On the Role of Sparse and Redundant Representations in Image Processing

Michael Elad, *Senior Member, IEEE,* Mário A.T. Figueiredo, *Senior Member, IEEE,* and Yi Ma, *Senior Member, IEEE*

Invited Paper

***Abstract*—Much of the progress made in image processing in the past decades can be attributed to better modeling of image content, and a wise deployment of these models in relevant applications. This path of models spans from the simple ℓ_2-norm smoothness, through robust, thus edge preserving, measures of smoothness (e.g. total variation), and till the very recent models that employ sparse and redundant representations.**

In this paper, we review the role of this recent model in image processing, its rationale, and models related to it. As it turns out, the field of image processing is one of the main beneficiaries from the recent progress made in the theory and practice of sparse and redundant representations. We discuss ways to employ these tools for various image processing tasks, and present several applications in which state-of-the-art results are obtained.

I. Introduction

A Close inspection of the progress made in the field of image processing in the past several decades reveals that much of it is a direct consequence of the better *image modeling* employed. Armed with a stronger and more reliable model, one can better handle applications ranging from sampling, denoising, restoration, and reconstruction in inverse problems, all the way to compression, detection, separation, and beyond. Indeed, the evolution of models for visual data is at the heart of the image processing literature.

What is a model and why do we need one? We provide an initial answer to these questions through a simple example of noise removal from an image. Given a noisy image, a denoising algorithm is essentially required to separate the noise form the (unknown) clean image. Such a separation clearly requires a close familiarity with the characteristics of both the noise and the original image. Knowing that the noise is additive, white, and Gaussian (AWG) is a good start, but far from being sufficient, since the underlying image may also behave like such noise, thereby making the separation of the two impossible. The additional information on the clean image content, that will allow separating it from the AWG noise, constitutes what we refer to in this paper as an *image model*. A classic example of such a model is the intuitive assumption that near-by pixels in "well-behaved" images exhibit strong correlation; i.e., natural images tend to be spatially (piecewise) smooth.

There is a long path of models and their usage in image processing. This path spans from the simple ℓ_2-norm of local differences (expressing smoothness), through robust and thus edge preserving measures of smoothness, such as the total variation [11], [48]. On a different track, it was observed that sparsity of the wavelet coefficients could be used as a reasonable image model [41], [43], [49], and this concept has been used frequently in the past decade. More recently, improved versions of this idea bring us to sparse and redundant representation modeling. Recent work on various image processing applications indicate that models based on sparse and redundant representations lead to state-of-the-art results, and encompass a persuasive potential to this field.

In this paper we explore the role in image processing of models based on sparse and redundant representations and their rationale. We also relate them to other models that are used successfully in image processing, such as the local-PCA (principal component analysis), and example-based techniques. We discuss ways to employ sparse and redundant representations in image processing tasks in general, and then concentrate on several key applications where this model is shown to lead to state-of-the-art results. More specifically, we discuss image denoising, image inpainting, image deblurring, and super-resolution reconstruction. We conclude this paper with a wish-list for this field, describing the research frontiers of this important and challenging arena of work.

II. Historical Perspective on Image Modeling

We start our discussion by motivating the quest for a model for images, then gradually build a case for a model that is based on sparse and redundant representations.

Consider a family of signals – a set of vectors $\mathcal{X} = \{\boldsymbol{x}_j,\ j = 0, 1, 2...\}$, such that $\boldsymbol{x}_j \in \mathbb{R}^n$. To make our discussion more concrete, we shall assume hereafter that each such signal is a $\sqrt{n} \times \sqrt{n}$ pixels image, representing natural and typical image content, with the corresponding vectors being obtained by lexicographically stacking the image pixels. While such images are very diverse vectors in $\mathbb{R}^n$, we have no reason to assume that they occupy the entire space. Said more accurately, assuming that the pixels in $\boldsymbol{x} \in \mathcal{X}$ have values in the range $[0, 1)$, these images do not populate or sample the hyper-cube $[0, 1)^n \subset \mathbb{R}^n$ uniformly. For example, as mentioned above, spatially smooth images occur much more

Michael Elad is with the Department of Computer Science, The Technion–Israel Institute of Technology, Haifa, Israel; email: *elad@cs.technion.ac.il*.

Mário Figueiredo is with the Instituto de Telecomunicações and the Department of Electrical and Computer Engineering, Instituto Superior Técnico, Lisboa, Portugal; email: *mario.figueiredo@lx.it.pt*.

Yi Ma is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, USA; email: *yima@uiuc.edu*.

often than highly non-smooth and disorganized images. This line of reasoning naturally leads to the Bayesian framework of imposing a probability density function (PDF) on the images – a 'prior' distribution $P(\boldsymbol{x})$ [44]. Priors are extensively used in image processing, serving in inverse problems, compression, anomaly detection, and more, namely because they provide a systematic way of measuring the probability of an image.

