Sparse Representations, Compressive Sensing and Dictionaries for Pattern Recognition

Vishal M. Patel
Center for Automation Research
UMIACS
University of Maryland
College Park, MD 20742
Email: pvishalm@umiacs.umd.edu

Rama Chellappa
Center for Automation Research
UMIACS
University of Maryland
College Park, MD 20742
Email: rama@umiacs.umd.edu

***Abstract*—In recent years, the theories of Compressive Sensing (CS), Sparse Representation (SR) and Dictionary Learning (DL) have emerged as powerful tools for efficiently processing data in non-traditional ways. An area of promise for these theories is object recognition. In this paper, we review the role of SR, CS and DL for object recognition. Algorithms to perform object recognition using these theories are reviewed.**

An important aspect in object recognition is feature extraction. Recent works in SR and CS have shown that if sparsity in the recognition problem is properly harnessed then the choice of features is less critical. What becomes critical, however, is the number of features and the sparsity of representation. This issue is discussed in detail.

I. Introduction

Sparse and redundant signal representations have recently drawn much interest in vision, signal and image processing [1], [2], [3]. This is due in part to the fact that signals and images of interest can be sparse or compressible in some dictionary. The dictionary can be either based on a mathematical model of the data or it can be learned directly from the data. It has been observed that learning a dictionary directly from training rather than using a predetermined dictionary (such as wavelet or Fourier) usually leads to better representation and hence can provide improved results in many practical applications such as restoration and classification.

In this paper, we summarize approaches to object recognition based on sparse representation, compressive sensing and dictionary learning. We first give an overview of these theories. Then, we show that how traditionally used features such as principle component analysis and linear discriminant analysis provide information as good as randomly projected features when classification is performed using these theories. What becomes important when such features are used is the dimension of features and how the sparse representation is computed. Finally, we present some simulation results to show the effectiveness of these methods for object recognition.

II. Background

In this section, we present an overview of sparse representations, compressive sampling and dictionary learning.

A. Sparse Representation

Sparse coding allows us to represent a signal as a linear combination of a few atoms of a dictionary. Suppose the signal $\mathbf{x} \in I\!R^m$ of measurements x_i satisfies

$$\mathbf{x} = \mathbf{D}\alpha \quad for \quad \mathbf{D} \in I\!R^{m \times n}, \quad m << n. \tag{1}$$

As $m << n$, the system (1) admits infinitely many solutions. One way of choosing a solution for (1) involves taking the solution that is 'smallest,' (in the l^2 sense), which corresponds to

$$\tilde{\alpha} = \mathbf{D}^T(\mathbf{D}\mathbf{D}^T)^{-1}\mathbf{x},$$

called the *pseudo inverse* of $\mathbf{D}$.

One of the popular ways of computing a solution of (1), involves finding the 'sparsest' vector. The sparsest solution may be obtained by solving the following problem

$$\min_{\omega} \|\omega\|_0 \quad subject \quad to \quad \mathbf{D}\omega = \mathbf{x}, \tag{2}$$

where $\|\alpha\|_0 := |\#\{i : \alpha_i \neq 0\}| < n$, which is a count for the number of nonzero elements in α. As the problem in (2) is NP hard alternative solutions are often sought. For instance, Basis Pursuit (see e.g. [4]) offers the solution via l_1 minimization as

$$\min_{\omega} \|\omega\|_1 \quad subject \quad to \quad \mathbf{D}\omega = \mathbf{x}. \tag{3}$$

The sparsest recovery is possible provided that certain conditions are met [5].

