INVITED PAPER

Sparse Representation for Computer Vision and Pattern Recognition

A relatively small sample of computer vision and pattern recognition information in applications such as face recognition is often sufficient to reveal the meaning the user desires.

By John Wright, *Member IEEE*, Yi Ma, *Senior Member IEEE*, Julien Mairal, *Member IEEE*, Guillermo Sapiro, *Senior Member IEEE*, Thomas S. Huang, *Life Fellow IEEE*, and Shuicheng Yan, *Senior Member IEEE*

ABSTRACT | Techniques from sparse signal representation are beginning to see significant impact in computer vision, often on nontraditional 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.

KEYWORDS | Compressed sensing; computer vision; pattern recognition; signal representations

Manuscript received March 25, 2009; revised December 29, 2009; accepted February 3, 2010. Date of publication April 29, 2010; date of current version May 19, 2010. The work of J. Wright and Y. Ma was supported in part by NSF, ONR, and a Microsoft Fellowship. The work of G. Sapiro was supported in part by ONR, NGA, NSF, DARPA, and ARO. The work of T. S. Huang was supported in part by IARPA VACE Program. The work of S. Yan was supported in part by the NRF/IDM under Grant NRF2008IDM-IDM004-029.

J. Wright was with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA. He is now with Visual Computing Group, Microsoft Research Asia, Beijing 100190, China (e-mail: jnwright@uiuc.edu).
Y. Ma is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA, and also with the Visual Computing Group, Microsoft Research Asia, Beijing 100190, China (e-mail: yima@uiuc.edu).
J. Mairal is with the INRIA-Willow project, Ecole Normale Supérieure, Laboratoire d'Informatique de l'Ecole Normale Supérieure (INRIA/ENS/CNRS UMR 8548), 75005 Paris, France (e-mail: julien.mairal@m4x.org).
G. Sapiro is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: guille@ece.umn.edu).
T. S. Huang is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: t-huang1@illinois.edu).
S. Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576 (e-mail: eleyans@nus.edu.sg).

Digital Object Identifier: 10.1109/JPROC.2010.2044470

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 [12].

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 [54], [61], [62], [83], [87], [96], image super-resolution [92], motion and data segmentation [37], [68], denoising

and inpainting [15], [56], [60], background modeling [20], [25], photometric stereo [69], and image classification [55], [57], [65], [84], [91], [93]. In almost all of these applications, using sparsity as a prior leads to state-of-the-art results.[1] Sparsity and the design of appropriate dictionaries and projections have influenced the development of both algorithms and of physical imaging systems [6]–[8], [76].

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 is found for the distribution, we should expect that a typical sample has a very sparse representation with respect to such a (possibly learned) basis.[2] Such a sparse representation, if computed correctly, might naturally encode semantic information about 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 into 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 in vision [95]. In this section, we will see how sparse representation and sparse error correction can be used to achieve robust face recognition in scenarios where well-controlled training images can be collected. 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 effective algorithms for face recognition. In turn, the face recognition example reveals new theoretical phenomena in sparse representation that may at first seem surprising.

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. For a detailed explanation of how such images can be obtained, see [83]. We stack the given N_i training images from the ith 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 [9], [42], 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 $\boldsymbol{D}$ for the entire training set as the concatenation of the $N = \sum_i N_i$ training samples of all c object classes

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

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

$$\boldsymbol{x} = \boldsymbol{D}\alpha_0 \in \mathbb{R}^m \quad (2)$$

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 ith 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

$$\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{e}_0 = \boldsymbol{D}\alpha_0 + \boldsymbol{e}_0 \quad (3)$$

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

[1] Although the main focus of this paper is computer vision, similarly unconventional (and noteworthy) applications of sparse representation arise in audio classification [40], [41], [45], bioinformatics [47], and human activity classification [90]. In all of these applications, the choice of dictionary remains critical.

[2] We use the term "basis" loosely here, since the *dictionary* can be overcomplete and its atoms are often not guaranteed to be independent.

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 [32], the desired solution $(\boldsymbol{\alpha}_0, \boldsymbol{e}_0)$ is not only sparse, but also it is the *sparsest* solution to the system of (3)

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

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 [17], [28], the authors of [87] proposed to seek this informative vector $\boldsymbol{\alpha}_0$ by solving the convex relaxation

$$\min \|\boldsymbol{\alpha}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subject to} \quad \boldsymbol{x} = \boldsymbol{D}\alpha + \boldsymbol{e} \tag{5}$$

where $\|\boldsymbol{\alpha}\|_1 \doteq \sum_i |\alpha_i|$. That work demonstrated empirically an interesting tendency of the ℓ^1-minimizer: as visualized in Fig. 1, sparse representation separates the identity of the face (red coefficients) from the error due to corruption or occlusion.

Once the ℓ^1-minimization problem has been solved (see, e.g., [16], [30], [46], and [80]), 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 [87] for details). Here, we present only a few representative results; a more thorough empirical evaluation can be found in [87]. Fig. 2(left) compares the recognition rate of this approach (labeled SRC) with several popular methods on the Extended Yale B Database [42] under varying levels of synthetic block occlusion.