Returning to the image denoising example, consider a given image $\boldsymbol{y}$, known to be a noisy version of a clean image $\boldsymbol{x}$, contaminated by an additive perturbation $\boldsymbol{v}$, i.e. $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v}$. Assuming that $\boldsymbol{v}$ has finite energy $\|\boldsymbol{v}\|_2 \leq \epsilon$, the unknown image $\boldsymbol{x}$ must be in the sphere $\|\boldsymbol{x}-\boldsymbol{y}\|_2 \leq \epsilon$. The optimization problem

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \; P(\boldsymbol{x}) \quad \text{subject to} \quad \|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \epsilon \tag{1}$$

leads to the most probable image $\hat{\boldsymbol{x}}$ in this sphere, which is an effective estimate of $\boldsymbol{x}$. This way the prior is exploited for solving the denoising problem.

Much effort has been allocated in the image processing community to forming adequate priors. One very common way to construct $P(\boldsymbol{x})$ is to guess its structure based on intuitive expectations on the data content. For example, the Gibbs distribution $P(\boldsymbol{x}) \propto \exp\{-\lambda\|\boldsymbol{L}\boldsymbol{x}\|_2^2\}$ uses a Laplacian matrix (defined as the linear space-invariant operation that applies the Laplacian filter to the image $\boldsymbol{x}$) to give an evaluation of the probability of the image $\boldsymbol{x}$. In such a prior, deviation from spatial smoothness, as measured by the Laplacian operator, is used for judging the probability of the signal. This prior is well-known and extensively used in image processing, and is known to be related to both *Tikhonov regularization* and *Wiener filtering* [2]. It is also worth mentioning that this is a particular instance of the more general class of image models known as Markov random fields [31].

The above smoothness prior is known to cause image over-smoothing, when used in various image enhancement and restoration tasks. The culprit is the squared ℓ^2-norm in the exponent of that prior, which strongly penalizes (i.e., makes highly unlikely) any large local differences such as edges, which are key features for visual perception [5]. One remedy for this problem was found to be the replacement of the squared ℓ^2-norm by a more robust measure, such as the ℓ^1-norm; by penalizing less any large values, the resulting probability density on $\boldsymbol{L}\boldsymbol{x}$ is allowed to have heavy tails. Thus, a prior of the form $P(\boldsymbol{x}) \propto \exp\{-\lambda\|\boldsymbol{L}\boldsymbol{x}\|_1\}$ was found to be more adequate, thus became popular in recent years. Similar in spirit is the total-variation prior that also promotes smoothness, but differently, by replacing the Laplacian with gradient norms, thereby using first derivatives rather than second ones. Interestingly, the adoption of the ℓ^1-norm measure is known to lead to an enforcement of sparsity of the signal/image derivatives.

Another property that can be used for constructing an image prior is sparsity of the wavelet transform, as was experimentally observed in natural images [43], [49]. The orthogonal wavelet transform of an image $\boldsymbol{x}$ is given by $\boldsymbol{T}\boldsymbol{x}$, where $\boldsymbol{T}$ is a specially designed orthogonal matrix ($\boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{T}\boldsymbol{T}^T = \boldsymbol{I}$) containing in its rows spatial derivatives of varying scale, thereby providing what is known as "multi-scale" analysis of the signal [41]. Therefore, the prior in this case becomes $P(\boldsymbol{x}) \propto \exp\{-\lambda\|\boldsymbol{T}\boldsymbol{x}\|_p^p\}$ with $p \leq 1$ to promote sparsity, and known to adequately model natural images. Here, the resemblance to the total variation and the Laplacian priors is evident, as in all these cases derivatives of some sort and a robust measure are combined in forming the prior $P(\boldsymbol{x})$.

One may adopt a common point of view towards these priors, regarding them as mere attempts to describe a random generator machine $\mathcal{M}$ that supposedly generates the signals of interest. This brings us to sparse and redundant representation modeling, which is a way of synthesizing signals according to a prior defined on the coefficients of the representation.

III. Sparse and Redundant Representation Modeling

A. The Model

Consider the linear system $\boldsymbol{D}\boldsymbol{\alpha} = \boldsymbol{x}$, where $\boldsymbol{D} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{\alpha} \in \mathbb{R}^m$, and interpret it as a way of constructing an image $\boldsymbol{x}$. Each of the m columns of $\boldsymbol{D}$ is a possible image in $\mathbb{R}^n$ – we refer to these columns as atomic images and to the matrix $\boldsymbol{D}$ as a dictionary of atoms. One can interpret $\boldsymbol{D}$ as the periodic table of the fundamental elements in the chemistry that describes our images.