One can adapt the above framework to noisy setting, where the measurements are contaminated with an error η obeying $\|\eta\|_2 < \epsilon$, that is

$$\mathbf{x} = \mathbf{D}\alpha + \eta \quad \text{for} \quad \|\eta\|_2 < \epsilon. \tag{4}$$

A stable solution could be obtained from

$$\min_{\alpha} \|\alpha\|_1 \quad \text{subject to} \quad \|Ax - y\|_2 < \epsilon. \tag{5}$$

Intuitively, the l_1-norm is the convex function closest to the l_0-(quasi)-norm, so this substitution is referred to as convex relaxation [5]. One hopes that the solution to the relaxation yields a good approximation of the ideal solution vector. The advantage of the new formulation is that it can be solved in

polynomial time with standard scientific softwares. One can also use greedy pursuits and iterative thresholding algorithms to solve the above problems [5], [6], [7], [8].

B. Compressive Sensing

Compressive sampling is a new concept in signal processing and information theory where one measures a small number of non-adaptive linear combinations of the signal. These measurements are usually much smaller than the number of samples that define the signal. From these small number of measurements, the signal is then reconstructed by a non-linear procedure [9], [10].

More precisely, suppose $\mathbf{x} \in I\!R^m$ is k-sparse in a basis (or a Dictionary) Ψ, so that $\mathbf{x} = \Psi\mathbf{x}_0$, with $\| \mathbf{x}_0 \|_0 = k \ll m$. In the case when $\mathbf{x}$ is compressible in Ψ, it can be well approximated by the best k-term representation. Consider a random $n \times m$ measurement matrix Φ with $n < m$ and assume that m measurements, that make up a vector $\mathbf{y}$, are made such that

$$\mathbf{y} = \Phi\mathbf{x} = \Phi\Psi\mathbf{x}_0 = \Theta\mathbf{x}_0.$$

According to CS theory, when Θ satisfies the restricted isometry property (RIP) [11], one can reconstruct $\mathbf{x}$ via its coefficients $\mathbf{x}_0$ by solving the following ℓ^1 minimization problem [9], [10]:

$$\hat{\mathbf{x}_0} = \arg \min_{\mathbf{x}_0 \in I\!R^N} \| \mathbf{x}_0 \|_1 \quad \text{subject to} \quad \mathbf{y} = \Phi\Psi\mathbf{x}_0. \tag{6}$$

A matrix Θ is said to satisfy the RIP of order k with constants $\delta_K \in (0, 1)$ if

$$(1 - \delta_k) \| \mathbf{v} \|_2^2 \leq \| \Theta\mathbf{v} \|_2^2 \leq (1 + \delta_k) \| \mathbf{v} \|_2^2 \tag{7}$$

for any $\mathbf{v}$ such that $\| \mathbf{v} \|_0 \leq k$.

One popular class of measurement matrices satisfying an RIP is the one consisting of i.i.d. Gaussian entries. It is a well known fact that if Φ is an $n \times m$ Gaussian matrix where $n > \mathcal{O}(k \log m)$ and Ψ is a sparsifying basis, then Θ satisfies the RIP with high probability. One can also use greedy pursuits and iterative soft or hard thresholding algorithms to recover signals from compressive measurements.

C. Dictionary Learning

It has been observed that learning a dictionary directly from training rather than using a predetermined dictionary (such as wavelet or Fourier) usually leads to better representation and hence can provide improved results in many practical image processing applications such as restoration and classification [3]. Designing dictionaries based on training is a much recent approach to dictionary design which is strongly motivated by recent advances in the sparse representation theory [12],[13],[3]. In dictionary learning methods, given a set of examples $\mathbf{B} = [\mathbf{x}_1, \cdots, \mathbf{x}_m]$, 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

$$(\hat{\mathbf{D}}, \hat{\Gamma}) = \arg \min_{\mathbf{D}, \Gamma} \|\mathbf{B} - \mathbf{D}\Gamma\|_F^2 \text{ subject to } \forall i \ \|\gamma_i\|_0 \leq T_0 \tag{8}$$

where γ_i represents a column of Γ. Here, $\|\mathbf{A}\|_F$ denotes the Frobenius norm defined as $\|\mathbf{A}\|_F = \sqrt{\sum_{ij} \mathbf{A}_{ij}^2}$. Two of the simplest algorithms for finding such dictionary are the method of optimal directions (MOD)[12] and the K-SVD [13] algorithm.

