Dictionaries for Image-based Recognition

Vishal M. Patel
UMIACS
University of Maryland
College Park, MD
Email: pvishalm@umiacs.umd.edu

Qiang Qiu
UMIACS
University of Maryland
College Park, MD
Email: qiu@umiacs.umd.edu

Rama Chellappa
UMIACS
University of Maryland
College Park, MD
Email: rama@umiacs.umd.edu

***Abstract*—In recent years, Sparse Representation (SR) and Dictionary Learning (DL) have emerged as powerful tools for efficiently processing of image and video data in non-traditional ways. An area of promise for these theories is object recognition. In this paper, we present an overview of SR and DR and examine several interesting object recognition approaches using these theories. We will also explore the use of non-linear kernel SR as well as DL methods in many computer vision problems including object recognition, multimodal biometrics recognition, and domain adaptation.**

I. Introduction

In recent years, the field of sparse representation and dictionary learning has undergone rapid development, both in theory and in algorithms [1], [2], [3], [4]. It has also been successfully applied to numerous image understanding applications [2], [3]. This is partly due to the fact that signals or images of interest, though high dimensional, can often be coded using few representative atoms in some dictionary. Given a redundant dictionary $\mathbf{D}$ and a signal $\mathbf{y}$, finding the sparsest representation of $\mathbf{y}$ in $\mathbf{D}$ entails solving the following optimization problem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \text{ subject to } \mathbf{y} = \mathbf{D}\mathbf{x}, \qquad (1)$$

where the $\|\mathbf{x}\|_0 := |\#\{i : x_i \neq 0\}|$, which is a count for the number of nonzero elements in $\mathbf{x}$. Problem (1) is NP-hard and cannot be solved in a polynomial time. Hence, approximate solutions are usually sought [3], [5], [6], [7]. For instance, Basis Pursuit [5] offers the solution via l_1 minimization as

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ subject to } \mathbf{y} = \mathbf{D}\mathbf{x}, \qquad (2)$$

where $\|\mathbf{x}\|_1 = \sum_i |x_i|$. The sparsest recovery is possible provided that certain conditions are met [8]. One can adapt the above framework to noisy setting, where the measurements are contaminated with an error η obeying $\|\boldsymbol{\eta}\|_2 < \epsilon$, that is

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \boldsymbol{\eta} \quad \text{for} \quad \|\boldsymbol{\eta}\|_2 < \epsilon. \qquad (3)$$

A stable solution can be obtained by solving the following optimization problem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ subject to } \|\mathbf{y} - \mathbf{D}\mathbf{x}\| < \epsilon. \qquad (4)$$

The dictionary $\mathbf{D}$ can be either based on a mathematical model of the data [3] or it can be trained directly from the data [9]. It has been observed that learning a dictionary directly from training rather than using a predetermined dictionary (such as wavelet or Gabor) usually leads to better representation and hence can provide improved results in many practical applications such as restoration and classification [1], [2], [10], [11], [12], [13], [14], [15]. Designing dictionaries based on training is a much recent approach to dictionary learning which is strongly motivated by recent advances in the SR theory.

In dictionary learning methods, given a set of examples $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_m]$, the objective is to find a dictionary that provides the best representation for each examples in this set. One can obtain this by solving the following optimization problem

$$(\hat{\mathbf{D}}, \hat{\mathbf{X}}) = \arg\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \text{ subject to } \forall i \; \|\mathbf{x}_i\|_0 \leq T_0$$

where $\mathbf{x}_i$ represents a column of $\mathbf{X}$ and T_0 is a sparsity parameter. Here, $\|\mathbf{A}\|_F$ denotes the Frobenius norm defined as $\|\mathbf{A}\|_F = \sqrt{\sum_{ij} |A_{ij}|^2}$. Several algorithms have been published in the literature that solve the above problem [1]. Two of the well-known algorithms for finding such dictionary are the method of optimal directions (MOD)[16] and the K-SVD [17] algorithm. Both MOD and K-SVD are iterative methods and they alternate between sparse-coding and dictionary update steps. See [17] and [16] for more details.

While dictionaries are often trained to obtain good reconstruction, training dictionaries with a specific discriminative criteria has also been considered. In what follows, we present several applications of SR and DL in object recognition, domain adaptation, and multimodal biometrics recognition.

II. Sparse Representation-based Classification

Sparse representation-based classification (SRC) [18] was one of the first methods that showed the effectiveness of SR for face recognition. The idea proposed in [18] is to create a dictionary matrix of the training samples as column vectors. The test sample is also represented as a column vector. Different dimensionality reduction methods are used to reduce the dimension of both the test vector and the vectors in the dictionary. One such approach for dimensionality reduction is random projections [19]. Random projections, using a generated sensing matrix, are taken of both the dictionary matrix and the test sample. It is then simply a matter of solving an ℓ_1 minimization problem in order to obtain the sparse solution. Once the sparse solution is obtained, it can provide information as to which training sample the test vector most closely relates to.

Let each image be represented as a vector in $\mathbb{R}^n$, $\mathbf{D}$ be the dictionary (i.e. training set) and $\mathbf{y}$ be the test image. The SRC algorithm is as follows:

1) Create a matrix of training samples $\mathbf{D} = [\mathbf{D}_1, ..., \mathbf{D}_k]$ for k classes, where $\mathbf{D}_i$ are the set of images of each class.
2) Reduce the dimension of the training images and a test image by any dimensionality reduction method. Denote the resulting dictionary and the test vector as $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{y}}$, respectively.
3) Normalize the columns of $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{y}}$.
4) Solve the following ℓ_1 minimization problem
$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}'} \parallel \mathbf{x}' \parallel_1 \quad \text{subject to} \quad \tilde{\mathbf{y}} = \tilde{\mathbf{D}}\mathbf{x}', \qquad (5)$$
5) Calculate the residuals
$$r_i(\tilde{\mathbf{y}}) = \|\tilde{\mathbf{y}} - \tilde{\mathbf{D}}\delta_i(\hat{\mathbf{x}})\|_2,$$
for $i = 1, ..., k$ where δ_i a characteristic function that selects the coefficients associated with the i^{th} class.
6) Identify($\mathbf{y}$)=$\arg\min_i r_i(\tilde{\mathbf{y}})$.

