boldsymbol{\alpha}_N^{(t)}]$ to denote the dictionary, confidence and coefficient matrices, respectively, in the t^{th} iteration.

Coefficient Update: In this step, we fix confidence matrix $\mathbf{P}^{(t)}$ and dictionary $\mathbf{D}^{(t)}$ at the t^{th} iteration and update the sparse coefficients of each data sample. This is equivalent to solving the following convex optimization problem

$$\boldsymbol{\Lambda}^{(t+1)} = \arg \min_{\boldsymbol{\Lambda}} \left\{ \sum_{i=1}^{N} \left(J_i + \gamma \|\boldsymbol{\alpha}_i^{(t)}\|_1 \right) + \lambda F \right\}, \tag{3}$$

which can be solved by using the Iterative Projection Method (IPM) proposed in [13].

Dictionary Update: We now update the dictionaries by fixing $\boldsymbol{\Lambda}^{(t+1)}$ and $\mathbf{P}^{(t)}$, and solve for $\mathbf{D}_c^{(t+1)}$ for each class. The optimization problem in (2) becomes

$$\mathbf{D}_c^{(t+1)} = \arg \min_{\mathbf{D}_c} \sum_{i=1}^{N} J_i \left(\mathbf{D}^{(t)}, \mathbf{P}^{(t)}, \boldsymbol{\Lambda}^{(t+1)} \right). \tag{4}$$

In addition to solving (4), we require that the dictionary atoms are of unit norm. Eq. (4) is essentially a quadratic programming problem which can be solved for one dictionary atom at a time using the approach proposed in [14].

Confidence Update: Keeping the updated dictionary $\mathbf{D}^{(t+1)}$ and the coefficient matrix $\boldsymbol{\Lambda}^{(t+1)}$ fixed, we propose the following approach for updating the confidence matrix. The confidence for each labeled sample is fixed based on (1) and is not updated. To update the confidence of unlabeled data, we first make the observation that a sample $\mathbf{x}_i$ which is well represented by the dictionary of class c, should have high confidence belonging to that class and vice versa. In other words, the confidence of a sample belonging to a particular class should be inversely proportional to the reconstruction error of the sample using the dictionary of the class.

For the unlabeled data $\mathbf{x}_i$ with coefficients $\boldsymbol{\alpha}_i = [\boldsymbol{\alpha}_{i1}^T, \dots, \boldsymbol{\alpha}_{iK}^T]^T$, we first compute the reconstruction error e_{ij} for each of the j classes as $e_{ij}^{(t)} = \|\mathbf{x}_i - \mathbf{D}_j^{(t)} \boldsymbol{\alpha}_{ij}^{(t)}\|_2^2$. At $(t+1)^{\text{th}}$ iteration, we update the confidence of a training sample $\mathbf{x}_i$ to belong to the j^{th} class, denoted by $P_{ij}^{(t+1)}$ as

$$P_{ij}^{(t+1)} = \begin{cases} \dfrac{\exp^{-\frac{e_{ij}^{(t)}}{\sigma^2}}}{\sum_{c=1}^{K} \exp^{-\frac{e_{ic}^{(t)}}{\sigma^2}}} & \text{if } \dfrac{\exp^{-\frac{e_{ij}^{(t)}}{\sigma^2}}}{\sum_{c=1}^{K} \exp^{-\frac{e_{ic}^{(t)}}{\sigma^2}}} > \theta \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where the negative exponent captures the inverse relationship between the confidence and the reconstruction error. The term in the denominator normalizes the value from 0 to 1 and the threshold θ controls the confidence level. Finally, we do not update the probability of all the samples, hence, we use a parameter which decides how many samples per class can be updated in confidence matrix $\mathbf{P}$ in one iteration. The entire procedure for S2D2 is summarized in Algorithm 1.

Classification: In this section, we explain how the class label of a test sample is predicted using the learned dictionaries. For a given test sample $\mathbf{x}_t$, we compute the corresponding sparse coefficients $\boldsymbol{\alpha}_t$ by solving the following ℓ_1 minimization problem: $\boldsymbol{\alpha}_t = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \quad s.t \quad \mathbf{x}_t = \mathbf{D}\boldsymbol{\alpha}$. To predict the class label of a test sample $\mathbf{x}_t$, we first compute the error ϵ_{ti} in its representation as: $\epsilon_{ti} = \|\mathbf{x}_t - \mathbf{D}\delta_i(\boldsymbol{\alpha}_t)\|_2^2 + 0.5\|\boldsymbol{\alpha}_t - \mathbf{m}_i\|_2^2$, where $\mathbf{m}_i$ is the mean coefficients for the i^{th} class and $\delta_i(.)$ is a characteristic function that selects coefficients corresponding to the i^{th} class. The predicted class label c_t of the test sample $\mathbf{x}_t$ is the one that minimizes this representation error $c_t = \arg\min_{i \in \{i1,2,\ldots,K\}} \epsilon_{ti}$.