Both MOD and K-SVD are iterative methods and they alternate between sparse-coding and dictionary update steps. First, a dictionary $\mathbf{D}$ with ℓ_2 normalized columns is initialized. Then, the main iteration is composed of the following two stages:

- *Sparse coding*: In this step, $\mathbf{D}$ is fixed and the following optimization problem is solved to compute the representation vector γ_i for each example $\mathbf{x}_i$

  $$i = 1, \cdots, m, \quad \min_{\gamma_i} \|\mathbf{x}_i - \mathbf{D}\gamma_i\|_2^2 \text{ s. t. } \|\gamma_i\|_0 \leq T_0.$$

  As discussed earlier, since the above problem is NP-hard, approximate solutions are usually sought. Any standard technique [4] can be used but a greedy pursuit algorithm such as orthogonal matching pursuit [6],[7] is often employed due to its efficiency [8].
- *Dictionary update*: This is where both MOD and K-SVD algorithms differ. The MOD algorithm updates all the atoms simultaneously by solving a quadratic problem whose solution is given by $\mathbf{D} = \mathbf{B}\Gamma^\dagger$, where $\Gamma^\dagger$ denotes the Moore-Penrose pseudo-inverse. Even though the MOD algorithm is very effective and usually converges in a few iterations, it suffers from the high complexity of the matrix inversion.
  In the case of K-SVD, the dictionary update is performed atom-by-atom in an efficient way rather than using a matrix inversion. It has been observed that the K-SVD algorithm requires fewer iterations to converge than the MOD method.

D. Discriminative Dictionary Learning

While dictionaries are often trained to obtain good reconstruction, training dictionaries with a specific discriminative criteria 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 in the late nineties [14]. Recently, similar algorithms for simultaneous sparse signal representation and discrimination have also been proposed [15], [16], [17]. In [17], Huang and Aviyente present a framework for signal classification that combines a discriminative method with a generative method using LDA and a pre-defined dictionary. A similar algorithm called supervised simultaneous orthogonal matching pursuit (SSOMP) is presented by Kokiopoulou and Frossard in [16].

Suppose that we are given C distinct classes and a set of m_i training images per class, $i \in \{1, \cdots, C\}$. We identify an $l \times q$ grayscale image as an N-dimensional vector, $\mathbf{x}$, which can be obtained by stacking its columns, where $N = l \times q$. Let

$$\mathbf{B}_i = [\mathbf{x}_{i1}, \cdots, \mathbf{x}_{im_i}] \in I\!R^{N \times m_i} \tag{9}$$

be an $N \times m_i$ matrix of training images corresponding to the i^{th} class. Similarly, we define a new matrix

$$\begin{aligned}\mathbf{A} &= [\mathbf{B}_1, ..., \mathbf{B}_C] \in I\!R^{N \times M} \\ &= [\mathbf{x}_{11}, ..., \mathbf{x}_{1m_1} | \mathbf{x}_{21}, ..., \mathbf{x}_{2m_2} | ..... | \mathbf{x}_{C1}, ..., \mathbf{x}_{Cm_C}],\end{aligned} \tag{10}$$

as concatenation of training samples from all the classes, where $M = \sum_{i=1}^{C} m_i$. Then, for dimensionality reduction, one can decompose $\mathbf{A}$ in the following form

$$\mathbf{A} = \mathbf{DS}, \;\; \mathbf{D} \in I\!R^{N \times R}, \;\; \mathbf{S} \in I\!R^{R \times M},$$

where $\mathbf{D}$ is drawn from a predefined dictionary $\tilde{\mathbf{D}}$ and $\mathbf{S}$ contains the coefficients. In other words, every column of $\mathbf{A}$ is represented in the same set of basis functions using different coefficients. This can be viewed as a dimensionality reduction step where each data sample is represented in the subspace spanned by the columns of $\mathbf{D}$, using only $R \ll N$ coefficients. Similarly, one can formulate a supervised dimensionality reduction problem as follows

$$\min_{\mathbf{DS}} \|\mathbf{A} - \mathbf{DS}\|_F^2 - \lambda J(\mathbf{D}) \;\; s.t \;\; \mathbf{D} \subseteq \tilde{\mathbf{D}},$$

where J denotes the cost function that captures the separability of different classes [16].