The assumption made in this method is that given sufficient training samples of the k^{th} class, $\tilde{\mathbf{D}}_k$, any new test image y that belongs to the same class will approximately lie in the linear span of the training samples from the class k. This implies that most of the coefficients not associated with class k in $\hat{\mathbf{x}}$ will be close to zero. Hence, α' is a sparse vector. This algorithm can also be extended to deal with occlusions and random noise. Furthermore, a method of rejecting invalid test samples can also be introduced within this framework [18]. In particular, to decide whether a given test sample is a valid sample or not, the notion of Sparsity Concentration Index (SCI) has been proposed in [18].

One of the main difficulties in iris biometric is that iris images acquired from a partially cooperating subject often suffer from blur, occlusion due to eyelids, and specular reflections. As a result, the performance of existing iris recognition systems degrade significantly on these images. Hence, it is essential to select good images before they are input to the recognition algorithm. To this end, the SRC framework was extended for cancelable iris biometric in [20], [21] that can select and recognize iris images in a single step.

In Figure 1, we display the iris images having the least SCI value for the blur, occlusion and segmentation error experiments performed on the real iris images in the University of Notre Dame ND dataset. As it can be observed, the low SCI images suffer from high amounts of distortion. The recognition performance of the SR based method for iris biometric [20] is summarized in Table I. As it can be seen from the table SRC provides the best recognition performance over that of NN and Libor Masek's iris identification source code.

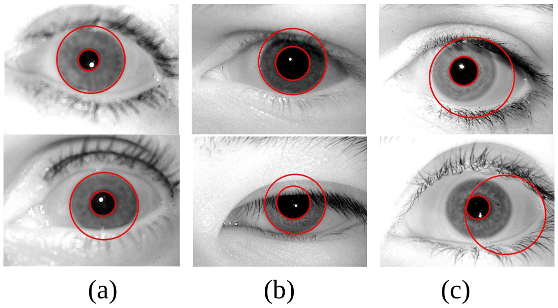

(a) (b) (c)

Fig. 1. Iris images with low SCI values in the ND dataset. Note that the images in (a), (b) and (c) suffer from high amounts of blur, occlusion and segmentation errors, respectively .

TABLE I
RECOGNITION RATE ON ND DATASET [20].

Image Quality	NN	Masek's Implementation	SRC
Good	98.33	97.5	99.17
Blured	95.42	96.01	96.28
Occluded	85.03	89.54	90.30
Seg. Error	78.57	82.09	91.36

A. Multimodal Multivariate Sparse Representation

The ideas presented in the above section can be extended to the case of multimodal multivariate sparse representation which is covered in this section. For simplicity, we present the multivariate sparse representation framework in terms of multimodal biometrics recognition [22], however, it can be used for any multimodal or multichannel classification problem [23]. An overview of the algorithm presented in [22] is shown in Fig. 2.

Consider a multimodal C-class classification problem with D different biometric traits. Suppose there are p_i training samples in each biometric trait. For each biometric trait $i = 1, \ldots, D$, we denote
$$\mathbf{X}^i = [\mathbf{X}^i_1, \mathbf{X}^i_2, \ldots, \mathbf{X}^i_C]$$
as an $n_i \times p_i$ dictionary of training samples consisting of C sub-dictionaries $\mathbf{X}^i_k$'s corresponding to C different classes. Each sub-dictionary
$$\mathbf{X}^i_j = [\mathbf{x}^i_{j,1}, \mathbf{x}^i_{j,2}, \ldots, \mathbf{x}^i_{j,p_j}] \in \mathbb{R}^{n \times p_j}$$
represents a set of training data from the ith modality labeled with the jth class. Note that n_i is the feature dimension of each sample and there are p_j number of training samples in class j. Hence, there are a total of $p = \sum_{j=1}^{C} p_j$ many samples in the dictionary $\mathbf{X}^i_C$. In multimodal biometrics recognition problem given test samples (a matrix) $\mathbf{Y}$, which consists of D different modalities $\{\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^D\}$ where each sample $\mathbf{Y}^i$ consists of d_i observations $\mathbf{Y}^i = [\mathbf{y}^i_1, \mathbf{y}^i_2, \ldots, \mathbf{y}^i_d] \in \mathbb{R}^{n \times d_i}$, the objective is to identify the class to which a test sample $\mathbf{Y}$ belongs to [22], [23].