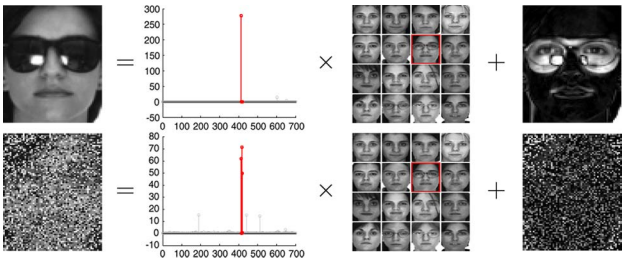

Fig. 1. *Overview of the face recognition approach. The method represents a test image (left), which is potentially occluded (top) or corrupted (bottom), as a sparse linear combination of all the training images (middle) plus sparse errors (right) due to occlusion or corruption. Red (darker) coefficients 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 (seven each) in the standard AR face database.*

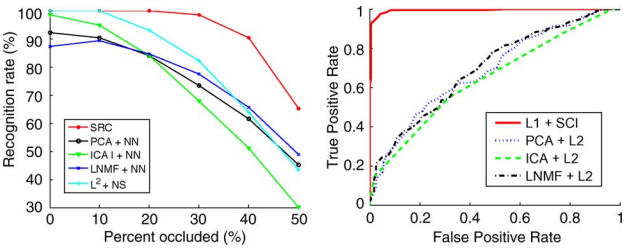

Fig. 2. *Face recognition and validation. (Left) Recognition rate of the ℓ^1-based method (labeled SRC), as well as principal component analysis (PCA) [82], independent component analysis [50], localized nonnegative matrix factorization (LNMF) [53], and nearest subspace (NS) [52] on the extended Yale 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 [87].*

Fig. 2 compares the sparsity-based approach outlined here with several popular methods from the literature[3]: the principal component analysis (PCA) approach of [82], independent component analysis (ICA) architecture I [50], and local nonnegative matrix factorization (LNMF) [53]. The first method 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. Fig. 2(left) also compares to the nearest subspace method [52], which makes similar use of linear illumination models, but does not correct sparse errors.

The ℓ^1-based approach achieves the highest overall recognition rate of the methods tested, with almost perfect recognition up to 30% occlusion and a recognition rate above 90% with 40% occlusion. Fig. 2(right) shows the validation performance of the various methods, under 30% contiguous occlusion, plotted as a receiver 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 also strong [see Fig. 1(bottom)], with recognition rates above 90% even at 70% corruption.

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

The 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 equation

[3]See [95] for a more thorough review of the vast literature on face recognition.

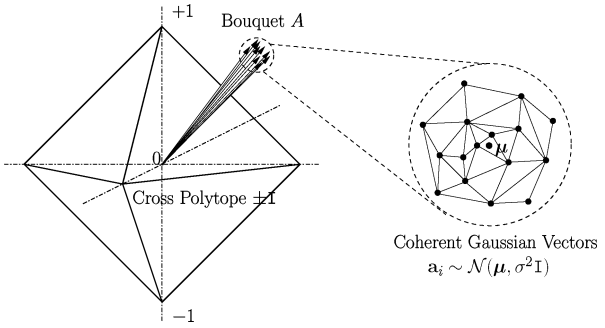

Fig. 3. *The "cross-and-bouquet" model. (Left) The bouquet D and the cross-polytope spanned by the matrix $\pm\mathtt{I}$. (Right) Tip of the bouquet magnified; it is modeled as a collection of i.i.d. Gaussian vectors with small variance σ^2 and common mean vector μ. The cross-and-bouquet polytope is spanned by vertices from both the bouquet D and the cross $\pm\mathtt{I}$ [86].*

(3) does not satisfy popular sufficient conditions for guaranteeing correct sparse recovery by ℓ^1-minimization.

In face recognition, the columns of $\boldsymbol{D}$ 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 Fig. 1 lie on $\mathbb{S}^{8063}$. 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 Fig. 3. The incoherence [29] and restricted isometry [17] properties that are so useful in providing performance guarantees for ℓ^1-minimization therefore do not hold for the "cross-and-bouquet" matrix $[\boldsymbol{D}\ \mathtt{I}]$. Also, the density of the desired solution is not uniform either: $\boldsymbol{\alpha}$ is usually a very sparse nonnegative 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 nonnegative sparse signal $\boldsymbol{\alpha}_0 \in \mathbb{R}^N$ from highly corrupted observations $\boldsymbol{x} \in \mathbb{R}^m$

$$\boldsymbol{x} = \boldsymbol{D}\alpha_0 + \boldsymbol{e}_0$$

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 $\boldsymbol{D}$ are independent identically distributed (i.i.d.) samples from a Gaussian distribution

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

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 $\text{WPG}_{\delta,\rho,C_0,\eta_0}$, 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., [28], 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 weak proportional growth (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 that 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 $\boldsymbol{e}_0$ with support J and signs $\boldsymbol{\sigma}$

$$(\boldsymbol{\alpha}_0, \boldsymbol{e}_0) = \arg\min \|\boldsymbol{\alpha}\|_1 + \|\boldsymbol{e}\|_1$$
$$\text{subject to} \quad \boldsymbol{D}\alpha + \boldsymbol{e} = \boldsymbol{D}\alpha_0 + \boldsymbol{e}_0 \tag{8}$$

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 $(\boldsymbol{\alpha}_0, \boldsymbol{e}_0)$ [27]. Understanding ℓ^1-recoverability at each $(I, J, \boldsymbol{\sigma})$ completely characterizes which solutions to $\boldsymbol{x} = D\boldsymbol{\alpha} + \boldsymbol{e}$ can be correctly recovered.

