SPARSE REPRESENTATION FOR COMPUTER VISION AND PATTERN RECOGNITION

By

John Wright

Yi Ma

Julien Mairal

Guillermo Sapiro

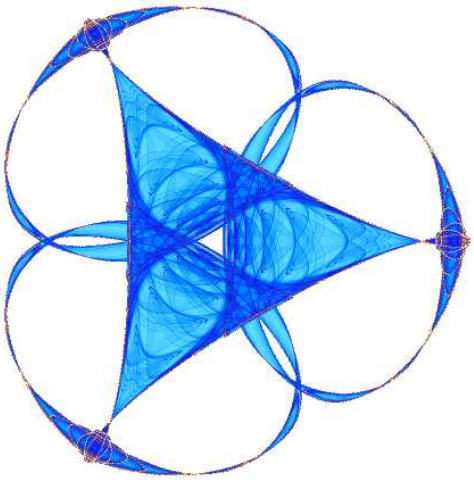
Thomas Huang

and

Shuicheng Yan

IMA Preprint Series # 2252

(May 2009)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA

400 Lind Hall
207 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612-624-6066 Fax: 612-626-7370
URL: http://www.ima.umn.edu

Sparse Representation For Computer Vision and Pattern Recognition

John Wright*, *Member*, Yi Ma*, *Senior Member*, Julien Mairal†, *Member*, Guillermo Sapiro‡, *Senior Member*, Thomas Huang§, *Life Fellow*, Shuicheng Yan¶, *Member*

***Abstract*—Techniques from sparse signal representation are beginning to see significant impact in computer vision, often on non-traditional applications where the goal is not just to obtain a compact high-fidelity representation of the observed signal, but also to extract semantic information. The choice of dictionary plays a key role in bridging this gap: unconventional dictionaries consisting of, or learned from, the training samples themselves provide the key to obtaining state-of-the-art results and to attaching semantic meaning to sparse signal representations. Understanding the good performance of such unconventional dictionaries in turn demands new algorithmic and analytical techniques. This review paper highlights a few representative examples of how the interaction between sparse signal representation and computer vision can enrich both fields, and raises a number of open questions for further study.**

I. Introduction

Sparse signal representation has proven to be an extremely powerful tool for acquiring, representing, and compressing high-dimensional signals. This success is mainly due to the fact that important classes of signals such as audio and images have naturally sparse representations with respect to fixed bases (i.e., Fourier, Wavelet), or concatenations of such bases. Moreover, efficient and provably effective algorithms based on convex optimization or greedy pursuit are available for computing such representations with high fidelity [10].

While these successes in classical signal processing applications are inspiring, in computer vision we are often more interested in the content or semantics of an image rather than a compact, high-fidelity representation. One might justifiably wonder, then, whether sparse representation can be useful at all for vision tasks. The answer has been largely positive: in the past few years, variations and extensions of ℓ^1 minimization have been applied to many vision tasks, including face recognition [71], image super-resolution [75], motion and data segmentation [33], [56], supervised denoising and inpainting [51] and background modeling [16], [21] and image classification [47], [48]. In almost all of these applications, using sparsity as a prior leads to state-of-the-art results.

The ability of sparse representations to uncover semantic information derives in part from a simple but important property of the data: although the images (or their features) are naturally very high dimensional, in many applications images belonging to the same class exhibit *degenerate structure*. That is, they lie on or near low-dimensional subspaces, submanifolds, or stratifications. If a collection of representative samples are found for the distribution, we should expect that a typical sample have a very sparse representation with respect to such a (possibly learned) basis.[1] Such a sparse representation, if computed correctly, could naturally encode the semantic information of the image.

However, to successfully apply sparse representation to computer vision tasks, we typically have to address the additional problem of *how to correctly choose the basis for representing the data*. This is different from the conventional setting in signal processing where a given basis with good property (such as being sufficiently incoherent) can be assumed. In computer vision, we often have to learn from given sample images a task-specific (often overcomplete) dictionary; or we have to work with one that is not necessarily incoherent. As a result, we need to extend the existing theory and algorithms for sparse representation to new scenarios.

This paper will feature a few representative examples of sparse representation in computer vision. These examples not only confirm that sparsity is a powerful prior for visual inference, but also suggest how vision problems could enrich the theory of sparse representation. Understanding why these new algorithms work and how well they work can greatly improve our insights to some of the most challenging problems in computer vision.

II. Robust Face Recognition: Confluence of Practice and Theory

Automatic face recognition remains one of the most visible and challenging application domains of computer vision [77]. Foundational results in the theory of sparse representation have recently inspired significant progress on this difficult problem.

*John Wright and Yi Ma are with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign. Address: 145 Coordinated Science Laboratory, 1308 West Main Street, Urbana, IL 61801. Email: jnwright@uiuc.edu, yima@uiuc.edu

†Julien Mairal is with the INRIA-Willow project, Ecole Normale Supérieure, Laboratoire d'Informatique de l'Ecole Normale Supérieure (INRIA/ENS/CNRS UMR 8548), 45, rue d'Ulm 75005, Paris, France. Email: julien.mairal@m4x.org

‡Guillermo Sapiro is with the Department of Electrical and Computer Engineering, University of Minnesota. Address: 200 Union Street SE, Minneapolis, MN 55455. Email: guille@ece.umn.edu

§ Thomas Huang is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign. Address: 2039 Beckman Institute, MC-251 405 N. Mathews, Urbana, IL 61801. Email: t-huang1@illinois.edu

¶ Shuicheng Yan is with the Department of Electrical and Computer Engineering, National University of Singapore. Address: Office E4-05-11, 4 Engineering Drive 3, 117576, Singapore. Email: eleyans@nus.edu.sg

[1]We use the term "basis" loosely here, since the *dictionary* can be overcomplete and, even in the case of just complete, there is no guarantee of independence between the atoms.

The key idea is a judicious choice of dictionary: representing the test signal as a sparse linear combination of *the training signals themselves*. We will first see how this approach leads to simple and surprisingly effective solutions to face recognition. In turn, the face recognition example reveals new theoretical phenomena in sparse representation that may seem surprising in light of prior results.

A. From Theory to Practice: Face Recognition as Sparse Representation

Our approach to face recognition assumes access to well-aligned training images of each subject, taken under varying illumination.[2] We stack the given N_i training images from the i-th class as columns of a matrix $\boldsymbol{D}_i \doteq [\boldsymbol{d}_{i,1}, \boldsymbol{d}_{i,2}, \ldots, \boldsymbol{d}_{i,N_i}] \in \mathbb{R}^{m \times N_i}$, each normalized to have unit ℓ^2 norm. One classical observation from computer vision is that images of the same face under varying illumination lie near a special low-dimensional subspace [6], [38], often called a *face subspace*. So, given a sufficiently expressive training set $\boldsymbol{D}_i$, a new image of subject i taken under different illumination and also stacked as a vector $\boldsymbol{x} \in \mathbb{R}^m$, can be represented as a linear combination of the given training: $\boldsymbol{x} \approx \boldsymbol{D}_i \boldsymbol{\alpha}_i$ for some coefficient vector $\boldsymbol{\alpha}_i \in \mathbb{R}^{N_i}$.