Let $\mathbf{\Gamma} = [\mathbf{\Gamma}^1, \mathbf{\Gamma}^2, \ldots, \mathbf{\Gamma}^D] \in \mathbb{R}^{p \times d}$ be the matrix formed by concatenating the coefficient matrices with $d = \sum_{i=1}^{D} d_i$, then we can seek for the row-sparse matrix $\mathbf{\Gamma}$ by solving the following ℓ_1/ℓ_q-regularized least square problem
$$\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \frac{1}{2} \sum_{i=1}^{D} \|\mathbf{Y}^i - \mathbf{X}^i\mathbf{\Gamma}^i\|_F^2 + \lambda\|\mathbf{\Gamma}\|_{1,q} \qquad (6)$$

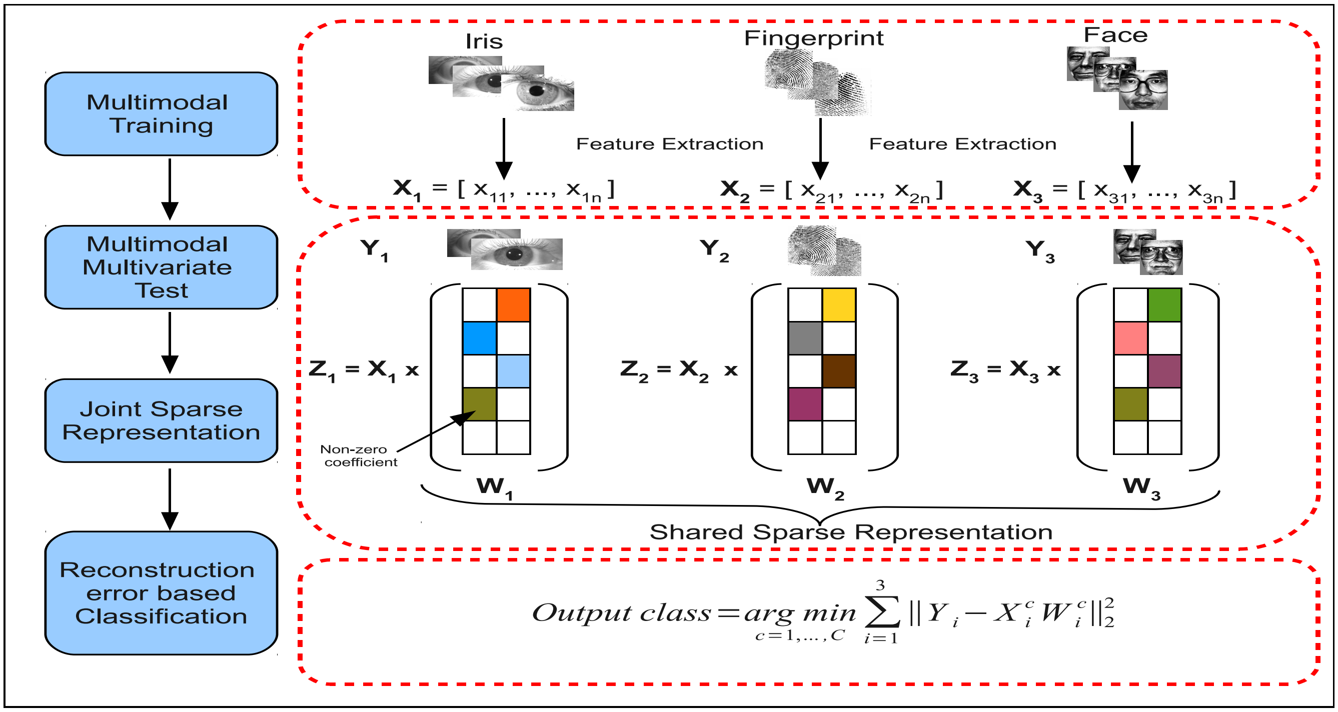

Multimodal Training
Multimodal Multivariate Test
Joint Sparse Representation
Reconstruction error based Classification
Iris
Fingerprint
Face
Feature Extraction
Feature Extraction
$\mathbf{X_1} = [\, x_{11}, \ldots, x_{1n} \,]$
$\mathbf{X_2} = [\, x_{21}, \ldots, x_{2n} \,]$
$\mathbf{X_3} = [\, x_{31}, \ldots, x_{3n} \,]$
$\mathbf{Y_1}$
$\mathbf{Y_2}$
$\mathbf{Y_3}$
$\mathbf{Z_1} = \mathbf{X_1} \times$
$\mathbf{Z_2} = \mathbf{X_2} \times$
$\mathbf{Z_3} = \mathbf{X_3} \times$
Non-zero coefficient
$\mathbf{W_1}$
$\mathbf{W_2}$
$\mathbf{W_3}$
Shared Sparse Representation
$Output\ class = \underset{c=1,\ldots,C}{arg\ min} \sum_{i=1}^{3} \| Y_i - X_i^c W_i^c \|_2^2$

Fig. 2. Overview of multimodal multivariate sparse representation-based classification algorithm [22].

where λ is a positive parameter and q is set greater than 1 to make the optimization problem convex. Here, $\|\boldsymbol{\Gamma}\|_{1,q}$ is a norm defined as $\|\boldsymbol{\Gamma}\|_{1,q} = \sum_{k=1}^{p} \|\boldsymbol{\gamma}^k\|_q$ where $\boldsymbol{\gamma}^k$'s are the row vectors of $\boldsymbol{\Gamma}$. Once $\hat{\boldsymbol{\Gamma}}$ is obtained, the class label associated to an observed vector is then declared as the one that produces the smallest approximation error

$$\hat{j} = \arg\min_{j} \sum_{i=1}^{D} \|\mathbf{Y}^i - \mathbf{X}^i \boldsymbol{\delta}_j^i(\boldsymbol{\Gamma}^i)\|_F^2, \tag{7}$$

where $\boldsymbol{\delta}_j^i$ is the matrix indicator function defined by keeping rows corresponding to the jth class and setting all other rows equal to zero. Note that the optimization problem (6) reduces to the conventional Lasso [24] when $D = 1$ and $d = 1$. The resulting classification algorithm reduces to SRC [18]. In the case, when $D = 1$ (6) is referred to as multivariate Lasso [25]. This model can be extended to handle noise and occlusion [22]. Furthermore, using the kernel methods, it can be made nonlinear [22].