[4]The nonnegativity of $\boldsymbol{\alpha}$ can be viewed as a consequence of convex cone models for illumination [42]; the existence of such a solution can be guaranteed by choosing training samples that span the cone of observable test illuminations [83].

In this language, the following characterization of the error correction capability of ℓ^1-minimization can be given [86].

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 $m \to \infty$

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

In other words, as long as the bouquet is sufficiently tight, asymptotically ℓ^1-minimization recovers any nonnegative sparse signal from almost any error with support size less than 100% [86]. This provides some theoretical corroboration to the strong 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 [83] addresses issues such as registration and alignment as well as obtaining sufficient training data of each subject, and integrates these results into a practical system for face recognition. However, it is important to realize that this work aims at scenarios such as access control where the training data can be controlled:

> The face recognition approach described here assumes that the training images have been carefully controlled and that the number of samples per class is sufficiently large. Outside these operating conditions, and in particular when only a single sample per class is available, it should not be expected to perform well.

This work does not address the problem of face recognition from unconstrained training, which arises in applications in personal photo organization and image search.

Although the cross-and-bouquet model explains much of the error correction ability of ℓ^1 minimization, 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 [57] 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 semisupervised learning. Popular spectral approaches to clustering start with a graph representing pairwise relationships between the data samples [74]. Manifold learning algorithms such as ISOMAP [77], locally linear embedding (LLE) [71], and Laplacian eigenmaps (LEs) [11] all rely on graphs constructed with different motivations [89]. Moreover, most popular subspace learning algorithms, e.g., PCA [49] and linear discriminant analysis (LDA) [10], can all be explained within the graph embedding framework [89]. Also, a number of semisupervised learning algorithms are driven by the regularizing graphs constructed over both labeled and unlabeled data [97].

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 semisupervised 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 [11] 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 [28] 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 data set is derived by representing each datum as a sparse linear combination 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 by $G = \{V, W\}$. Here, V is the set of N vertices, each of which is identified with a sample in $\boldsymbol{X}$, and $W = [w_{ij}] \in \mathbb{R}^{N \times N}$ is 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 as follows.

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

$$\min_{\boldsymbol{\alpha}^i} \|\boldsymbol{\alpha}^i\|_1 \quad \text{subject to} \quad \boldsymbol{x}_i = \boldsymbol{D}^i \boldsymbol{\alpha}^i \tag{9}$$

where matrix $\boldsymbol{D}^i = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_N, \mathrm{I}] \in \mathbb{R}^{m \times (m+N-1)}$ and $\boldsymbol{\alpha}^i \in \mathbb{R}^{m+N-1}$.
3) **Graph weight setting**: Set $w_{ij} = \alpha^i_j$ (nonnegativity constraints may be imposed for α^i_j in optimization 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 semisupervised learning [22], [88].

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 methods, ℓ^1-graph can also be integrated with graph-based algorithms for tasks such as data clustering, subspace learning, and semisupervised 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 [74]. 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 [74] 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)

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

and solved via generalized eigenvalue decomposition.

3) Semisupervised Learning With ℓ^1-Graph: Semisupervised learning has attracted a great deal of recent attention. The main idea is to improve classifier performance by

Table 1 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
USPS : 7	**0.962**	0.381	0.724
FOR. : 7	**0.763**	0.621	0.593
ETH. : 7	**0.605**	0.371	0.522

Table 2 Classification Error Rates (In Percent) 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 [10]
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**

using additional unlabeled training samples to characterize the intrinsic geometry of the observation space (see, for example, [66] for the application of sparse models for semisupervised 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

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

where $S_w(P)$ and $S_b(P)$ measure the within-class scatter and the interclass scatter of the labeled data, respectively, and $\gamma \in (0,1)$ is a coefficient that balances the supervised term and the ℓ^1-graph regularization term (see also [70]).

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 [48] (200 samples are selected for each class), forest covertype database [1] (100 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 1 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.

Our next experiment concerns data classification based on low-dimensional projections. Table 2 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, ten samples for each class are randomly selected as gallery set and the remaining ones are used for testing.

Finally, we evaluate the effectiveness of the ℓ^1-graph in semisupervised learning scenarios. Table 3 compares results with the ℓ^1-graph to several alternative graph constructions. We make two observations: 1) the ℓ_1-graph-based semisupervised learning algorithm generally achieves the lowest error rates compared to semisupervised learning based on more conventional graphs, and 2) semisupervised 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 semisupervised algorithms are based on the supervised algorithm marginal Fisher analysis (MFA) [89].

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,

Table 3 Classification Error Rates (In Percent) for Semisupervised 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

algorithms combining sparse representation and spectral clustering also achieve state-of-the-art results on extensive public data sets [37], [68]. 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.