Algorithm 1: Algorithm for Learning Semi-Supervised Discriminative Dictionaries (S2D2)

Input: Labeled Data $\mathcal{L} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, N_l\}$,
Unlabeled Data $\mathcal{U} = \{\mathbf{x}_i, i = N_l + 1, \ldots, N_l + N_u\}$
Output: Dictionary $\mathbf{D}^* = [\mathbf{D}_1^* | \mathbf{D}_2^* | \ldots \mathbf{D}_K^*]$
Algorithm:
1. Initialization:
   a. Initialize the dictionary $\mathbf{D}_c^{(0)}, \forall c$ and coefficients $\boldsymbol{\Lambda}^{(0)}$ using labeled training data.
   b. Initialize the confidence variable for each training sample using Eq. (1).
2. Repeat for fixed number of iteration (or until convergence):
   a. **Update the coefficients** $\boldsymbol{\Lambda}^{(t+1)}$ based on Eq. (3).
   b. **Update the dictionary** $\mathbf{D}_c^{(t+1)}$ for each class c using Eq. (4).
   c. **Update the confidence matrix** $\mathbf{P}$ - For each unlabeled sample $\mathbf{x}_i, i = N_l + 1, \ldots, N_l + N_u$, update $\mathbf{P}_{ij}$ using Eq. (5).
3. Return $\mathbf{D}^* = \mathbf{D}^{(\mathrm{T_c})}$, where $\mathrm{T_c}$ is the iteration number at which the learning algorithm converges.

4. EXPERIMENTAL RESULTS

To illustrate the effectiveness of our method, we present experimental results on some of the publicly available databases such as the USPS digit dataset [15], the AR face dataset [16] and the Kimia's object dataset [17]. A comparison with other existing object recognition methods in [12] suggests that the discriminative dictionary learning algorithm known as Fisher Discriminant Dictionary Learning (FDDL) is among the best. Hence, we treat it as state-of-the-art and use it as a bench mark for comparisons. We also compare our method with that of Support Vector Machines (SVM) as well as a semi-supervised extension of SVM known as (S3VM) [10]. In all of our experiments, the parameter θ is set equal to 0.7 to avoid using low confidence samples for dictionary update. For face recognition and object recognition λ and γ are to 0.05 and 0.005, respectively. However for digit recognition we observe best result at $\lambda = 0.001$ and $\gamma = 0.6$. The number of iterations are set to a maximum value of 15.

Datasets	SVM	S3VM	FDDL	S2D2
USPS Digit	74.47	75.61	79.24	**85.61**
AR Face	68.24	77.54	74.25	**85.98**
shapes216	84.26	84.26	86.11	**87.96**

Table 1. Recognition accuracy for the proposed method, compared to competing ones for the applications of digit and face recognition.

Digit Recognition: The USPS digit dataset [15] consists of binary images of hand written digits from 0 to 9. This dataset contains 7291 training samples and 2007 test samples. From the training data, four samples are randomly chosen as the labeled samples and the rest of the training data is used as the unlabeled data. The original images are of size 16×16 which forms the feature vector of dimension 256. We add a maximum of 50 unlabeled samples per class at each iterations.

The recognition accuracies are shown in the first row of Table 1. Observe that the proposed method outperforms the other methods by more than 6%. The improvement in performance compared to SVM and FDDL is due to the fact that we utilize the unlabeled data for updating dictionaries in the training stage. Being supervised techniques, the performance of SVM and FDDL reduces when the available labeled samples are small. Unlike S3VM which assigns hard labels to the unlabeled data points at each iteration, the proposed method assigns only a soft probability of class for each unlabeled data.The reason why S2D2 performs better than S3VM is probably because the soft assignment approach is more robust to labeling errors when compared to hard assignment.

In Fig. 2, we show a labeled image from each class on the left and the most confident unlabeled samples on the right. Note these unlabeled samples are the ones which contribute the most to the dictionary update. As can be observed, the most confident unlabeled samples belong to the correct class. Furthermore, there is a significant variation between the labeled and the highest confident unlabeled samples for each class. This demonstrates that the proposed method increases the diversity of the original training set, which also explains the improved performance obtained by our method.

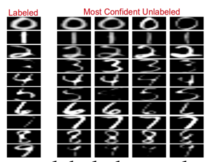
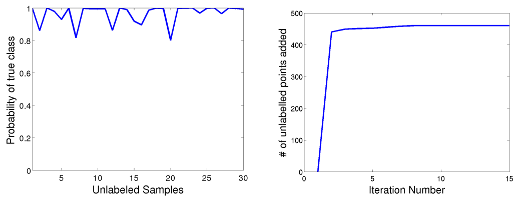


Fig. 2. Most confident unlabeled samples on the USPS dataset. Highest confident unlabeled samples belong to the correct class and also add diversity to the training set.

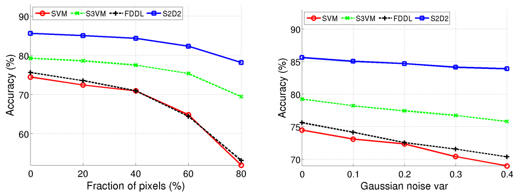
Illustration of soft label update of P: To further analyze the performance of the proposed method on the USPS digit dataset, we illustrate the change in $\mathbf{P}$ matrix over successive iterations. Ideally, one would like to use each unlabeled data to update its true class dictionary with confidence 1 and not to update the other dictionaries. In order to study how close the proposed method gets to the ideal condition, we plot the confidence of true class of randomly selected unlabeled data, i.e. the updated probability corresponding to the true class in Fig. 3(a). As explained previously, we add more unlabeled samples at each iterations, so, we also illustrate how this number increases over iterations in Fig. 3(b).