III. Discriminative Dictionary Learning

Dictionaries can be trained for both reconstruction and discrimination applications. In the late nineties, Etemand and Chellappa proposed a Linear Discriminant Analysis (LDA) based basis selection and feature extraction algorithm for classification using wavelet packets [26]. Recently, similar algorithms for simultaneous sparse signal representation and discrimination have also been proposed in [27], [28]. The basic idea in learning a discriminative dictionary is to add an LDA type of discrimination on the sparse coefficients which essentially enforces separability among dictionary atoms of different classes. Some of the other methods for learning discriminative dictionaries include [29], [30], [31], [32], [27], [11]. Additional techniques may be found within these references.

A. Information-theoretic Dictionary Learning

In particular, a dictionary learning method based on information maximization principle was proposed in [33] for action recognition. Given the initial dictionary D^o, the objective is to compress it into a dictionary D^* of size k, which encourages the signals from the same class to have very similar sparse representations.

Let L denote the labels of M discrete values, $L \in [1, M]$. Given a set of dictionary atoms D^*, define $P(L|D^*) = \frac{1}{|D^*|}\sum_{d_i \in D^*} P(L|d_i)$. For simplicity, denote $P(L|d^*)$ as $P(L_{d^*})$, and $P(L|D^*)$ as $P(L_{D^*})$. To enhance the discriminative power of the learned dictionary, the following objective function is considered

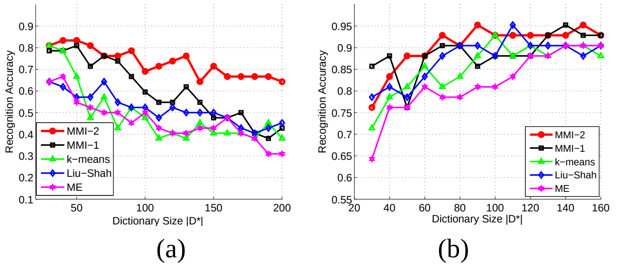
$$\arg\max_{D^*} I(D^*; D^o \backslash D^*) + \lambda I(L_{D^*}; L_{D^o \backslash D^*}) \tag{8}$$

where $\lambda \geq 0$ is the parameter to regularize the emphasis on appearance or label information and I denotes mutual information. One can approximate (8) as

$$\begin{aligned} \arg \max_{d^* \in D^o \backslash D^*} & [H(d^*|D^*) - H(d^*|\bar{D}^*)] \\ & + \lambda[H(L_{d^*}|L_{D^*}) - H(L_{d^*}|L_{\bar{D}^*})], \end{aligned} \tag{9}$$

where H denotes entropy. One can easily notice that the above formulation also forces the classes associated with d^* to be most different from classes already covered by the selected atoms D^*; and at the same time, the classes associated with d^* are most representative among classes covered by the remaining atoms. Thus the learned dictionary is not only compact, but also covers all classes to maintain the discriminability.

In Fig. 3, we present the recognition accuracy on the Keck gesture dataset with different dictionary sizes and over different global and local features [33]. Leave-one-person-out setup is used. That is, sequences performed by a person are left out, and the average accuracy is reported. Initial dictionary size $|D^o|$ is chosen to be twice the dimension of the input signal and sparsity 10 is used in this set of experiments. As can be seen the mutual information-based method, denoted as MMI-2 outperforms the other methods.



(a) (b)

Fig. 3. Recognition accuracy on the Keck gesture dataset with different features and dictionary sizes (shape and motion are global features. STIP is a local feature.) [33]. The recognition accuracy using initial dictionary D^o: (a) 0.23 (b) 0.42. In all cases, the MMI-2 (red line) outperforms the rest.

B. *Kernel Dictionary Learning*

Linear representations are almost always inadequate for representing nonlinear data arising in many practical applications. For example, many types of descriptors in computer vision have intrinsic nonlinear similarity measure functions. The most popular ones include the spatial pyramid descriptor which uses a pyramid match kernel, and the region covariance descriptor which uses a Riemannian metric as the similarity measure between two descriptors. Both of these distance measures are highly non-linear. Unfortunately, the traditional dictionary learning methods, e.g. MOD and K-SVD, are based on linear models. This inevitably leads to poor performances for many datasets, e.g., object classification of Caltech-101 [34] dataset, even when discriminant power is taken into account during the training. This motivates us to study non-linear kernel sparse representations for object representation and classification [9].

Let $\Phi : \mathbb{R}^N \rightarrow \mathcal{F} \subset \mathbb{R}^{\tilde{N}}$ be a non-linear mapping from $\mathbb{R}^N$ into a dot product space $\mathcal{F}$. One can learn a non-linear dictionary $\mathbf{D}$ in the feature space $\mathcal{F}$ by solving the following optimization problem:

$$\arg\min_{\mathbf{D},\mathbf{X}} \ \|\Phi(\mathbf{Y}) - \mathbf{D}\mathbf{X}\|_F^2 \ \ s.t \ \ \|\mathbf{x}_i\|_0 \leq T_0, \forall i. \qquad (10)$$

where $\mathbf{D} \in \mathbb{R}^{\tilde{N} \times K}$ is the sought dictionary, $\mathbf{X} \in \mathbb{R}^{K \times n}$ is a matrix whose ith column is the sparse vector $\mathbf{x}_i$ corresponding to the sample $\mathbf{y}_i$, with maximum of T_0 non-zero entries. It was shown in [34], that there exists an optimal solution $\mathbf{D}^*$ to the problem (10) that has the following form:

$$\mathbf{D}^* = \Phi(\mathbf{Y})\mathbf{A} \qquad (11)$$

for some $\mathbf{A} \in \mathbb{R}^{n \times K}$. As a result, one can seek an optimal dictionary through optimizing $\mathbf{A}$ instead of $\mathbf{D}$. By substituting Eq. 11 into Eq. 10, the problem can be re-written as follows:

$$\arg\min_{\mathbf{A},\mathbf{X}} \ \|\Phi(\mathbf{Y}) - \Phi(\mathbf{Y})\mathbf{A}\mathbf{X}\|_F^2 \ \ s.t \ \ \|\mathbf{x}_i\|_0 \leq T_0, \forall i. \quad (12)$$