Note that approaches mentioned above are based on predefined dictionary. In contrast with these methods, Rodriguez and Sapiro present [15] a method that learns a non-parametric dictionary which is efficient for simultaneous sparse representation as well as class discrimination. Other methods have also been proposed for learning discriminative dictionaries [18], [19], [20], [21], [22], and [23]. In particular, a dictionary learning method based on information maximization principle was proposed in [24] for action recognition. The objective function in [24] maximizes the mutual information between what has been learned and what remains to be learned in terms of appearance information and class distribution for each dictionary item. A Gaussian Process (GP) model is proposed for sparse representation to optimize the dictionary objective function. The sparse coding property allows a kernel with a compact support in GP to realize a very efficient dictionary learning process. Hence an action video can be described by a set of compact and discriminative action attributes. In [25] a discriminative K-SVD method was proposed for face recognition. This framework was recently extended for object recognition in [26]. Additional techniques for discriminative dictionary learning may be found within these references.

E. Feature Extraction

Extraction of relevant low dimensional features of an object is an important issue in object recognition. Many methods have been developed for transforming high-dimensional features into lower dimensional feature space. Some of them include principle component analysis, linear discriminant analysis and locality preserving projections [27], [28], [29]. Advances in SR and CS have shown that the precise choice of features is no longer critical. What is critical is that the dimension of the features and the sparsity of the representation. It was shown that, even random features contain enough information to correctly classify any test sample [29]. This is partly motivated by the following lemma [30], [31], [32]

Lemma 1. *(Johnson-Lindenstrauss) Let $\epsilon \in (0, 1)$ be given. For every set S of $\sharp(S)$ points in $I\!R^N$, if n is a positive integer such that $n > n_0 = O\left(\frac{\ln(\sharp(S))}{\epsilon^2}\right)$, there exists a Lipschitz mapping $f : R^N \rightarrow R^n$ such that*

$$(1-\epsilon)\|\mathbf{u} - \mathbf{v}\|^2 \leq \|f(\mathbf{u}) - f(\mathbf{v})\|^2 \leq (1+\epsilon)\|\mathbf{u} - \mathbf{v}\|^2 \tag{11}$$

for all $\mathbf{u}, \mathbf{v} \in S$.

This lemma essentially states that, a set S of points in $I\!R^N$ can be embedded into a lower-dimensional Euclidean space $I\!R^n$ such that the pairwise distance of any two points is approximately maintained. In fact, it can be shown that f can be taken as a linear mapping represented by an $n \times N$ matrix Φ whose entries are randomly drawn from certain probability distributions. This in turn implies that it is possible to change the original form of the data and still preserve its statistical characteristics useful for recognition. One can clearly see the link between JL-lemma and the restricted isometry property [32].

Let Φ be an $n \times N$ random matrix with $n \leq N$ such that each entry $\phi_{i,j}$ of Φ is an independent realization of q, where q is a random variable on a probability measure space (Ω, ρ). It has been shown that given any set of points S, the following are some of the matrices that will satisfy (11) with high probability, provided n satisfies the condition of the Lemma 1 [31]:

- $n \times N$ random matrices Φ whose entries $\phi_{i,j}$ are independent realizations of Gaussian random variables $\phi_{i,j} \sim N\left(0, \frac{1}{n}\right)$.
- Independent realizations of ± 1 Bernoulli random variables

$$\phi_{i,j} \doteq \begin{cases} +1/\sqrt{n}, & \text{with probability } \frac{1}{2} \\ -1/\sqrt{n}, & \text{with probability } \frac{1}{2}. \end{cases}$$