IV. DICTIONARY LEARNING FOR IMAGE ANALYSIS

The previous sections examined applications in vision and machine learning in which a sparse representation in an overcomplete dictionary consisting of the samples themselves yielded semantic information. This is an extremely useful idea for clustering and classification, especially for problems such as face recognition and motion segmentation where the data have linear or piecewise linear structure. However, for applications such as inpainting or denoising, the identity of the given training samples is less important—they only serve as a means to an end. Moreover, in applications such as general image classification, it is less clear that images in one class should follow a single linear model. In such applications, it may be possible to learn more relevant dictionaries by optimizing a task-specific objective function. Such dictionaries have the added advantage of often being much more compact than the original training set, allowing more efficient online processing. This section provides an overview of approaches to learning such dictionaries, as well as their many 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], [38], and [63], and of representing the actual data in terms of them, e.g., [21], [24], and [34], have been developed in recent years, leading to state-of-the-art results in many signal and image processing tasks [13], [36], [51], [57], [60], [66]. We refer the reader to [12] for a recent review on the subject.

The actual dictionary plays a critical role, and it has been shown again and again that learned 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 [38], denoising [4], [60], and classification [57]. 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 [32], [35], [44], [72], [81]. Dictionaries can be learned by locally optimizing these and related objectives [33], [67]. 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), and $\boldsymbol{D} \in \mathbb{R}^{m\times K}$ 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 $\boldsymbol{\alpha}_j \in \mathbb{R}^K$, 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 $\boldsymbol{X} \simeq \boldsymbol{DA}$ 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

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

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 [39], [43], [67], [79], and [98]). 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 [78]. 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 (nonconvex) Lorentzian penalty function, motivated either by further approximating the ℓ^0 by ℓ^1 [19], or by considering a mixture of Laplacians prior for the coefficients in $\boldsymbol{A}$ and exploiting MDL concepts [67], instead of the more classical Laplacian prior.[5]

[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.

Since (11) is not simultaneously convex in $\{\boldsymbol{A}, \boldsymbol{D}\}$, coordinate-descent-type optimization techniques have been proposed [4], [38]. These approaches have been extended for multiscale dictionaries and color images in [60], leading to state-of-the-art results. See Fig. 4 for an example of color image denosing with this approach, and [58] and [60] for numerous additional examples, comparisons, and applications in image demosaicing, image inpainting, and image denoising. An example of a learned dictionary is shown in Fig. 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 one tenth of the signal dimension m.

State-of-the-art results obtained in [60] are "shared" with those in [23], which extends the nonlocal means approach developed in [5] and [14]. Interestingly, the two frameworks are quite related, since they both use patches as building blocks (in [60], the sparse coding is applied to all overlapping image patches), and while a dictionary is learned in [60] from a large data set, the patches of the processed image itself are the "dictionary" in nonlocal means. The sparsity constraint in [60] is replaced by a proximity constraint and other processing steps in [14] and [23]. The exact relationship and the combination of nonlocal means with sparsity modeling has been recently exploited by Mairal *et al.* [55] to further improve on these results. They also developed a very fast online dictionary learning approach.

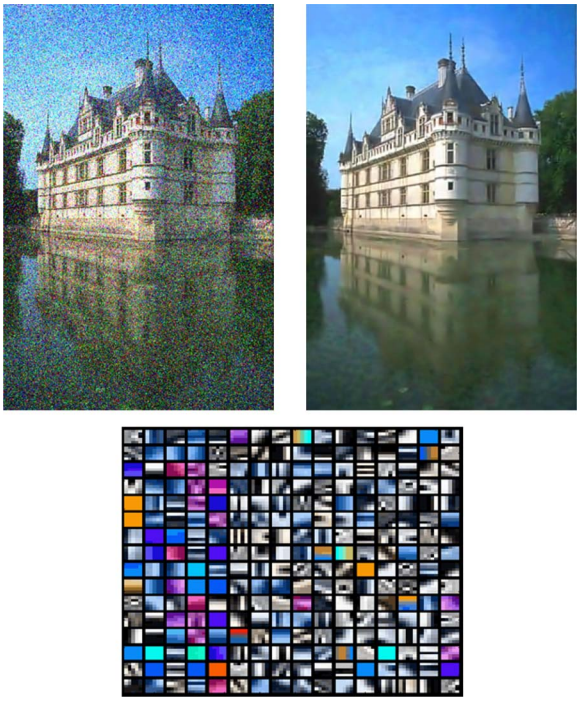

Fig. 4. *Image denoising via sparse modeling and dictionary learned from a standard set of color images.*

Fig. 5. *Image classification via sparse modeling. Two classes have been considered, "bikes" and "background," and the dictionaries where trained in a semisupervised fashion.*

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 [64] and [66], 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 [55], [57], [59], [70], [75] where an additional term is added in (11) to encourage the learning of dictionaries that are most relevant to the task at hand. The dictionary learning then becomes task dependent and (semi)supervised. In the case of [70], 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 [55], 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 [59] for learning to detect edges as part of an image classification system. These frameworks have been extended in [57], 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 data sets. An example from [55] of the detection of patches corresponding to bikes from the popular Gratz data set is shown in

Fig. 5. The reader is referred to [55], [57], [59], and [70] 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], [18], [31] 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 nonadaptive 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 [33], 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 nonparametric 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 [33] 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 (Fig. 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 predefined sensing environment (see also [35], [73], and [85]).