In order to see the advantage of this formulation over the original one, we will examine the objective function. Through some manipulation, the cost function can be re-written as:

$$\|\Phi(\mathbf{Y}) - \Phi(\mathbf{Y})\mathbf{A}\mathbf{X}\|_F^2 = \mathbf{tr}((\mathbf{I} - \mathbf{A}\mathbf{X})^T \mathbb{K}(\mathbf{Y},\mathbf{Y})(\mathbf{I} - \mathbf{A}\mathbf{X})),$$

where $\mathbb{K}(\mathbf{Y},\mathbf{Y})$ is a kernel matrix whose elements are computed from $\kappa(i,j) = \Phi(\mathbf{y}_i)^T\Phi(\mathbf{y}_j)$. It is apparent that the objective function is feasible since it only involves a matrix of finite dimension $\mathbb{K} \in \mathbb{R}^{n \times n}$, instead of dealing with a possibly infinite dimensional dictionary. An important property of this formulation is that the computation of $\mathbb{K}$ only requires dot products. Therefore, we are able to employ *Mercer* kernel functions to compute these dot products without carrying out the mapping Φ.

To solve the above optimization problem for learning non-linear dictionaries, we have proposed variants of MOD and K-SVD algorithms in the feature space [34]. The procedure essentially involves two stages: sparse coding and dictionary update in the feature space. For sparse coding, we propose non-linear version of orthogonal matching pursuit algorithm [34]. Once sparse codes are found in the feature space, we update the dictionary atoms in an efficient way.

A synthetic experiment was cone to examine the effectiveness of a learned dictionary in the feature space in [34]. A dictionary is learned from 1500 data samples generated from a 2-dimensional parabola

$$\{y = [y_1, y_2] \in \mathbb{R}^2 \mid y_2 = y_1^2\}.$$

Columns 2-4 in Fig. 4 show level curves of the projection coefficients for three different dictionary atoms. The level curves are obtained as follows. First, every point $y \in \mathbb{R}^2$ is projected onto the selected dictionary atom to get the projection coefficients. Then, points with the same projection coefficients are grouped together and are shown with the same color map. Coefficients of the kernel K-SVD (Bottom row of columns 2-4 in Fig. 4) change most dramatically along the main directions of data's variation, while coefficients of the linear K-SVD do not. This observation implies that non-linear dictionary learning method can provide good representation for data with non-linear structures.

1) Non-linear Discriminative Dictionary Learning: The optimization problem (12) is purely generative. It does not explicitly promote the discrimination which is important for many classification tasks. Using the kernel trick, when the data is transformed into a high dimensional feature space, the data from different classes may still overlap. Hence, generative dictionaries may lead to poor performance in classification

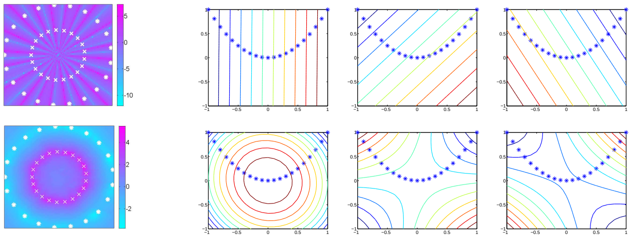


Fig. 4. Left: Comparison of error ratio for K-SVD and kernel K-SVD (common logarithm scale). Right: Comparison between contours of linear K-SVD and kernel K-SVD for three different dictionary atoms. In both figures, the first row corresponds to K-SVD and the second row corresponds to kernel K-SVD.

even when data is non-linearly mapped to a feature space. To overcome this, a method for designing non-linear dictionaries that are simultaneously generative and discriminative was proposed in [35].

Figure 5 presents an important comparison in terms of the discriminative power of learning a discriminative dictionary in the feature space where kernel LDA type of discriminative term has been included in the objective function. A scatter plot of the sparse coefficients obtained using different approaches show that such a discriminative dictionary is able to learn the underlying non-linear sparsity of data as well as it provides more discriminative representation. See [35], [34] for more details on the design of non-linear kernel dictionaries.

C. Unsupervised Dictionary Learning

Dictionary learning techniques for unsupervised clustering have also gained some traction in recent years. In [36], 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 [37]. Additional sparsity motivated subspace clustering methods include [38], [39], [40].

In particular, scale and in-plane rotation invariant clustering approach, which extends the dictionary learning and sparse representation framework for clustering and retrieval of images was proposed in [13]. Figure 6 presents and overview this approach [13]. Given a database of images $\{\mathbf{x}_j\}_{j=1}^N$ and the number of clusters K, the Radon transform [41] is used to find scale and rotation invariant features. It then uses sparse representation methods to simultaneously cluster the data and learn dictionaries for each cluster. One of the main features of this method is that it is effective for both texture and shape-based images. Various experiments in [13] demonstrated the effectiveness of this approach in image retrieval experiments, where the significant improvements in performance are achieved.