- Independent realizations of related distributions such as

$$\phi_{i,j} \doteq \begin{cases} +\sqrt{3/n}, & \text{with probability } \frac{1}{6} \\ 0, & \text{with probability } \frac{2}{3} \\ -\sqrt{3/n}, & \text{with probability } \frac{1}{6}. \end{cases}$$

III. Algorithms and Applications

To illustrate the effectiveness of SR, CS and DL methods for object recognition, in this section, we highlight some of the results on face recognition [29], [33] and action recognition [24].

A. Face recognition

Sparse representation-based classification (SRC) [29] was one of the first methods that showed the effectiveness of SR and CS for face recognition. The idea 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. In particular, 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. This algorithm was shown to be robust to noise and occlusion.

The recognition rates achieved by the SRC method for face recognition with different features and dimensions are summarized in Table I on the extended Yale B Dataset [34]. As it can be seen from Table I the SRC method achieves the best recognition rate of 98.09% with randomfaces of dimension 504. Note that the recognition rate does not change significantly with different features provided that the dimension of the feature is high enough. This can be seen from the last column of Table I. The SRC framework was extended for cancelable iris biometric in [35].

TABLE I
RECOGNITION RATES (IN %) OF SRC ALGORITHM [29] ON THE EXTENDED YALE B DATABASE.

Dimension	30	56	120	504
Eigen	86.5	91.63	93.95	96.77
Laplacian	87.49	91.72	93.95	96.52
Random	82.60	91.47	95.53	98.09
Downsample	74.57	86.16	92.13	97.10
Fisher	86.91	-	-	-

The SRC method uses training samples as dictionary. It recognizes faces by solving an optimization problem over the set of images enrolled into the database. This solution trades robustness and size of the database against computational efficiency. To deal with this, a dictionary-based face recognition (DFR) algorithm was recently proposed in [33]. This method consists of two main stages. In the first stage, given training samples from each class, class specific dictionaries are trained with some fixed number of atoms. In the second stage, a novel test image is projected onto the span of the atoms in each learned dictionary. The residual vectors are then used for classification. Furthermore, assuming the Lambertian reflectance model for the facial surface, a relighting approach is introduced within this framework so that one can add many elements to gallery and robustness to illumination changes can be realized. This method was shown to be very efficient and effective in recognizing face images under varying illumination.

The average rank-1 results obtained using various methods are summarized in Table II on the PIE database [36]. The average rank-1 recognition rate achieved by DFR method is 99% and it outperforms the other competitive methods that follow similar experimental setting.

B. Action recognition

In [24], an information maximization-based dictionary learning method was proposed for action recognition. Given

TABLE II
AVERAGE RANK-1 RECOGNITION RATES (RR) (IN %) OF DIFFERENT METHODS ON THE PIE DATABASE [37].

Method	DFR	MA[38]	MB[38]	[39]
RR	**99**	93	96	94

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

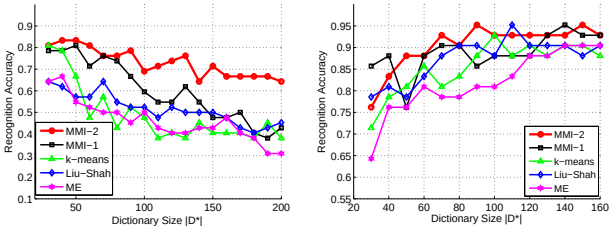
$$\arg\max_{D^*} I(D^*; D^o \backslash D^*) + \lambda I(L_{D^*}; L_{D^o \backslash D^*}) \tag{12}$$

where $\lambda \geq 0$ is the parameter to regularize the emphasis on appearance or label information and I denotes mutual information. One can approximate (12) as

$$\begin{aligned} \arg \max_{d^* \in D^o \backslash D^*} & [H(d^*|D^*) - H(d^*|\bar{D}^*)] \\ & + \lambda [H(L_{d^*}|L_{D^*}) - H(L_{d^*}|L_{\bar{D}^*})], \end{aligned} \tag{13}$$