The problem becomes more interesting and more challenging if the identity of the test sample is initially unknown. We define a new matrix D for the entire training set as the concatenation of the $N = \sum_i N_i$ training samples of all c object classes:

$$\boldsymbol{D} \doteq [\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_c] = [\boldsymbol{d}_{1,1}, \boldsymbol{d}_{1,2}, \ldots, \boldsymbol{d}_{k,N_k}]. \quad (1)$$

Then the linear representation of $\boldsymbol{x}$ can be rewritten in terms of all training samples as

$$\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}_0 \quad \in \mathbb{R}^m, \quad (2)$$

where $\boldsymbol{\alpha}_0 = [0, \cdots, 0, \boldsymbol{\alpha}_i^T, 0, \ldots, 0]^T \in \mathbb{R}^N$ is a coefficient vector whose entries are all zero except for those associated with the i-th class. The special support pattern of this coefficient vector is highly informative for recognition: ideally, it precisely identifies the subject pictured. However, in practical face recognition scenarios, the search for such an informative coefficient vector $\boldsymbol{\alpha}_0$ is often complicated by the presence of partial corruption or occlusion: gross errors affect some fraction of the image pixels. In this case, the above linear model (2) should be modified as

$$\boldsymbol{x} \;=\; \boldsymbol{x}_0 + \boldsymbol{e}_0 \;=\; \boldsymbol{D}\,\boldsymbol{\alpha}_0 + \boldsymbol{e}_0, \quad (3)$$

where $\boldsymbol{e}_0 \in \mathbb{R}^m$ is a vector of errors – a fraction, ρ, of its entries are nonzero.

Thus, face recognition in the presence of varying illumination and occlusion can be treated as the search for a certain sparse coefficient vector $\boldsymbol{\alpha}_0$, in the presence of a certain sparse error $\boldsymbol{e}_0$. The number of unknowns in (3) exceeds the number of observations, and we cannot directly solve for $\boldsymbol{\alpha}_0$. However, under mild conditions [28], the desired solution $(\boldsymbol{\alpha}_0, \boldsymbol{e}_0)$ is not only sparse, it is the *sparsest* solution to the system of equations (3):

$$(\boldsymbol{\alpha}_0, \boldsymbol{e}_0) = \arg\min \|\boldsymbol{\alpha}\|_0 + \|\boldsymbol{e}\|_0 \quad \text{subj} \quad \boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{e}. \quad (4)$$

Here, the ℓ^0 "norm" $\|\cdot\|_0$ counts the number of nonzeros in a vector. Originally inspired by theoretical results on equivalence between ℓ^1 and ℓ^0-minimizations [13], [24], in [71] the authors proposed to seek this informative vector $\boldsymbol{\alpha}_0$ by solving the convex relaxation

$$\min \|\boldsymbol{\alpha}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subj} \quad \boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{e}, \quad (5)$$

where $\|\boldsymbol{\alpha}\|_1 \doteq \sum_i |\alpha_i|$. That work reported striking empirical results: the ℓ^1-minimizer, visualized in Figure 1, has a strong tendency to separate the identity of the face (red coefficients) from the error due to corruption or occlusion.

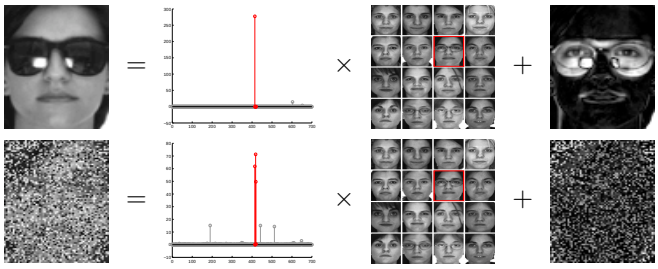


Fig. 1. **Overview of the face recognition approach.** The method represents a test image (left), which is potentially occluded (top) or corrupted (bottom), as a sparse linear combination of all the training images (middle) plus sparse errors (right) due to occlusion or corruption [71]. Red (darker) coefficients correspond to training images of the correct individual. The algorithm determines the true identity (indicated with a red box at second row and third column) from 700 training images of 100 individuals (7 each) in the standard AR face database.

Once the ℓ^1-minimization problem has been solved (see, e.g., [9], [26], [30]), classification (identifying the subject pictured) or validation (determining if the subject is present in the training database) can proceed by considering how strongly the recovered coefficients concentrate on any one subject (see [71] for details). Here, we present only a few representative results; a more thorough empirical evaluation can be found in [71]. Figure 2 (left) compares the recognition rate of this approach (labeled SRC) with several popular methods on the Extended Yale B Database [38] under varying levels of synthetic block occlusion.

Figure 2 compares the sparsity-based approach outlined here with several popular methods from the literature[3]: the Principal Component Analysis (PCA) approach of [67], Independent Component Analysis (ICA) [43], and Local Nonnegative Matrix Factorization (LNMF) [46]. The first provides a standard baseline of comparison, while the latter two methods are more directly suited for occlusion, as they produce lower-dimensional feature sets that are spatially localized. Figure 2 left also compares to the Nearest Subspace method [45], which makes similar use of linear illumination models, but is not based on sparsity and does not correct sparse errors.

The ℓ^1-based approach achieves the highest overall recognition rate of the methods tested, with almost perfect recognition

[2]For a detailed explanation of how such images can be obtained, see [68].

[3]See [77] for a more thorough review of the vast literature on face recognition.

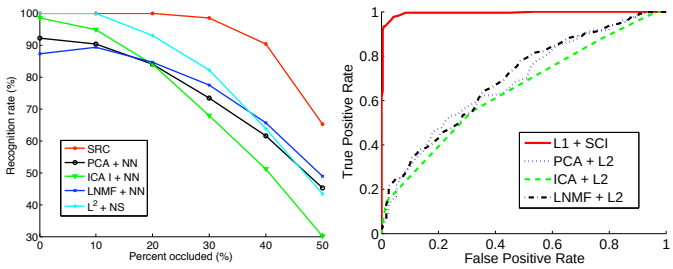


Fig. 2. **Face recognition and validation.** Left: Recognition rate of the ℓ^1-based method (labeled SRC), as well as Principal Component Analysis (PCA) [67], Independent Component Analysis [43], Localized Nonnegative Matrix Factorization (LNMF) [46] and Nearest Subspace (NS) [45] on the Extended Yale B Face Database under varying levels of contiguous occlusion. Right: Receiver Operating Characteristic (ROC) for validation with 30% occlusion. In both scenarios, the sparse representation-based approach significantly outperforms the competitors [71].

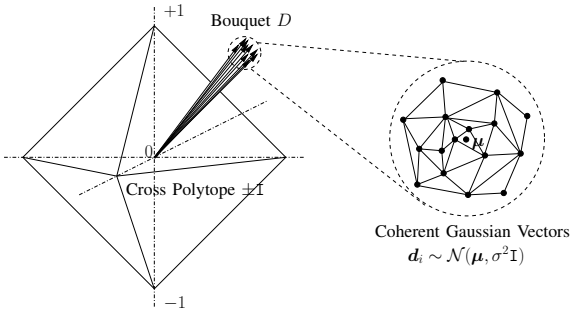


Fig. 3. **The "cross-and-bouquet" model.** Left: the bouquet $\boldsymbol{D}$ and the crosspolytope spanned by the matrix $\pm\mathtt{I}$. Right: tip of the bouquet magnified; it is modeled as a collection of iid Gaussian vectors with small variance σ^2 and common mean vector $\boldsymbol{\mu}$. The cross-and-bouquet polytope is spanned by vertices from both the bouquet $\boldsymbol{D}$ and the cross $\pm\mathtt{I}$ [70].

up to 30% occlusion and a recognition rate above 90% with 40% occlusion. Figure 2 (right) shows the validation performance of the various methods, under 30% contiguous occlusion, plotted as a Reciever Operating Characteristic (ROC) curve. At this level of occlusion, the sparsity-based method is the only one that performs significantly better than chance. The performance under random pixel corruption is even more striking (see Figure 1, bottom), with recognition rates above 90% even at 70% corruption.