D. Dictionary Learning from Partially Labeled Data

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

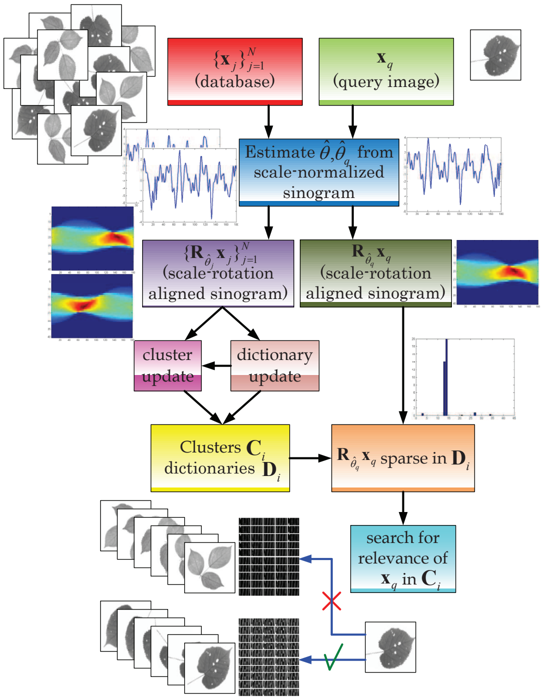

$\{\mathbf{x}_j\}_{j=1}^N$ (database)
$\mathbf{x}_q$ (query image)
Estimate $\hat{\theta}, \hat{\theta}_q$ from scale-normalized sinogram
$\{\mathbf{R}_{\hat{\theta}_j}\mathbf{x}_j\}_{j=1}^N$ (scale-rotation aligned sinogram)
$\mathbf{R}_{\hat{\theta}_q}\mathbf{x}_q$ (scale-rotation aligned sinogram)
cluster update
dictionary update
Clusters $\mathbf{C}_i$ dictionaries $\mathbf{D}_i$
$\mathbf{R}_{\hat{\theta}_q}\mathbf{x}_q$ sparse in $\mathbf{D}_i$
search for relevance of $\mathbf{x}_q$ in $\mathbf{C}_i$

Fig. 6. Overview of simultaneous scale and in-plane rotation invariant clustering and dictionary learning method [13].

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 [42] for an excellent survey of recent efforts on semi-supervised learning.

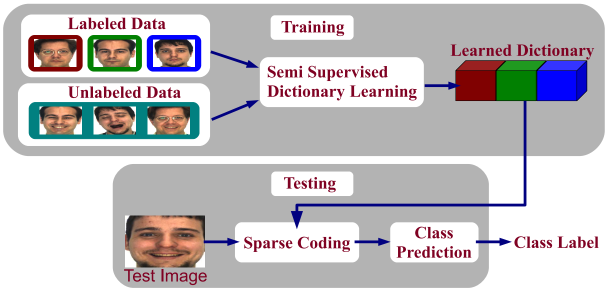


Fig. 7. Block diagram illustrating Semi-Supervised Dictionary Learning [43].

Two of the most popular methods for semi-supervised learning are Co-Training [44] and Semi-Supervised Support Vector Machines (S3VM) [45]. 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. Semi-

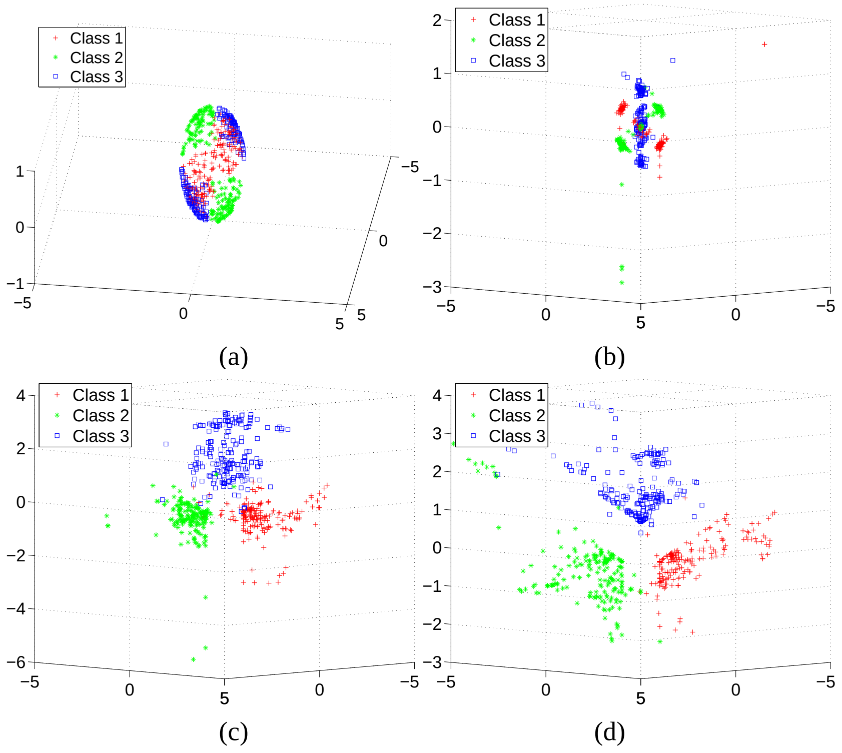


Fig. 5. A synthetic example showing the significance of learning a discriminative dictionary in feature space for classification. (a) Synthetic data which consists of linearly non separable 3D points on a sphere. Different classes are represented by different colors. (b) Sparse coefficients from K-SVD projected onto learned SVM hyperplanes. (c) Sparse coefficients from a non-linear dictionary projected onto learned SVM hyperplanes. (d) Sparse coefficients from non-linear discriminative kernel dictionary projected onto learned SVM hyperplanes [35].

supervised support vector machines consider the labels of the unlabeled data as additional unknowns and jointly optimizes over the classifier parameters and the unknown labels in the SVM framework [46].