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 nonlinear extension of metric learning (see [94] for bibliography on the subject and [75] 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., [26] and references therein, is yet another interesting area for future research.

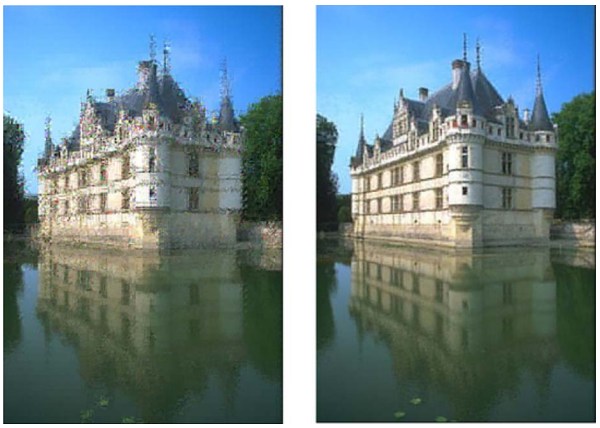

Fig. 6. *Simultaneously learning the dictionary and sensing matrices (right) significantly outperforms classical CS, where, for example, a random sensing matrix is used in conjunction with an independently learned dictionary (left).*

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

J. Wright and Y. Ma would like to thank their colleagues on the work of face recognition, A. Ganesh, S. Sastry, A. ang, A. Wagner, and Z. Zhou, and on the work on

motion segmentation, S. Rao, R. Tron, and R. Vidal. G. Sapiro would like to thank his partners and teachers in the journey of sparse modeling: F. Bach, J. Duarte, M. Elad, F. Lecumberry, J. Ponce, I. Ramirez, F. Rodriguez, and A. Szlam. Special thanks go to J. Duarte, F. Lecumberry, and I. Ramirez for producing the images in the dictionary learning section. S. Yan would like to thank H. Wang, B. Cheng, and J. Yang for the work of ℓ^1-graph.

REFERENCES

[1] J. Blackard, D. Dean, and C. Anderson, "Forest coverType dataset," 1998. [Online]. Available: http://kdd.ics.uci.edu/databases/covertype/covertype.data.html

[2] B. Liebe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2003, pp. 409–415. [Online]. Available: http://tahiti.mis.informatik.tu-darmstadt.de/oldmis/Research/Projects/categorization/eth80-db.html

[3] *Compressive Sensing Resources*, Rice University, Accessed 2010. [Online]. Available: 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. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[5] S. P. Awate and R. T. Whitaker, "Unsupervised, information-theoretic, adaptive image filtering for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 364–376, Mar. 2006.

[6] P. Baheti and M. Neifeld, "Feature-specific structured imaging," *Appl. Opt.*, vol. 45, no. 28, pp. 7382–7391, 2006.

[7] P. Baheti and M. Neifeld, "Adaptive feature specific imaging: A face recognition example," *Appl. Opt.*, vol. 47, no. 10, pp. 821–831, 2008.

[8] P. Baheti and M. Neifeld, "Random projections based feature-specific structured imaging," *Opt. Exp.*, vol. 16, no. 3, pp. 1764–1776, 2009.

[9] R. Basri and D. Jacobs, "Lambertian reflection and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 3, pp. 218–233, Mar. 2003.

[10] P. Belhumeur, J. Hespanda, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[11] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, pp. 1373–1396, 2003.

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

[13] O. Bryt and M. Elad, "Compression of facial images using the K-SVD algorithm," *J. Vis. Commun. Image Represent.*, vol. 19, pp. 270–283, 2008.

[14] A. Buades, B. Coll, and J. Morel, "A review of image denoising algorithms, with a new one," *SIAM J. Multiscale Model. Simul.*, vol. 4, no. 2, pp. 490–530, 2005.

[15] J. Cai, H. Ji, X. Liu, and Z. Shen, "Blind motion deblurring from a single image using sparse approximation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 104–111.

[16] E. Candes and J. Romberg, ℓ^1-magic: Recovery of sparse signals via convex programming. [Online]. Available: http://www.acm.caltech.edu/l1magic/

[17] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[18] E. J. Candès, "Compressive sampling," in *Proc. Int. Congr. Mathematicians*, Madrid, Spain, 2006, vol. 3, pp. 1433–1452.

[19] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ^1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5–6, pp. 877–905, 2008.

[20] V. Cevher, A. C. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa, "Compressive sensing for background subtraction," presented at the Eur. Conf. Comput. Vis., Marseille, France, Oct. 12–18, 2008.

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

[22] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. Huang, "Learning with ℓ^1-graph for image analysis," *IEEE Trans. Image Process.*, 2010.

[23] 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 Int. Conf. Image Process.*, San Antonio, TX, Sep. 2007, pp. 313–316.

[24] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, pp. 1413–1457, 2004.

[25] M. Dikmen and T. Huang, "Robust estimation of foreground in surveillance video by sparse error estimation," in *Proc. 19th Int. Conf. Pattern Recognit.*, 2008. DOI: 10.1109/ICPR.2008.4761910.