B. From Practice to Theory: Dense Error Correction by ℓ^1-Minimization

The strong empirical results alluded to in the previous section seem to demand a correspondingly strong theoretical justification. However, a more thoughtful consideration reveals that the underdetermined system of linear equations (3) does not satisfy popular sufficient conditions for guaranteeing correct sparse recovery by ℓ^1-minimization.

In face recognition, the columns of A are highly correlated: they are all images of *some* face. As m becomes large (i.e. the resolution of the image becomes high), the convex hull spanned by all face images of all subjects is only an extremely tiny portion of the unit sphere $\mathbb{S}^{m-1}$. For example, the images in Figure 1 lie on $\mathbb{S}^{8,063}$. The smallest inner product with their normalized mean is 0.723; they are contained within a spherical cap of volume $\leq 1.47 \times 10^{-229}$. These vectors are tightly bundled together as a "bouquet," whereas the standard pixel basis $\pm\mathtt{I}$ with respect to which we represent the errors $\boldsymbol{e}$ forms a "cross" in $\mathbb{R}^m$, as illustrated in Figure 3. The incoherence [25] and restricted isometry [13] properties that are so useful in providing performance guarantees for ℓ^1-minimization therefore do not hold for the "cross-and-bouquet" matrix $[\boldsymbol{D}\ \mathtt{I}]$ (similarly, conditions that guarantee sparse recovery via greedy techniques such as orthogonal matching pursuit are also often violated by these type of dictionaries). Also, the density of the desired solution is not uniform either: $\boldsymbol{\alpha}$ is usually a very sparse non-negative vector[4], but $\boldsymbol{e}$ could be dense (with a fraction nonzeros close to one) and have arbitrary signs. Existing results for recovering sparse signals suggest that ℓ^1-minimization may have difficulty in dealing with such signals, contrary to its empirical success in face recognition.

In an attempt to better understand the face recognition example outlined above, we consider the more abstract problem of recovering such a non-negative sparse signal $\boldsymbol{\alpha}_0 \in \mathbb{R}^N$ from highly corrupted observations $\boldsymbol{x} \in \mathbb{R}^m$:

$$\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}_0 + \boldsymbol{e}_0,$$

where $\boldsymbol{e}_0 \in \mathbb{R}^m$ is a vector of errors of arbitrary magnitude. The model for $\boldsymbol{D} \in \mathbb{R}^{m\times N}$ should capture the idea that it consists of small deviations about a mean, hence a "bouquet." We can model this by assuming the columns of D are iid samples from a Gaussian distribution:

$$\begin{aligned} \boldsymbol{D} = [\boldsymbol{d}_1 \dots \boldsymbol{d}_N] \in \mathbb{R}^{m\times N}, \quad \boldsymbol{d}_i \sim_{iid} \mathcal{N}\left(\boldsymbol{\mu}, \tfrac{\nu^2}{m}\mathtt{I}_m\right), \\ \|\boldsymbol{\mu}\|_2 = 1, \qquad \|\boldsymbol{\mu}\|_\infty \leq C_\mu m^{-1/2}. \end{aligned} \tag{6}$$

Together, the two assumptions on the mean force $\boldsymbol{\mu}$ to remain incoherent with the standard basis (or "cross") as $m \to \infty$.

We study the behavior of the solution to the ℓ^1-minimization (5) for this model, in the following asymptotic scenario:

Assumption 1 (Weak Proportional Growth): A sequence of signal-error problems exhibits weak proportional growth with parameters $\delta > 0, \rho \in (0,1), C_0 > 0, \eta_0 > 0$, denoted $\mathrm{WPG}_{\delta,\rho,C_0,\eta_0}$, if as $m \to \infty$,

$$\frac{N}{m} \to \delta, \quad \frac{\|\boldsymbol{e}_0\|_0}{m} \to \rho, \quad \|\boldsymbol{\alpha}_0\|_0 \leq C_0\, m^{1-\eta_0}. \tag{7}$$

This should be contrasted with the "total proportional growth" (TPG) setting of, e.g., [24], in which the number of nonzero entries in the signal $\boldsymbol{\alpha}_0$ also grows as a fixed fraction of the dimension. In that setting, one might expect a sharp phase transition in the combined sparsity of $(\boldsymbol{\alpha}_0, \boldsymbol{e}_0)$ that can be recovered by ℓ^1-minimization. In WPG, on the other hand, we observe a striking phenomenon not seen in TPG: the correction of arbitrary fractions of errors. This comes at the expense of the stronger assumption that $\|\boldsymbol{\alpha}_0\|_0$ is sublinear, an assumption that is valid in some real applications such as the face recognition example above.

In the following, we say the cross-and-bouquet model is ℓ^1-*recoverable at* $(I, J, \boldsymbol{\sigma})$ if for all $\boldsymbol{\alpha}_0 \geq 0$ with support I and

[4]The nonnegativity of $\boldsymbol{\alpha}$ can be viewed as a consequence of convex cone models for illumination [38]; the existence of such a solution can be guaranteed by choosing training samples that span the cone of observable test illuminations [68].

$\boldsymbol{e}_0$ with support J and signs $\boldsymbol{\sigma}$,

$$(\boldsymbol{\alpha}_0, \boldsymbol{e}_0) \;=\; \arg\min \; \|\boldsymbol{\alpha}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subject to} \;\; \boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{e} = \boldsymbol{D}\boldsymbol{\alpha}_0 + \boldsymbol{e}_0, \; (8)$$

and the minimizer is uniquely defined. From the geometry of ℓ^1-minimization, if (8) does not hold for some pair $(\boldsymbol{\alpha}_0, \boldsymbol{e}_0)$, then it does not hold for any $(\boldsymbol{\alpha}, \boldsymbol{e})$ with the same signs and support as $(\boldsymbol{\alpha}_0, \boldsymbol{e}_0)$ [23]. Understanding ℓ^1-recoverability at each $(I, J, \boldsymbol{\sigma})$ completely characterizes which solutions to $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{e}$ can be correctly recovered. In this language, the following characterization of the error correction capability of ℓ^1-minimization can be given [70]:

Theorem 1 (Error Correction with the Cross-and-Bouquet): For any $\delta > 0$, $\exists\, \nu_0(\delta) > 0$ such that if $\nu < \nu_0$ and $\rho < 1$, in $\mathrm{WPG}_{\delta,\rho,C_0,\eta_0}$ with $\boldsymbol{D}$ distributed according to (6), if the error support J and signs $\boldsymbol{\sigma}$ are chosen uniformly at random, then as $m \to \infty$,

$$\mathbb{P}_{\boldsymbol{D},J,\sigma}\left[\ell^1\text{-recoverability at } (I, J, \sigma) \;\; \forall\, I \in \binom{[N]}{k_1}\right] \to 1.$$

In other words, as long as the bouquet is sufficiently tight, asymptotically ℓ^1-minimization recovers any non-negative sparse signal from almost any error with support size less than 100% [70]. This provides some theoretical corroboration to the strong practical and empirical results observed in the face recognition example, especially in the presence of random corruption.