(a) (b)

Fig. 3. Analysis of confidence matrix. (a) Probability of true class. (b) Points added over iterations.

Performance in the presence of missing and noisy pixels: To further evaluate the robustness of the proposed method, we compute the recognition performance of the proposed method when pixels in the image are either missing or corrupted by noise. In the missing data experiment, we set pixels at random locations to zero for test images in the digit recognition application. The number of corrupted pixels is varied and we plot the corresponding accuracy in Fig. 4(a). Note that the recognition accuracy falls as expected when the amount of missing pixels is increased. But the fall in accuracy is much lower for the proposed technique when compared to other methods. This clearly demonstrates the improved robustness of the proposed method compared to competing methods. Similarly to study the robustness of our method in the presence of noise, we add independent and identically distributed Gaussian noise to the pixels. We vary the variance of the added noise and compute the recognition accuracy for all the methods. The results are shown in Fig. 4(b). We observed a similar improvement in robustness of the proposed technique.



(a) (b)

Fig. 4. Accuracy for two kinds of corruption for digit recognition. (a) Accuracy vs Missing Data. (b) Accuracy vs noise variance.

Face Recognition: For face recognition, we use the AR face dataset [16], which contains over 4000 faces of 126 people captured under varying conditions of illumination, facial expressions and occlusion. We choose fifty females and fifty males for our experiments. We perform a Principal Component Analysis (PCA) based dimensionality reduction on the intensity images to obtain 300 dimensional PCA feature vectors. Each class has seven training samples and seven test samples. Out of the seven training samples, we randomly choose two to form the labeled data and use the remaining five as the unlabeled data. We add a maximum of four unlabeled data during the dictionary learning stage, since the available unlabeled data per class is limited for the dataset. We present the recognition accuracies in the second row of Table 1. The proposed method performs significantly better than the competing techniques.

Object Recognition: In the final set of experiments, we use Kimia's object dataset [17] which has 18 object categories each with 12 binary shapes. We randomly chose six images per class for training and the remaining six for testing. Furthermore, we randomly picked four images per class as the labeled data and the remaining two as the unlabeled data. Each image is resized to 16×16 and intensity values are used as features. The classification rates for all the algorithms are compared in Table 1. We see that the proposed method performs better than the other methods. These results clearly demonstrate that the performance of discriminative dictionary learning methods can be improved significantly by using unlabeled data, when the available labeled data is limited.

5. CONCLUSION

We proposed a method that utilizes unlabeled training data for learning discriminative dictionaries. The proposed method iteratively estimates the confidence of unlabeled samples belonging to each of the classes and uses it to refine the learned dictionaries. Experiments using various publicly available datasets demonstrate the improved accuracy and robustness to noise and missing information of the proposed method compared to state-of-the-art dictionary learning techniques.

6. REFERENCES

[1] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031 –1044, june 2010.

[2] K. Etemand and R. Chellappa, "Separability-based multiscale basis selection and feature extraction for signal and image classification," *IEEE Transactions on Image Processing*, vol. 7, no. 10, pp. 1453–1465, Oct. 1998.

[3] V. M. Patel and R. Chellappa, "Sparse representations, compressive sensing and dictionaries for pattern recognition," in *Asian Conference on Pattern Recognition (ACPR)*, 2011.

[4] P. Sprechmann and G. Sapiro, "Dictionary learning and sparse coding for unsupervised clustering," in *IEEE ICASSP*, 2010.

[5] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *IEEE CVPR*, 2010.

[6] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52 –68, march 2011.

[7] Y. C. Chen, C. S. Sastry, V. M. Patel, P. J. Phillips, and R. Chellappa, "Rotation invarient simultaneous clustering and dictionary learning," in *IEEE ICASSP*, 2012.

[8] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-supervised learning*, Adaptive computation and machine learning. MIT Press, 2006.

[9] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *ACM COLT*, 1998.

[10] V. Sindhwani and S. S. Keerthi, "Large scale semi-supervised linear svms," in *ACM SIGIR*, 2006.

[11] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[12] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *IEEE ICCV*, 2011.

[13] L. Rosasco, A. Verri, M. Santoro, S. Mosci, and S. Villa, "Iterative projection methods for structured sparsity regularization," *MIT-CSAIL-TR-2009-050, CBCL-282*, 2009.

[14] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *IEEE ICIP*, 2010.

[15] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, pp. 550–554, May 1994.

[16] A.M. Martinez and R. Benavente, "The AR face database," *CVC Tech. Report No. 24,*, 1998.

[17] T. B. Sebastian, P. N. Klein, and B. B. Kimia, "Recognition of shapes by editing shock graphs," in *IEEE ICCV*, 2001.