[26] Y. Dong, D. Liu, D. Dunson, and L. Carin, "Bayesian multi-task compressive sensing with Dirichlet process priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009.

[27] D. Donoho, "Neighborly polytopes and sparse solution of underdetermined linear equations," 2005. preprint.

[28] D. Donoho, "For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–829, 2006.

[29] D. Donoho and M. Elad, "Optimal sparse representation in general (non-orthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Acad. Sci.*, pp. 2197–2202, Mar. 2003.

[30] D. Donoho and Y. Tsaig, Fast solution of ℓ^1-norm minimization problems when the solution may be sparse, 2006, preprint. [Online]. Available: http://www.stanford.edu/tsaig/research.html

[31] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[32] D. L. Donoho and M. Elad, "Optimal sparse representation in general (nonorthogoinal) dictionaries via l1 minimization," in *Proc. Nat. Acad. Sci.*, Mar. 2003, vol. 100, pp. 2197–2202.

[33] J. M. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1395–1408, Jul. 2009.

[34] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Stat.*, vol. 32, no. 2, pp. 407–499, 2004.

[35] M. Elad, "Optimized projections for compressed-sensing," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5695–5702, Dec. 2007.

[36] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 54, no. 12, pp. 3736–3745, Dec. 2006.

[37] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2790–2797.

[38] K. Engan, S. O. Aase, and J. H. Husoy, "Frame based signal compression using method of optimal directions (MOD)," in *Proc. IEEE Int. Symp. Circuits Syst.*, Jul. 1999, vol. 4, DOI: 10.1109/ISCAS.1999.779928.

[39] M. A. T. Figueiredo, "Adaptive sparseness using Jeffreys prior," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2001, pp. 697–704.

[40] J. Gemmeke and B. Cranen, "Noise robust digit recognition using sparse representations," in *Int. Speech Commun. Assoc. Tutorial Res. Workshop*, 2008.

[41] J. Gemmeke and B. Cranen, "Using sparse representations for missing data imputation in noise robust speech recognition," in *Proc. EUSIPCO*, 2008. [Online]. Available: http://www.eurasip.org/Proceedings/Eusipco/Eusipco2008/papers/1569102857.pdf

[42] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[43] R. Giryes, Y. C. Eldar, and M. Elad, "Automatic parameter setting for iterative shrinkage methods," in *Proc. IEEE 25th Conv. Electron. Electr. Eng. Israel*, Dec. 2008, pp. 820–824.

[44] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.

[45] R. Grosse, R. Raina, H. Kwong, and A. Ng, "Shift-invariant sparse coding for audio classification," in *Proc. Conf. Uncertainty Artif. Intell.*, 2007.

[46] E. Hale, W. Yin, and Y. Zhang, "Fixed point continuation for ℓ_1-minimizatoin: Methodology and convergence," *SIAM J. Optim.*, vol. 19, no. 3, pp. 1107–1130, 2008.

[47] X. Hang and F. Wu, "Sparse representation for classification of tumors using gene

expression data," *J. Biomed. Biotechnol.*, 2009, Article ID 403689, 6 p.

[48] J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994.

[49] I. Joliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[50] 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. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1977–1981, Dec. 2005.

[51] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 957–968, Jun. 2005.

[52] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[53] S. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 1, pp. 207–212.

[54] X. Li, T. Jia, and R. Zhang, "Expression-insensitive 3D face recognition using sparse representation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2575–2582.

[55] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Learning discriminative dictionaries for local image analysis," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2008, DOI: 10.1109/CVPR.2008.4587652.

[56] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009.

[57] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Advances in Neural Information Processing Systems*, vol. 21, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2009.

[58] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, Jan. 2008.

[59] J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce, "Discriminative sparse image models for class-specific edge detection and image interpretation," presented at the Eur. Conf. Comput. Vis., Marseille, France, Oct.12–18, 2008.

[60] J. Mairal, G. Sapiro, and M. Elad, "Learning multiscale sparse representations for image and video restoration," *SIAM Multiscale Modeling and Simulation*, vol. 7, no. 1, pp. 214–241, Apr. 2008.

[61] X. Mei, H. Ling, and D. Jacobs, "Sparse representation of cast shadows via l1-regularized least squares," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009.

[62] P. Nagesh and B. Li, "A compressive sensing approach for expression-invariant face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1518–1525.

[63] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vis. Res.*, vol. 37, pp. 3311–3325, 1997.

[64] G. Peyre, "Sparse modeling of textures," 2007, *preprint ceremade 2007-15*.

[65] A. Quattoni, M. Collins, and T. Darrell, "Transfer learning for image classification with sparse prototype representation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2008. DOI: 10.1109/CVPR.2008.4587637.

[66] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 759–766.

[67] I. Ramirez, F. Lecumberry, and G. Sapiro, Sparse modeling with universal priors and learned incoherent dictionaries. [Online]. Available: http://www.ima.umn.edu/preprints/sep2009/2279.pdf

[68] 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 *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2008, DOI: 10.1109/CVPR.2008.4587437.

[69] D. Reddy, A. Agrawal, and R. Chellappa, "Enforcing integrability by error correction using ℓ_1-minimization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2350–2357.

[70] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries," Univ. Minnesota, Minneapolis, MN, Tech. Rep./IMA Preprint, Dec. 2007.