C. Remarks on Sparsity-Based Recognition

The theoretical justifications of this approach discussed here have inspired further practical work in this direction. The work reported in [68] addresses issues such as pose and alignment as well as obtaining sufficient training data of each subject, and integrates these results into a practical system for face recognition that achieves state-of-the-art results. Moreover, while in this section we have focused on the interplay between theory and practice in one particular application, face recognition, similar ideas have seen application on a number of problems in and even beyond vision, e.g., in sensor networks and human activity classification [74] as well as speech recognition [36], [37].

Although the cross-and-bouquet model has successfully explained the error correction ability of ℓ^1 minimization in this application, the striking discriminative power of the sparse representation (see also sections III and IV) still lacks rigorous mathematical justification. Better understanding this behavior seems to require a better characterization of the internal structure of the bouquet and its effect on the ℓ^1-minimizer. To the best of our knowledge, this remains a wide open topic for future investigation.

III. ℓ^1-Graphs

The previous section showed how for face recognition, a representation of the test sample in terms of the training samples themselves yielded useful information for recognition. Whereas before, this representation was motivated via linear illumination models, we now consider a more general setting in which an explicit linear model is absent. Here, the sparse coefficients computed by ℓ^1-minimization are used to characterize relationships between the data samples, in order to accomplish various machine learning tasks. The key idea is to accomplish this by interpreting the coefficients as weights in a directed graph, which we term the ℓ^1-*graph* (see also [48] for a graphical model interpretation of the sparse representation approach for image classification described in Section IV).

A. Motivations

An informative graph, directed or undirected, is critical for graph-based machine learning tasks such as data clustering, subspace learning, and semi-supervised learning. Popular spectral approaches to clustering start with a graph representing pairwise relationships between the data samples [61]. Manifold learning algorithms such as ISOMAP [63], Locally Linear Embedding (LLE) [58], and Laplacian Eigenmaps (LE) [8], all rely on graphs constructed with different motivations [73]. Moreover, most popular subspace learning algorithms, *e.g.*, Principal Component Analysis (PCA) [42] and Linear Discriminant Analysis (LDA) [7], can all be explained within the graph embedding framework [73]. Also, a number of semi-supervised learning algorithms are driven by the regularizing graphs constructed over both labeled and unlabeled data [78].

Most of the works described above rely on one of two popular approaches to graph construction: the k-nearest-neighbor method and the ε-ball method. The first assigns edges between each data point and its k-nearest neighbors, whereas the second assigns edges between each data point and all samples within its surrounding ε-ball. From a machine learning perspective, the following graph characteristics are desirable:

1) *High discriminating power*. For data clustering and label propagation in semi-supervised learning, the data from the same cluster/class are expected to be assigned large connecting weights. The graphs constructed in those popular ways however, often fail to capture piecewise linear relationships between data samples in the same class.
2) *Sparsity*. Recent research on manifold learning [8] shows that a sparse graph characterizing locality relations can convey the valuable information for classification. Also for large-scale applications, a sparse graph is the inevitable choice due to storage limitations.
3) *Adaptive neighborhood*. It often happens that the available data are inadequate and do not evenly distribute, resulting in different neighborhood structure for different data points. Both the k-nearest-neighbor and ε-ball methods (in general) use a fixed global parameter to determine the neighborhoods for all the data, and thus do not handle situations where an adaptive neighborhood is required.

Enlightened by recent advances in our understanding of sparse coding by ℓ^1 optimization [24] and in applications such as the face recognition example described in the previous section, we propose to construct the so-called ℓ^1-graph via sparse data coding, and then harness it for popular graph-based machine learning tasks. An ℓ^1 graph over a dataset is derived

by encoding each datum as the sparse representation of the remaining samples, and automatically selects the most informative neighbors for each datum. The sparse representation computed by ℓ^1-minimization naturally satisfies the properties of sparsity and adaptivity. Moreover, we will see empirically that characterizing linear relationships between data samples via ℓ^1-minimization can significantly enhance the performance of existing graph-based learning algorithms.

B. ℓ^1-Graph Construction

We represent the sample set as a matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{m \times N}$, where N is the sample number and m is the feature dimension. We denote the ℓ^1-graph as $G = \{\boldsymbol{X}, W\}$, where $\boldsymbol{X}$ is the vertex set and $W = [w_{ij}] \in \mathbb{R}^{N \times N}$ the edge weight matrix. The graph is constructed in an unsupervised manner, with a goal of automatically determining the neighborhood structure as well as the corresponding connection weights for each datum.

Unlike the k-nearest-neighbor and ε-ball based graphs in which the edge weights characterize pairwise relations, the edge weights of ℓ^1-graph are determined in a group manner, and the weights related to a certain vertex characterize how the rest samples contribute to the sparse representation of this vertex. The procedure to construct the ℓ^1-graph is:

1) **Inputs:** The sample set $\boldsymbol{X}$.
2) **Sparse coding:** For each sample $\boldsymbol{x}_i$, solve the ℓ^1 norm minimization problem

$$\min_{\boldsymbol{\alpha}^i} \|\boldsymbol{\alpha}^i\|_1, \ s.t. \ \ \boldsymbol{x}_i = \boldsymbol{D}^i \boldsymbol{\alpha}^i, \tag{9}$$

where matrix $\boldsymbol{D}^i = [\boldsymbol{x}_1, ..., \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, ..., \boldsymbol{x}_N, \mathtt{I}] \in \mathbb{R}^{m \times (m+N-1)}$ and $\boldsymbol{\alpha}^i \in \mathbb{R}^{m+N-1}$.
3) **Graph weights setting:** $W_{ij} = \alpha^i_j$ (nonnegativity constraints may be imposed if for similarity measurement) if $i > j$, and $W_{ij} = \alpha^i_{j-1}$ if $i < j$.

For data with linear or piecewise-linear class structure the sparse representation conveys important discriminative information, which is automatically encoded in the ℓ^1-graph. The derived graph is naturally sparse – the sparse representation computed by ℓ^1-minimization never involves more than m nonzero coefficients, and may be especially sparse when the data have degenerate or low-dimensional structure. The number of neighbors selected by ℓ^1-graph is adaptive to each data point, and these numbers are automatically determined by the ℓ^1 optimization process. Thus, the ℓ^1-graph possesses all the three characteristics of a desired graph for data clustering, subspace learning, and semi-supervised learning [18], [72].

C. ℓ^1-Graph for Machine Learning Tasks

An informative graph is critical for achieving high performance with graph-based learning algorithms. Similar to conventional graphs constructed by k-nearest-neighbor or ε-ball method, ℓ^1-graph can also be integrated with graph-based algorithms for tasks such as data clustering, subspace learning, and semi-supervised learning. In the following sections, we show how ℓ^1-graphs can be used for each of these purposes.

1) Spectral clustering with ℓ^1-graph: Data clustering is the partitioning of samples into subsets, such that the data within each subset are similar to each other. Some of the most popular algorithms for this task are based on spectral clustering [61]. Using the ℓ^1-graph, the algorithm can automatically derive the similarity matrix from the calculation of these sparse codings (namely $w_{ij} = \alpha^i_j$). Inheriting the property of greater discriminating power from ℓ^1-graph, the spectral clustering based on ℓ^1-graph has greater potential to correctly separate the data into different clusters. Based on the derived ℓ^1-graph, the spectral clustering [61] process can be performed in the same way as for conventional graphs.