An interesting method to learn discriminative dictionaries for classification in a semi-supervised manner was recently proposed in [43]. Figure 7 shows the block diagram of this method [43] which uses both labeled and unlabeled data. While learning a dictionary, probability distribution is maintained 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 method robust to label assignment errors. See [43] for more details on the optimization of the partially labeled dictionary learning.

IV. Domain Adaptive Dictionary Learning

When designing dictionaries, training and testing domains may be different, e.g., different view points and illumination conditions. In [12], a function learning framework is presented for the task of transforming a dictionary learned from one visual domain to the other, while maintaining a domain-invariant sparse representation of a signal. An overview of this method is shown in Fig. 8.

Denote P signals observed in N different domains as $\{\mathbf{Y_1}, ..., \mathbf{Y_N}\}$, where $\mathbf{Y_i} = [\mathbf{y_{i1}}, ..., \mathbf{y_{iP}}]$, $\mathbf{y_{ip}} \in \mathbb{R}^n$. Thus, $\mathbf{y_{ip}}$ denotes the p^{th} signal observed in the i^{th} domain. Let $\mathbf{D_i}$ denote the dictionary for the i^{th} domain, where $\mathbf{D_i} = [\mathbf{d_{i1}} ... \mathbf{d_{iK}}]$, $\mathbf{d_{ik}} \in \mathbb{R}^n$. The domain dictionary learning problem can be formulated as

$$\arg\min_{\{\mathbf{D_i}\}_\mathbf{i}^\mathbf{N},\mathbf{X}} \sum_i^N \|\mathbf{Y_i} - \mathbf{D_i X}\|_F^2 \quad s.t.\ \forall p\ \|\mathbf{x}_p\|_o \leq T, \qquad (13)$$

where $\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x_P}]$, $\mathbf{x_p} \in \mathbb{R}^K$, are the sparse codes and T is a sparsity constant. The set of domain dictionary $\{\mathbf{D_i}\}_i^N$ learned through (13) enable the same sparse codes $\mathbf{x_p}$ for a signal $\mathbf{y_p}$ observed across N different domains to achieve domain adaptation.

A parametric function is used to model domain dictionaries $\mathbf{D_i}$ as follows

$$\mathbf{D_i} = F(\boldsymbol{\theta_i}, \mathbf{W}), \qquad (14)$$

where $\boldsymbol{\theta_i}$ denotes a vector of domain parameters, e.g., view point angles, illumination conditions, etc., and $\mathbf{W}$ denotes the dictionary function parameters [12]. Applying (14) to (13), one can formulate the domain dictionary function learning as follows

$$\arg\min_{\mathbf{W},\mathbf{X}} \sum_i^N \|\mathbf{Y_i} - F(\boldsymbol{\theta_i}, \mathbf{W})\mathbf{X}\|_F^2 \quad s.t.\ \forall p\ \|\mathbf{x}_p\|_o \leq T. \qquad (15)$$

Various linear and non-linear dictionary function learning models are considered in [12] and the optimization problem is

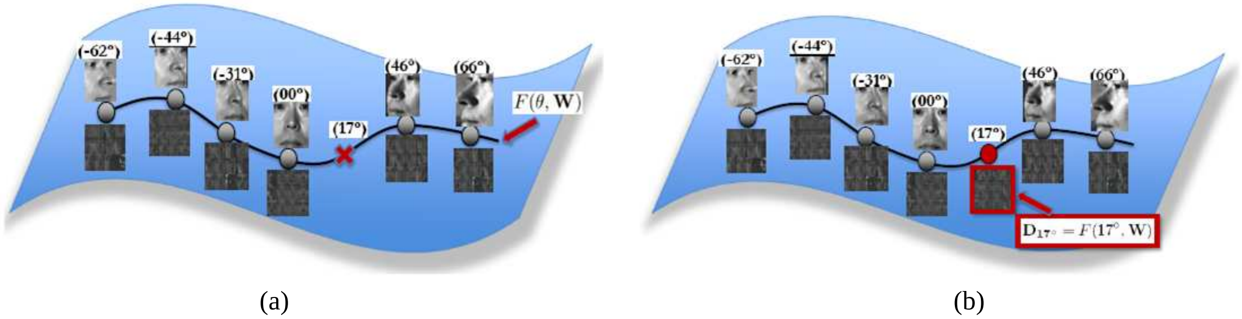


Fig. 8. Overview of the domain adaptive dictionary learning approach proposed in [12]. Consider example dictionaries corresponding to faces at different azimuths. (a) shows a depiction of example dictionaries over a curve on a dictionary manifold. Given example dictionaries, the approach presented in [12] learns the underlying dictionary function $F(\theta, \mathbf{W})$. In (b), the dictionary corresponding to a domain associated with observations is obtained by evaluating the learned dictionary function at corresponding domain parameters.

solved using a simple three step procedure. See [12] for more details on the optimization of (15) and experimental results on various datasets.

V. Conclusion

In this paper, we reviewed some of the approaches to object recognition based on the recently introduced theories of SR and DL. Furthermore, through the use of Mercer kernels, we showed how sparse representation and dictionary learning methods can be made non-linear. Even though, the main emphasis was given to object recognition, these methods can offer compelling solutions to other computer vision and machine learning problems such as matrix factorization, tracking, object detection, weakly supervised learning [43] and object recognition from video [47].

An excellent review of SR and DL from the view of analysis co-sparse model as well as a discussion on the open problems in this field can be found in [48].