[71] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 22, pp. 2323–2326, 2000.

[72] K. Schnass and P. Vandergheynst, "Dictionary preconditioning for greedy algorithms," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1994–2002, May 2008.

[73] M. Seeger, "Bayesian inference and optimal design in the sparse linear model," *J. Mach. Learn. Res.*, vol. 9, pp. 759–813, 2008.

[74] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[75] A. Szlam and G. Sapiro, "Discriminative k-metrics," 2009, *preprint*.

[76] D. Takhar, J. Laska, M. Wakin, M. Duarte, D. Baron, S. Sarvotham, K. Kelly, and R. Baraniuk, "A new compressive imaging camera architecture using optical domain compression," in *Proc. Comput. Imaging IV SPIE Electron. Imaging*, 2006, pp. 43–52.

[77] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[78] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. R. Stat. Soc.*, vol. 58, no. 1, pp. 267–288, 1996.

[79] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[80] J. Tropp and S. Wright, "Computational methods for sparse solution of linear inverse problems," California Inst. Technol., Pasadena, CA, ACM Tech. Rep. 2009-1, 2009.

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

[82] M. Turk and A. Pentland, "Face recognition using Eigenfaces," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 1991, pp. 586–591.

[83] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, "Towards a practical face recognition system: Robust registration and illumination by sparse representation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 597–604.

[84] C. Wang, S. Yan, L. Zhang, and H. Zhang, "Multi-label sparse coding for automatic image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1643–1650.

[85] Y. Weiss, H. Chang, and W. Freeman, "Learning compressed sensing," in *Proc. Allerton Conf. Commun. Control Comput.*, 2007. [Online]. Available: http://www.cs.huji.ac.il/~yweiss/allerton-final.pdf

[86] J. Wright and Y. Ma, "Dense error correction via ℓ^1-minimization," *IEEE Trans. Inf. Theory*, 2009.

[87] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 210–227, Feb. 2009.

[88] S. Yan and H. Wang, "Semi-supervised learning by sparse representation," in *SIAM Int. Conf. Data Mining*, pp. 792–801.

[89] S. Yan, D. Xu, B. Zhang, Q. Yang, H. Zhang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[90] A. Yang, R. Jafari, S. Sastry, and R. Bajcsy, "Distributed recognition of human actions using wearable motion sensor networks," *J. Ambient Intell. Sensor Environ.*, vol. 1, no. 2, pp. 103–115, 2009.

[91] A. Yang, S. Maji, K. Hong, P. Yan, and S. Sastry, "Distributed compression and fusion of nonnegative sparse signals for multiple-view object recognition," in *Proc. Int. Conf. Inf. Fusion*, 2009, pp. 1867–1874.

[92] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image superresolution as sparse representation of raw patches," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2008. DOI: 10.1109/CVPR.2008.4587647.

[93] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1794–1801.

[94] L. Yang, Distance metric learning: A comprehensive survey. [Online]. Available: http://www.cse.msu.edu/~yangliu1/frame_survey_v2.pdf

[95] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, pp. 399–458, 2003.

[96] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma, "Face recognition with contiguous occlusion using markov random fields," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009.

[97] X. Zhu, "Semi-supervised learning literature survey," Dept. Comput. Sci., Univ. Wisconsin, Madison, WI, Tech. Rep. 1530, 2005.

[98] H. Zou, "The adaptive LASSO and its oracle properties," *J. Amer. Stat. Assoc.*, vol. 101, pp. 1418–1429, 2006.

ABOUT THE AUTHORS

John Wright (Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, in 2009.

Currently, he is a Researcher with the Visual Computing group at Microsoft Research Asia, Beijing, China. His graduate work focused on developing efficient and provably correct algorithms for error correction with high-dimensional data, and on their application in automatic face recognition. His research interests encompass a number of topics in vision and signal processing, including minimum description length methods for clustering and classification, error correction and inference with nonideal data, video analysis and tracking, as well as face and object recognition.

Dr. Wright received a number of awards and honors, including a UIUC Distinguished Fellowship, Carver Fellowship, Microsoft Research Fellowship, the UIUC Martin Award for Outstanding Graduate Research, and the Lemelson-Illinois Prize for Innovation.

Yi Ma (Senior Member, IEEE) received two B.S. degrees in automation and applied mathematics from Tsinghua University, Beijing, China, in 1995 and the M.S. degree in electrical engineering and computer science, the M.A. degree in mathematics, and the Ph.D. degree in electrical engineering and computer science from the University of California at Berkeley, in 1997, 2000, and 2000, respectively.

Currently, he is an Associate Professor at the Electrical and Computer Engineering Department, University of Illinois at Urbana-Champaign, Urbana. Since January 2009, he has also been Research Manager of the Visual Computing group, Microsoft Research Asia, Beijing, China. His main research interest is in computer vision, high-dimensional data analysis, and systems theory. He is the first author of the popular vision textbook *An Invitation to 3-D Vision* (New York: Springer-Verlag, 2003).