2) Subspace learning with ℓ^1-graph: Subspace learning algorithms search for a projection matrix $P \in \mathbb{R}^{m \times d}$ (usually $d \ll m$) such that distances in the projected space are as informative as possible for classification. If the dimension of the projected space is large enough, then linear relationships between the training samples may be preserved, or approximately preserved. The pursuit of a projection matrix that simultaneously respects the sparse representations of all of the data samples can be formulated as an optimization problem (closely related to the problem of metric learning)

$$\min \sum_{i=1}^{N} \left\| P^T \boldsymbol{x}_i - \sum_{j=1}^{N} w_{ij} P^T \boldsymbol{x}_j \right\|_2^2 \quad \text{subj} \quad P^T \boldsymbol{X} \boldsymbol{X}^T P = I \tag{10}$$

and solved via generalized eigenvalue decomposition.

3) Semi-supervised Learning with ℓ^1-graph: Semi-supervised learning has attracted a great deal of recent attention. The main idea is to improve classifier performance by using additional unlabeled training samples to characterize the intrinsic geometry of the observation space (see for example [54] for the application of sparse models for semi-supervised learning problems). For classification algorithms that rely on optimal projections or embeddings of the data, this can be achieved by adding a regularization term to the objective function that forces the embedding to respect the relationships between the unlabeled data.

In the context of ℓ^1-graphs, we can modify the classical LDA criterion to also demand that the computed projection respects the sparse coefficients computed by ℓ^1-minimization:

$$\min_P \frac{\gamma S_w(P) + (1-\gamma) \sum_{i=1}^{N} \|P^T \boldsymbol{x}_i - \sum_{j=1}^{N} w_{ij} P^T \boldsymbol{x}_j\|_2^2}{S_b(P)},$$

where $S_w(P)$ and $S_b(P)$ measure the within-class scatter and inter-class scatter of the labeled data respectively, and $\gamma \in (0, 1)$ is a coefficient that balances the supervised term and ℓ^1-graph regularization term (see also [57]).

D. Experimental Results

In this section, we systematically evaluate the effectiveness of the ℓ^1-graph in the machine learning scenarios outlined above. The USPS handwritten digit database [41] (200 samples are selected for each class), forest covertype database [1] (120 samples are selected for each class), and ETH-80 object recognition database [2] are used for the experiments. Note that all the results reported here are from the best tuning of

all possible algorithmic parameters, and the results on the first two databases are the averages of ten runs while the results on ETH-80 are from one run.

Table I compares the accuracy of spectral clustering based on the ℓ^1-graph with spectral algorithms based on a number of alternative graph constructions, as well as the simple baseline of K-means. The clustering results from ℓ^1-graph based spectral clustering algorithm are consistently much better than the other algorithms tested.

TABLE I
CLUSTERING ACCURACIES (NORMALIZED MUTUAL INFORMATION) FOR SPECTRAL CLUSTERING ALGORITHMS BASED ON ℓ^1-GRAPH, GAUSSIAN-KERNEL GRAPH (G-G), LE-GRAPH (LE-G), AND LLE-GRAPH (LLE-G), AS WELL AS PCA+K-MEANS (PCA+KM).

Cluster #	ℓ^1-graph	G-g	LE-g	LLE-g	PCA+Km
USPS : 7	**0.962**	0.381	0.724	0.565	0.505
FOR. : 7	**0.763**	0.621	0.619	0.603	0.602
ETH. : 7	**0.605**	0.371	0.522	0.478	0.428

Our next experiment concerns data classification based on low-dimensional projections. Table II compares the classification accuracy of the ℓ^1-graph based subspace learning algorithm with several more conventional subspace learning algorithms. The following observations emerge: 1) the ℓ^1-graph based subspace learning algorithm is superior to all the other evaluated unsupervised subspace learning algorithms, and 2) ℓ^1-graph based subspace learning algorithm generally performs a little worse than the *supervised* algorithm Fisherfaces, but on the forest covertype database, ℓ^1-graph based subspace learning algorithm is better than Fisherfaces. Note that all the algorithms are trained on all the data available, and the results are based on nearest neighbor classifier; for all experiments, 10 samples for each class are randomly selected as gallery set and the remaining ones are used for testing.

TABLE II
COMPARISON CLASSIFICATION ERROR RATES (%) FOR DIFFERENT SUBSPACE LEARNING ALGORITHMS. LPP AND NPE ARE THE LINEAR EXTENSIONS OF LE AND LLE RESPECTIVELY.

Gallery #	PCA	NPE	LPP	ℓ^1-graph-SL	Fisherfaces [7]
USPS : 10	37.21	33.21	30.54	21.91	**15.82**
FOR. : 10	27.29	25.56	27.32	**19.76**	21.17
ETH. : 10	47.45	45.42	44.74	38.48	**13.39**

Finally, we evaluate the effectiveness of the ℓ^1 graph in semi-supervised learning scenarios. Table III compares results with the ℓ^1-graph to several alternative graph constructions. We make two observations: 1) the ℓ_1-graph based semi-supervised learning algorithm generally achieves the lowest error rates compared to semi-supervised learning based on more conventional graphs, and 2) semi-supervised learning based on the ℓ^1-graph and the graph used in LE algorithm can generally bring accuracy improvements compared to the counterpart without harnessing extra information from unlabeled data. Note that all the semi-supervised algorithms are based on the supervised algorithm Marginal Fisher Analysis (MFA) [73].

E. Remarks on ℓ^1-Graphs

Although in this section we have illustrated with a few generic examples the potential of ℓ^1-graphs for some general problems in machine learning, the idea of using sparse coefficients computed by ℓ^1-minimization for clustering has already found good success in the classical vision problem of segmenting multiple motions in a video, where low-dimensional self-expressive representations can be motivated by linear camera models. In that domain, algorithms combining sparse representation and spectral clustering also achieve state-of-the-art results on extensive public data sets [33], [56]. Despite apparent empirical successes, precisely characterizing the conditions under which ℓ^1-graphs can better capture certain geometric or statistic relationships among data remains an open problem. We expect many interesting and important mathematical problems may arise from this rich research field. The next section further investigates the use of sparse representations for image classification, including exploiting the sparse coefficients with respect to learned dictionaries.

TABLE III
COMPARISON CLASSIFICATION ERROR RATES (%) FOR SEMI-SUPERVISED ALGORITHMS ℓ^1-GRAPH (ℓ^1-G), LE-GRAPH (LE-G), AND LLE-GRAPH (LLE-G), SUPERVISED (MFA) AND UNSUPERVISED LEARNING (PCA) ALGORITHMS.

Labeled #	ℓ^1-g	LLE-g	LE-g	MFA	PCA
USPS : 10	**25.11**	34.63	30.74	34.63	37.21
FOR. : 10	**17.45**	24.93	22.74	24.93	27.29
ETH. : 10	**30.79**	38.83	34.54	38.83	47.45

IV. DICTIONARY LEARNING FOR IMAGE ANALYSIS

The previous sections examined applications in vision and machine learning in which a sparse representation in an over-complete dictionary consisting of the samples themselves yielded semantic information. For many applications, however, rather than simply using the data themselves, it is desirable to use a compact dictionary that is obtained from the data by optimizing some task-specific objective function. This section provides an overview of approaches to learning such dictionaries, as well as their applications in computer vision and image processing.

A. Motivations

As detailed in the previous sections, *sparse modeling* calls for constructing efficient representations of data as a (often linear) combination of a few typical patterns (atoms) learned from the data itself. Significant contributions to the theory and practice of learning such collections of atoms (usually called dictionaries or codebooks), e.g., [4], [34], [52], and of representing the actual data in terms of them, e.g., [17], [20], [30], have been developed in recent years, leading to state-of-the-art results in many signal and image processing tasks [11], [32], [44], [48], [51], [54]. We refer the reader to [10] for a recent review on the subject.