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. 1, we present the recognition accuracy on the Keck gesture dataset with different dictionary sizes and over different global and local features [24]. 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. 1. Recognition accuracy on the Keck gesture dataset with different features and dictionary sizes (shape and motion are global features. STIP is a local feature.) [24]. 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.

IV. Conclusion

In this paper, we reviewed some of the approaches to object recognition based on the recently introduced theories of sparse representation, compressed sensing and dictionary learning. Furthermore, we discussed that the type of features is flexible when sparse representation-based classification is used for object recognition. What is important is the dimension of features and the sparsity of representation. Even though, the main emphasis was given to object recognition, these methods can offer compelling solutions to other computer vision problems such as clustering, matrix factorization, tracking and object detection.

Acknowledgment

This work was partially supported by a MURI grant from the Office of Naval Research under the grant N00014-08-1-0638.

References

[1] 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.

[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] 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.

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

[5] 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.

[6] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[7] 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.

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

[9] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[10] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[11] ——, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, August 2006.

[12] 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.

[13] 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.

[14] 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.

[15] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries," *Tech. Report, University of Minnesota*, Dec. 2007.

[16] E. Kokiopoulou and P. Frossard, "Semantic coding by supervised dimensionality reduction," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 806–818, Aug. 2008.

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

[18] M. Ranzato, F. Haung, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[19] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," *Advances in Neural Information Processing Systems*, Vancouver, B.C., Canada 2006.

[20] 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*, Anchorage, AL, June 2008.

[21] J. Mairal, M. Leordeanu, F. Bach, M. Herbert, and J. Ponce, "Discriminative sparse image models for class-specific edge detection and image interpretation," *Proc. of the European Conference on Computer Vision*, 2008.

[22] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," *International Conference on Machine Learning*, Montreal, Canada, June 2009.

[23] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," *Advances in Neural Information Processing Systems*, Vancouver, B.C., Canada, Dec. 2008.

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

[25] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, june 2010, pp. 2691 –2698.

[26] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, june 2011, pp. 1697 –1704.

[27] R. Duda and P. Hart, *Pattern Classification*. Wiley, 2001.

[28] X. He and P. Niyogi, "Locality preserving projections," *Advances in Neural Information Processing Systems*, 2003.

[29] 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.

[30] W. B.Johnson and J. Lindenstrauss, "Extensions of lipschitz maps into a hilbert space," *Contemp. Math.*, pp. 189–206, 1984.

[31] D. Achlioptas, "Database-friendly random projections," *ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems*, pp. 274–281, 2001.

[32] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, Dec. 2008.

[33] V. M. Patel, T. Wu, S. Biswas, P. J. Phillips, and R. Chellappa, "Illumination robust dictionary-based face recognition," in *IEEE International Conf. on Image Process.*, 2011.

[34] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Ilumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, June 2001.

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

[36] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression database," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.

[37] V. M. Patel, T. Wu, S. Biswas, R. Chellappa, and P. J. Phillips, "Dictionary-based face recognition," *UMIACS Tech. Report, UMIACS-TR-2010-07/CAR-TR-1030, University of Maryland, College Park*, July 2010.

[38] S. K. Zhou, G. Aggarwal, and D. W. Jacobs, "Appearance characterization of linear lambertian objects, generalized photometric stereo, and illumination-invariant face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 230–245, Feb. 2007.

[39] S. Biswas, G. Aggarwal, and R. Chellappa, "Robust estimation of albedo for illumination-invariant matching and shape recovery," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 884–899, Mar. 2009.