Dr. Ma received the David Marr Best Paper Prize at the International Conference on Computer Vision 1999, the Longuet-Higgins Best Paper Prize at the European Conference on Computer Vision 2004, and the Sang Uk Lee Best Student Paper Award with his students at the Asian Conference on Computer Vision in 2009. He also received the CAREER Award from the National Science Foundation in 2004 and the Young Investigator Award from the Office of Naval Research in 2005. He is an associate editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and has served as the chief guest editor for special issues for the PROCEEDINGS OF IEEE and the IEEE SIGNAL PROCESSING MAGAZINE. He will also serve as Program Chair for the 2013 International Conference on Computer Vision, Sydney, Australia. He is a member of the Association for Computing Machinery (ACM), the Society for Industrial and Applied Mathematics (SIAM), and the American Society for Engineering Education (ASEE).

Julien Mairal (Member, IEEE) received the graduate degree from Ecole Polytechnique and Ecole Nationale Supérieure des Téléécommunications, Paris, France, in 2005 and the M.S. degree from the Ecole Normale Supérieure, Cachan, France, in 2007. He is currently working towards the Ph.D. degree under the supervision of J. Ponce and F. Bach at Ecole Normale Supérieure, Paris, France.

His research interests include machine learning, computer vision, and image processing.

Guillermo Sapiro (Senior Member, IEEE) was born in Montevideo, Uruguay, on April 3, 1966. He received the B.Sc. (*summa cum laude*), M.Sc., and Ph.D. degrees from the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, Israel, in 1989, 1991, and 1993, respectively.

After postdoctoral research at Massachusetts Institute of Technology (MIT), Cambridge, he became Member of Technical Staff at the research facilities of HP Labs in Palo Alto, CA. He is currently with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, where he holds the position of the Distinguished McKnight University Professor and Vincentine Hermes-Luh Chair in Electrical and Computer Engineering. He works on differential geometry and geometric partial differential equations, both in theory and applications in computer vision, computer graphics, medical imaging, and image analysis. He has authored and coauthored numerous papers in this area and has written a book *Geometric Partial Differential Equations and Image Analysis* published by Cambridge University Press, January 2001.

Dr. Sapiro recently coedited a special issue of the IEEE TRANSACTIONS ON IMAGE PROCESSING and a second one in the *Journal of Visual Communication and Image Representation*. He was awarded the Gutwirth Scholarship for Special Excellence in Graduate Studies in 1991, the Ollendorff Fellowship for Excellence in Vision and Image Understanding Work in 1992, the Rothschild Fellowship for Post-Doctoral Studies in 1993, the Office of Naval Research Young Investigator Award in 1998, the Presidential Early Career Awards for Scientist and Engineers (PECASE) in 1998, and the National Science Foundation Career Award in 1999. He is a member of the Society for Industrial and Applied Mathematics (SIAM). He is the funding Editor-in-Chief of the *SIAM Journal on Imaging Sciences*.

Thomas S. Huang (Life Fellow, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1956 and the M.S. (in 1960) and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He was on the Faculty of the Department of Electrical Engineering, MIT, from 1963 to 1973 and on the Faculty of the School of Electrical Engineering and Director of its Laboratory for Information and Signal Processing at Purdue University from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, Urbana, where he is now William L. Everitt Distinguished Professor of Electrical and Computer Engineering, Research Professor at the Coordinated Science Laboratory, and Technology and Co-Chair of the Institute's major research theme Human Computer Intelligent Interaction at the Beckman Institute for Advanced Science. His professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 21 books and over 600 papers in network theory, digital filtering, image processing, and computer vision.

Dr. Huang is a member of the National Academy of Engineering, a member of the Academia Sinica, Republic of China, a foreign member of the Chinese Academies of Engineering and Sciences, and a Fellow of the International Association of Pattern Recognition of the IEEE and the Optical Society of America. Among his many honors and awards are Honda Lifetime Achievement Award, IEEE Jack Kilby Signal Processing Medal, and the KS Fu Prize of the International Association for Pattern Recognition.

Shuicheng Yan (Senior Member, IEEE) received the Ph.D. degree from the School of Mathematical Sciences, Peking University, Beijing, China, in 2004.

He spent three years as a Postdoctoral Fellow at the Chinese University of Hong Kong, Hong Kong, and then at the University of Illinois at Urbana-Champaign, Urbana. Currently, he is an Assistant Professor at the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. In recent years, his research interests have focused on computer vision (biometrics, surveillance, and internet vision), multimedia (video event analysis, image annotation, and media search), machine learning (feature extraction, sparsity/nonnegativity analysis, and large-scale machine learning), and medical image analysis. He has authored or coauthored over 140 technical papers over a wide range of research topics.

Dr. Yan has served on the editorial board of the *International Journal of Computer Mathematics*. He also served as Guest Editor of the special issue for *Pattern Recognition Letters*, and has been serving as the Guest Editor of the special issue for *Computer Vision and Image Understanding*. He has served as Co-Chair of the IEEE International Workshop on Video-Oriented Object and Event Classification (VOEC'09) held in conjunction with the 2009 International Conference on Computer Vision (ICCV'09). He is the Special Session Chair of the 2010 Pacific-Rim Symposium on Image and Video Technology, and the Local Arrangement Chair of the 2010 IEEE International Conference on Multimedia & Expo. He is an associate editor of the IEEE Transactions on Circuits and Systems for Video Technology.