The actual dictionary plays a critical role, and it has been shown again and again that learned and data adaptive dictionaries significantly outperform off-the-shelf ones such as wavelets. Current techniques for obtaining such dictionaries mostly involve their optimization in terms of the task to be performed, e.g., representation [34], denoising [4], [51], and classification [48]. Theoretical results addressing the stability

and consistency of the sparse solutions (*active set* of selected atoms), as well as the efficiency of the coding algorithms, are related to intrinsic properties of the dictionary such as the mutual coherence, the cumulative coherence, and the Gram matrix norm of the dictionary [28], [31], [40], [59], [66]. Dictionaries can be learned by locally optimizing these and related objectives [29], [55]. In this section, we present basic concepts associated with dictionary learning, and provide illustrative examples of algorithm performance.

B. Sparse Modeling for Image Reconstruction

Let $\boldsymbol{X} \in \mathbb{R}^{m \times N}$ be a set of N column data vectors $\boldsymbol{x}_j \in \mathbb{R}^m$ (e.g., image patches), $\boldsymbol{D} \in \mathbb{R}^{m \times K}$ be a dictionary of K atoms represented as columns $\boldsymbol{d}_k \in \mathbb{R}^m$. Each data vector $\boldsymbol{x}_j$ will have a corresponding vector of reconstruction coefficients $\boldsymbol{\alpha}_j \in \mathbb{R}^K$ (in contrast with the cases described in previous sections, K will now be orders of magnitude smaller than N), which we will treat as columns of a matrix

$$\boldsymbol{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N] \in \mathbb{R}^{K \times N}.$$

The goal of *sparse modeling* is to design a dictionary $\boldsymbol{D}$ such that $\boldsymbol{X} \simeq \boldsymbol{D}\boldsymbol{A}$ with $\|\boldsymbol{\alpha}_j\|_0$ sufficiently small (usually below some threshold) for all or most data samples $\boldsymbol{x}_j$. For a fixed $\boldsymbol{D}$, the computation of $\boldsymbol{A}$ is called *sparse coding*.

We begin our discussion with the standard ℓ^0 *or* ℓ^1 *penalty* modeling problem,

$$(\boldsymbol{A}^*, \boldsymbol{D}^*) = \arg\min_{\boldsymbol{A},\boldsymbol{D}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{A}\|_F^2 + \lambda \|\boldsymbol{A}\|_p, \qquad (11)$$

where $\|\cdot\|_F$ denotes Frobenius norm and $p = 0, 1$. The cost function to be minimized in (11) consists of a quadratic *fitting term* and an ℓ^0 or ℓ^1 *regularization term* for each column of $\boldsymbol{A}$, the balance of the two being defined by the *penalty parameter* λ (this parameter has been studied in [35], [39], [55], [65], [79]). As mentioned above, the ℓ^1 norm can be used as an approximation to ℓ^0, making the problem convex in $\boldsymbol{A}$ while still encouraging sparse solutions [64]. While for reconstruction we found that the ℓ^0 penalty often produces better results, ℓ^1 leads to more stable active sets and is preferred for the classification tasks introduced in the next section. In addition, these costs can be replaced by a (non-convex) Lorentzian penalty function, motivated either by further approximating the ℓ^0 by ℓ^1 [15], or by considering a mixture of Laplacians prior for the coefficients in $\boldsymbol{A}$ and exploiting MDL concepts [55], instead of the more classical Laplacian prior.[5]

Since (11) is not simultaneously convex in $\{\boldsymbol{A}, \boldsymbol{D}\}$, coordinate descent type optimization techniques have been proposed [4], [34]. These approaches have been extended for multiscale dictionaries and color images in [51], leading to state-of-the-art results. See Figure 4 for an example of color image denosing with this approach, and [49], [51] for numerous additional examples, comparisons, and applications in image demosaicing, image inpainting, and image denoising. An example of a learned dictionary is shown in Figure 4 as well ($K = 256$). It is important to note that for image denoising, overcomplete dictionaries are used, $K > m$, and the patch sizes vary from 7×7, $m = 49$, to 20×20, $m = 400$ (in the multiscale case), with a sparsity of about 1/10th of the signal dimension m.

[5]The expression (11) can be derived from a MAP estimation with a Laplacian prior for the coefficients in $\boldsymbol{A}$ and a Gaussian prior for the sparse representation error.

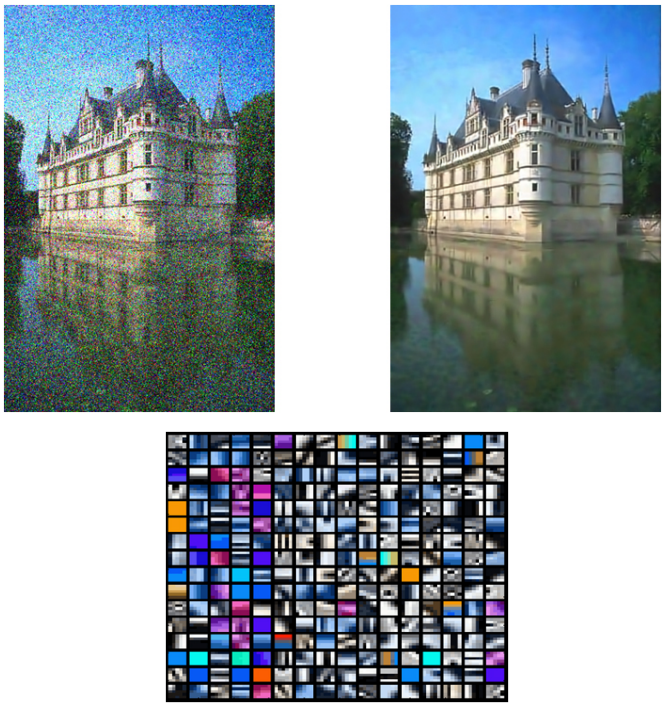


Fig. 4. **Image denoising** via sparse modeling and dictionary learned from a standard set of color images [49].

State-of-the-art results obtained in [51] are "shared" with those in [19], which extends the non-local means approach developed in [5], [12]. Interestingly, the two frameworks are quite related, since they both use patches as building blocks (in [51], the sparse coding is applied to all overlapping image patches), and while a dictionary is learned in [51] from a large dataset, the patches of the processed image itself are the "dictionary" in non-local means. The sparsity constraint in [51] is replaced by a proximity constraint and other processing steps in [12], [19]. The exact relationship and the combination of non-local-means with sparsity modeling has been recently exploited by the authors of [47] to further improve on these results. The authors also developed a very fast on-line dictionary learning approach.

C. Sparse Modeling for Image Classification

While image representation and reconstruction has been the most popular goal of sparse modeling and dictionary learning, other important image science applications are starting to be addressed by this framework, in particular, classification and detection. In [53], [54] the authors use the reconstruction/generative formulation (11), exploiting the quality of the representation and/or the coefficients $\boldsymbol{A}$ for the classification tasks. This generative only formulation can be augmented by discriminative terms [47], [48], [50], [57], [62] where an additional term is added in (11) to encourage the learning of

Fig. 5. **Image classification** via sparse modeling. Two classes have been considered, "bikes" and "background," and the dictionaries where trained in a semi-supervised fashion [47].