Acknowledgment

This work was partially supported by an Office of Naval Research grant N00014-12-1-0124.

References

[1] R. Rubinstein, A. Bruckstein, and M. Elad, “Dictionaries for sparse representation modeling,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045 –1057, june 2010.

[2] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. 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.

[3] M. Elad, M. Figueiredo, and Y. Ma, “On the role of sparse and redundant representations in image processing,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972 –982, june 2010.

[4] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[5] S. Chen, D. Donoho, and M. Saunders, “Atomic decomposition by basis pursuit,” *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1998.

[6] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, “Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition,” *1993 Conference Record of the 27th Asilomar Conference on Signals, Systems and Computers*, pp. 40–44, 1993.

[7] J. A. Tropp, “Greed is good: Algorithmic results for sparse approximation,” *IEEE Trans. Info. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[8] A. M. Bruckstein, D. L. Donoho, and M. Elad, “From sparse solutions of. systems of equations to sparse. modeling of signals and images,” *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[9] B. A. Olshausen and D. J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[10] V. M. Patel, T. Wu, S. Biswas, P. J. Phillips, and R. Chellappa, “Dictionary-based face recognition under variable lighting and pose,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 954–965, June 2012.

[11] H. V. Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, “Sparse embedding: A framework for sparsity promoting dimensionality reduction,” in *European Conference on Computer Vision*, 2012.

[12] Q. Qiu, V. M. Patel, P. Turaga, and R. Chellappa, “Domain adaptive dictionary learning,” in *European Conference on Computer Vision*, 2012.

[13] Y.-C. Chen, C. S. Sastry, V. M. Patel, P. J. Phillips, and R. Chellappa, “Rotation invariant simultaneous clustering and dictionary learning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

[14] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, “Dictionary-based face recognition from video,” *European Conference on Computer Vision*, October 2012.

[15] S. Shekhar, V. M. Patel, and R. Chellappa, “Synthesis-based recognition of low resolution faces,” in *International Joint Conference on Biometrics*, oct. 2011, pp. 1 –6.

[16] K. Engan, S. O. Aase, and J. H. Husoy, “Method of optimal directions for frame design,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, pp. 2443–2446, 1999.

[17] M. Aharon, M. Elad, and A. M. Bruckstein, “The k-svd: an algorithm for designing of overcomplete dictionaries for sparse representation,” *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[18] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[19] J. Pillai, V. Patel, R. Chellappa, and N. Ratha, “Sectored random projections for cancelable iris biometrics,” in *IEEE International Conference on Acoustics Speech and Signal Processing*, march 2010, pp. 1838 –1841.

[20] J. Pillai, V. M. Patel, and R. Chellappa, “Sparsity inspired selection and recognition of iris images,” *Third IEEE International Conference on Biometrics : Theory, Applications and Systems*, 2009.

[21] J. Pillai, V. Patel, R. Chellappa, and N. Ratha, “Secure and robust iris recognition using random projections and sparse representations,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1877–1893, Sept. 2011.

[22] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, “Joint sparsity-based robust multimodal biometrics recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted 2012.

[23] N. H. Nguyen, N. M. Nasrabadi, and T. D. Tran, “Robust multi-sensor classification via joint sparse representation,” in *International Conference on Information Fusion*, 2011.

[24] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Jour-*

nal of the Royal Statistical Society: Series B, vol. 58, no. 1, pp. 267–288, 1996.

[25] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, vol. 68, no. 1, pp. 49–67, 2006.

[26] 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.

[27] X. F. M. Yang, L. Zhang and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," 2011, *ICCV*.

[28] K. Huang and S. Aviyente, "Sparse representation for signal classification," *NIPS*, vol. 19, pp. 609–616, 2007.

[29] 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, april 2012.

[30] J. Mairal, F. Bach, J. Pnce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2008.

[31] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," *Advances in Neural Information Processing Systems*, 2008.

[32] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Computer Vision and Pattern Recognition*, 2010.

[33] Q. Qiu, Z. Jiang, and R. Chellappa, "Sparse dictionary-based representation and recognition of action attributes," *International Conference on Computer Vision*, 2011.

[34] H. V. Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Kernel dictionary learning," in *ICASSP*, 2012.

[35] A. Shrivastava, H. V. Nguyen, V. M. Patel, and R. Chellappa, "Design of non-linear discriminative dictionaries for image classification," in *Asian Conference on Computer Vision*, 2012.

[36] P. Sprechmann and G. Sapiro, "Dictionary learning and sparse coding for unsupervised clustering," *Proc. IEEE Conf. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2042–2045, March 2010.

[37] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3501–3508, June 2010.

[38] E. Elhamifar and R. Vidal, "Sparse subspace clustering," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 2790–2797, June 2009.

[39] S. Rao, R. Tron, R. Vidal, and Y. Ma, "Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, June 2008.

[40] M. Soltanolkotab and E. J. Candnes, "A geometric analysis of subspace clustering with outliers," *Preprint*, 2011.

[41] S. Helgason, *The Radon Transfrom*. Birkhauser, Boston, 1980.

[42] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-supervised learning*, ser. Adaptive computation and machine learning. MIT Press, 2006. [Online]. Available: http://books.google.com/books?id=kfqvQgAACAAJ

[43] A. Shrivastava, J. K. Pillai, V. M. Patel, and R. Chellappa, "Learning discriminative dictionaries with partially labeled data," in *International Conference on Image Processing (ICIP)*, 2012.

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

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

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

[47] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition from video," in *European Conference on Computer Vision*, 2012.

[48] M. Elad, "Sparse and redundant representation modeling - what next?" *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 922–928, 2012.