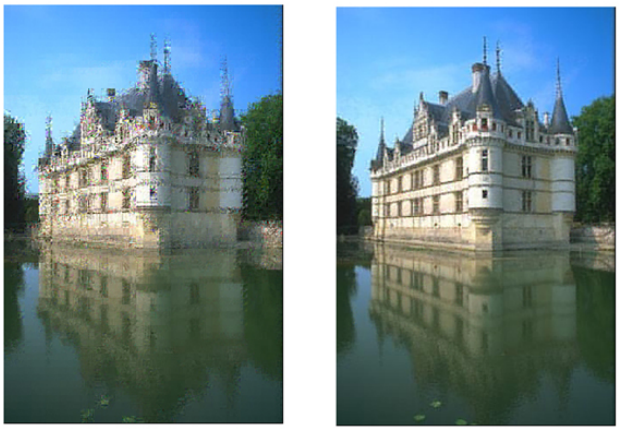


Fig. 6. **Simultaneously learning the dictionary and sensing matrices** (right figure) significantly outperforms classical CS, where for example a random sensing matrix is used in conjunction with an independently learned dictionary (left figure) [29].

dictionaries that are most relevant to the task at hand. The dictionary learning then becomes task-dependent and (semi-) supervised. In the case of [57] for example, a Fisher-discriminant type term is added in order to encourage signals (images) from different classes to pick different atoms from the learned dictionary. In [47], multiple dictionaries are learned, one per class, so that each class's dictionary provides a good reconstruction for its corresponding class and a poor one for the other classes (simultaneous positive and negative learning). This idea was then applied in [50] for learning to detect edges as part of an image classification system. These frameworks have been extended in [48], where a graphical model interpretation and connections with kernel methods are presented as well for the novel sparse model introduced there. Of course, adding such new terms makes the actual optimization even more challenging, and the reader is referred to those papers for details.

This framework of adapting the dictionary to the task, combining generative with discriminative terms for the case of classification, has been shown to outperform the generic dictionary learning algorithms, achieving state-of-the-art results for a number of standard datasets. An example from [47] of the detection of patches corresponding to bikes from the popular Gratz dataset is shown in Figure 5. The reader is referred to [47], [48], [50], [57] for additional examples and comparisons with the literature.

D. Learning to Sense

As we have seen, learning overcomplete dictionaries that facilitate a sparse representation of the data as a liner combination of a few atoms from such dictionary leads to state-of-the-art results in image and video restoration and classification. The emerging area of compressed sensing (CS), see [3], [14], [27] and references therein, has shown that sparse signals can be recovered from far fewer samples than required by the classical Shannon-Nyquist Theorem. The samples used in CS correspond to linear projections obtained by a sensing projection matrix. It has been shown that, for example, a non-adaptive random sampling matrix satisfies the fundamental theoretical requirements of CS, enjoying the additional benefit of universality. A projection sensing matrix that is optimally designed for a certain class of signals can further improve the reconstruction accuracy or further reduce the necessary number of samples. In [29], the authors extended the formulation in (11) to design a framework for the joint design and optimization, from a set of training images, of the non-parametric dictionary and the sensing matrix Φ,

$$(\boldsymbol{A}^*, \boldsymbol{D}^*, \Phi^*) = \arg\min_{\boldsymbol{A},\boldsymbol{D},\Phi} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{A}\|_F^2 + \lambda_1 \|Y - \Phi\boldsymbol{D}\boldsymbol{A}\|_F^2 + \lambda_2 \left\|(\Phi\boldsymbol{D})^T(\Phi\boldsymbol{D}) - I\right\|_F^2 + \lambda_3 \|\boldsymbol{A}\|_p .$$

In this formulation we include the sensing matrix Φ in the optimization, the sensed signal Y obtained from the data $\boldsymbol{X}$ via $Y = \Phi\boldsymbol{X}$, and the critical term that encourages orthogonality of the components of the effective dictionary $\Phi\boldsymbol{D}$, as suggested by the critical restricted isometry property in CS (see [29] for details on the optimization of this functional). This joint optimization outperforms both the use of random sensing matrices and those matrices that are optimized independently of the learning of the dictionary, Figure 6. Particular cases of the proposed framework include the optimization of the sensing matrix for a given dictionary as well as the optimization of the dictionary for a pre-defined sensing environment (see also [31], [60], [69]).

E. Remarks on Dictionary Learning

In this section we briefly discussed the topic of dictionary learning. We illustrated with a number of examples the importance of learning the dictionary for the task as well as the processing and acquisition pipeline. Sparse modeling, and in particular the (semi-) supervised case, can be considered as a non-linear extension of metric learning (see [76] for bibliography on the subject and [62] for details on the connections between sparse modeling and metric learning). Such interesting connection brings yet another exciting aspect into the ongoing sparse modeling developments. The connection with (regression) approaches based on Dirichlet priors, e.g., [22] and references therein, is yet another interesting area for future research.

V. Final Remarks

The examples considered in this paper illustrate several important aspects in the application of sparse representation to problems in computer vision. First, sparsity provides a powerful prior for inference with high-dimensional visual data that have intricate low-dimensional structures. Methods like ℓ^1-minimization offer computational tools to extract such structures and hence help harness the semantics of the data. As we have seen in the few highlighted examples, if properly applied, algorithms based on sparse representation can often achieve state-of-the-art performance. Second, the key to realizing this power is choosing the dictionary in such a way that sparse representations with respect to the dictionary correctly reveal the semantics of the data. This can be done implicitly, by building the dictionary from data with linear or locally linear structure, or explicitly, by optimizing various measures of how informative the dictionary is. Finally, rich data and problems in computer vision provide new examples for the theory of sparse representation, in some cases demanding new mathematical analysis and justification. Understanding the performance of the resulting algorithms can greatly enrich our understanding of both sparse representation and computer vision.

Acknowledgment

JW and YM thank their colleagues on the work of face recognition, A. Ganesh, S. Sastry, A. Yang, A. Wagner, and Z. Zhou and the work on motion segmentation, S. Rao, R. Tron, and R. Vidal. Their work is partially supported by NSF, ONR, and a Microsoft Fellowship.

GS thanks his partners and teachers in the journey of sparse modeling, F. Bach, J. Duarte, M. Elad, F. Lecumberry, J. Mairel, J. Ponce, I. Ramirez, F. Rodriguez, and A. Szlam. J. Duarte, F. Lecumberry, J. Mairal, and I. Ramirez produced the images and results in the dictionary learning section. GS is partially supported by ONR, NGA, NSF, NIH, DARPA, and ARO.

SC thanks Huan Wang, Bin Cheng, and Jianchao Yang for the work of ℓ^1-graph. His work is partially supported by NRF/IDM grant NRF2008IDM-IDM004-029.

The work of TSH was supported in part by IARPA VACE Program.

References

[1] *http://kdd.ics.uci.edu/databases/covertype/covertype.data.html*.
[2] *http://www.vision.ethz.ch/projects/categorization/*.
[3] Compressive sensing resources, http://www.dsp.ece.rice.edu/cs/.
[4] M. Aharon, M. Elad, and A. M. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Trans. SP*, 54(11):4311–4322, November 2006.
[5] S. P. Awate and R. T. Whitaker. Unsupervised, information-theoretic, adaptive image filtering for image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(3):364–376, 2006.
[6] R. Basri and D. Jacobs. Lambertian reflection and linear subspaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(3):218–233, 2003.
[7] P. Belhumeur, J. Hespanda, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.
[8] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 2003.
[9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
[10] 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*, 2008. To appear.
[11] O. Bryt and M. Elad. Compression of facial images using the K-SVD algorithm. *Journal of Visual Communication and Image Representation*, 19:270–283, 2008.
[12] A. Buades, B. Coll, and J. Morel. A review of image denoising algorithms, with a new one. *SIAM Journal of Multiscale Modeling and Simulation*, 4(2):490–530, 2005.
[13] E. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Information Theory*, 51(12), 2005.
[14] E. J. Candès. Compressive sampling. In *Proc. of the International Congress of Mathematicians*, volume 3, Madrid, Spain, 2006.
[15] E. J. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted l1 minimization. *J. Fourier Anal. Appl.*, 2008. To appear.
[16] V. Cevher, A. C. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa. Compressive sensing for background subtraction. In *ECCV*, Marseille, France, 12–18 October 2008.
[17] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.
[18] B. Cheng, J. Yang, S. Yan, and T. Huang. One step beyond sparse coding: Learning with ℓ^1-graph. *under review*.
[19] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Color image denoising by sparse 3d collaborative filtering with grouping constraint in luminance-chrominance space. In *Proc. IEEE International Conference on Image Processing (ICIP)*, San Antonio, Texas, USA, September 2007. to appear.
[20] I. Daubechies, M. Defrise, and C. D. Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. on Pure and Applied Mathematics*, 57:1413–1457, 2004.
[21] M. Dikmen and T. Huang. Robust estimation of foreground in surveillance video by sparse error estimation. In *International Conference on Image Processing*, 2008.
[22] Y. Dong, D. Liu, D. Dunson, and L. Carin. Bayesian multi-task compressive sensing with dirichlet process priors. submitted to *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2009.
[23] D. Donoho. Neighborly polytopes and sparse solution of underdetermined linear equations. *preprint*, 2005.
[24] D. Donoho. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. on Pure and Applied Math*, 59(6):797–829, 2006.
[25] D. Donoho and M. Elad. Optimal sparse representation in general (non-orthogonal) dictionaries via ℓ^1 minimization. *Proceedings of the National Academy of Sciences of the United States of America*, pages 2197–2202, March 2003.
[26] D. Donoho and Y. Tsaig. Fast solution of ℓ^1-norm minimization problems when the solution may be sparse. preprint, *http://www.stanford.edu/ tsaig/research.html*, 2006.
[27] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52:1289–1306, April 2006.
[28] D. L. Donoho and M. Elad. Optimal sparse representation in general (nonorthogoinal) dictionaries via l1 minimization. In *Proc. of the National Academy of Sciences*, volume 100, pages 2197–2202, March 2003.
[29] J. M. Duarte-Carvajalino and G. Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Trans. Image Processing*, 2009, to appear.
[30] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.
[31] M. Elad. Optimized projections for compressed-sensing. *IEEE Trans. SP*, 55(12):5695–5702, Dec. 2007.
[32] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. IP*, 54(12):3736–3745, December 2006.
[33] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.
[34] K. Engan, S. O. Aase, and J. H. Husoy. Frame based signal compression using method of optimal directions (MOD). In *Proc. of the IEEE Intern. Symposium Circuits Syst.*, volume 4, July 1999.
[35] M. A. T. Figueiredo. Adaptive sparseness using Jeffreys prior. In *Adv. NIPS*, pages 697–704, 2001.
[36] J. Gemmeke and B. Cranen. Noise robust digit recognition using sparse representations. In *ISCA ITRW*, 2008.
[37] J. Gemmeke and B. Cranen. Using sparse representations for missing data imputation in noise robust speech recognition. In *EUSIPCO*, 2008.
[38] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. on Pattern Analysis and Machine Intelligence*,

23(6):643–660, 2001.
[39] R. Giryes, Y. C. Eldar, and M. Elad. Automatic parameter setting for iterative shrinkage methods. In *IEEE 25-th Convention of Electronics and Electrical Engineers in Israel (IEEEI'08)*, Dec 2008.
[40] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Trans. Inf. Theory*, 49:3320–3325, 2003.
[41] J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.
[42] I. Joliffe. Principal component analysis. *Springer-Verlag, New York*, 1986.
[43] J. Kim, J. Choi, J. Yi, and M. Turk. Effective representation using ICA for face recognition robust to local distortion and partial occlusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(12):1977–1981, 2005.
[44] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans. PAMI*, 27(6):957–968, 2005.
[45] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.
[46] S. Li, X. Hou, H. Zhang, and Q. Cheng. Learning spatially localized, parts-based representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2001.
[47] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Learning discriminative dictionaries for local image analysis. In *Proc. IEEE CVPR*, 2008.
[48] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Adv. NIPS*, volume 21, 2009.
[49] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Trans. IP*, 17(1):53–69, January 2008.
[50] J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce. Discriminative sparse image models for class-specific edge detection and image interpretation. In *European Conference on Computer Vision*, Marseille, France, 2008.
[51] J. Mairal, G. Sapiro, and M. Elad. Learning multiscale sparse representations for image and video restoration. *SIAM MMS*, 7(1):214–241, April 2008.
[52] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37:3311–3325, 1997.
[53] G. Peyre. Sparse modeling of textures. *Preprint Ceremade 2007-15*, 2007.
[54] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *ICML*, pages 759–766, 2007.
[55] I. Ramirez, F. Lecumberry, and G. Sapiro. Sparse modeling with mixture priors and learned incoherent dictionaries. *pre-print*, 2009.
[56] S. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, and corrupted trajectories. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.
[57] F. Rodriguez and G. Sapiro. Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries. Technical report, University of Minnesota, December 2007. IMA Preprint, www.ima.umn.edu.
[58] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(22):2323–2326, 2000.
[59] K. Schnass and P. Vandergheynst. Dictionary preconditioning for greedy algorithms. *IEEE Trans. SP*, 56(5):1994–2002, 2008.
[60] M. Seeger. Bayesian inference and optimal design in the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.
[61] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
[62] A. Szlam and G. Sapiro. Discriminative k-metrics. *pre-print*, 2009.
[63] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
[64] R. Tibshirani. Regression shrinkage and selection via the LASSO. *JRSS*, 58(1):267–288, 1996.
[65] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.
[66] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. IT*, 50(10):2231–2242, October 2004.
[67] M. Turk and A. Pentland. Eigenfaces for recognition. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 1991.
[68] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma. Toward a practical face recognition system: Robust pose and illumination by sparse representation. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.
[69] Y. Weiss, H. Chang, and W. Freeman. Learning compressed sensing. In *Allerton Conference on Communications, Control and Computing*, 2007.
[70] J. Wright and Y. Ma. Dense error correction via ℓ^1-minimization. preprint, submitted to *IEEE Transactions on Information Theory*, 2008.
[71] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2009.
[72] S. Yan and H. Wang. Semi-supervised learning by sparse representation. *SIAM International Conference on Data Mining, SDM*.
[73] S. Yan, D. Xu, B. Zhang, Q. Yang, H. Zhang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.
[74] A. Yang, R. Jafari, S. Sastry, and R. Bajcsy. Distributed recognition of human actions using wearble motion sensor networks. *Journal of Amibient Intelligence and Sensor Environments*, 2009.
[75] J. Yang, J. Wright, T. Huang, and Y. Ma. Image superresolution as sparse representation of raw patches. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.
[76] L. Yang. Distance metric learning: A comprehensive survey. *http://www.cse.msu.edu/∼yangliu1/frame_survey_v2.pdf*.
[77] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, pages 399–458, 2003.
[78] X. Zhu. Semi-supervised learning literature survey. *Technical Report 1530, Department of Computer Sciences, University of Wisconsin, Madison*, 2005.
[79] H. Zou. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.