The multiplication of $\boldsymbol{D}$ by a sparse vector $\boldsymbol{\alpha}$ with $k_0 \ll m$ non-zeros produces a linear combination of k_0 atoms with varying weights, generating an image $\boldsymbol{x}$. The sparsity of $\boldsymbol{\alpha}$ can be measured by the ℓ_0^0 "norm" $\|\boldsymbol{\alpha}\|_0^0$, which is the limit of ℓ_p^p as $p \to 0$. In fact, $\lim_{p\to 0}\|\boldsymbol{\alpha}\|_p^p = |\{i,\ \alpha_i \neq 0\}|$, the number of non-zeros components of $\boldsymbol{\alpha}$. We shall refer to the vector $\boldsymbol{\alpha}$ that generates $\boldsymbol{x}$ as its representation, since it describes which atoms and what "portions" thereof were used for its construction. This process of linearly combining atoms to form an image (think of it as a molecule in the signal chemistry) may be referred to as *atom composition*.

Consider the set of all the possible placements of $k_0 \ll m$ non-zeros, which has cardinality $\binom{m}{k_0} = m!/(k_0!(m-k_0)!)$. Assume that samples from this set are drawn with uniform probability. Assume further that, given their positions, each non-zero entry in $\boldsymbol{\alpha}$ is drawn independently from the zero-mean, τ-variance, Gaussian distribution $\mathcal{N}(0,\tau)$. This gives us a complete definition of the PDF of $\boldsymbol{\alpha}$, as well as a probabilistic generative model for images $\boldsymbol{x}$ – this constitutes the random signal generator $\mathcal{M}(\boldsymbol{D}, k_0, \tau)$. Note that the image $\boldsymbol{x}$ is the outcome of a mixture of Gaussians, each of dimension k_0, and with equal probabilities.

An important addition to the definition of the model $\mathcal{M}$ could be postulating a random perturbation (noise) vector $\boldsymbol{e} \in \mathbb{R}^n$ with bounded power $\|\boldsymbol{e}\|_2 \leq \epsilon$, such that $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}+\boldsymbol{e}$. Such an additive perturbation may be useful because the model $\mathcal{M}(\boldsymbol{D}, k_0, \tau)$ imposes a too strong restriction on the images, and using it in practical applications will necessarily lead to a mismatch between actual observed images and their imposed model. With the introduction of this perturbation, which enables such a mismatch, we refer hereafter to the model proposed as $\mathcal{M}(\boldsymbol{D}, k_0, \tau, \epsilon)$.

B. The Basics on Using This Model

How do we practice image processing with sparse and redundant representation modeling? Suppose we have an

image $\boldsymbol{x}$, assumed to have been generated by the model $\mathcal{M}(\boldsymbol{D}, k_0, \tau, \epsilon)$, and that the parameters of the model are known. There are numerous image processing tasks that could be of interest to us. As shown next, common to these is the need to solve a problem we refer to hereafter as $\mathcal{P}_0(\boldsymbol{D}, \boldsymbol{x}, \delta)$, which has the form

$$\mathcal{P}_0(\boldsymbol{D}, \boldsymbol{x}, \delta): \ \min_{\boldsymbol{\alpha}} \ \|\boldsymbol{\alpha}\|_0^0 \quad \text{subject to } \|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\alpha}\|_2 \leq \delta. \quad (2)$$

Solving this problem corresponds to finding the sparsest vector $\boldsymbol{\alpha}$ that explains $\boldsymbol{x}$ as a linear combination of columns from $\boldsymbol{D}$, with weight vector $\boldsymbol{\alpha}$, with an error no larger than δ. This process is known as *atomic decomposition* [12].

How would one go about solving $\mathcal{P}_0(\boldsymbol{D}, \boldsymbol{x}, \delta)$? A direct solution is prohibitive, as it requires a combinatorial search of exponential size for considering all possible supports. Various practical alternative techniques for approximating the solution of this problem have been proposed in recent years. Some rely on a relaxation that replaces $\|\boldsymbol{\alpha}\|_0^0$ by $\|\boldsymbol{\alpha}\|_1$, yielding the problem $\mathcal{P}_1(\boldsymbol{D}, \boldsymbol{x}, \delta)$:

$$\mathcal{P}_1(\boldsymbol{D}, \boldsymbol{x}, \delta): \ \min_{\boldsymbol{\alpha}} \ \|\boldsymbol{\alpha}\|_1 \quad \text{subject to } \|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\alpha}\|_2 \leq \delta. \quad (3)$$

Another class of methods adopts greedy schemes in which the unknown support is obtained sequentially; detailed descriptions and references may be found in [6]. The important thing to mention in this respect is the fact that a thorough theoretical study leads to guarantees for such approximation methods to perform well (see [6] and the many references therein), thus making the solution of $\mathcal{P}_0(\boldsymbol{D}, \boldsymbol{x}, \delta)$ accessible.

Returning to image processing tasks, here are few key core examples where sparse and redundant representation modeling is used for handling commonly encountered problems:

- **Compression**: Nominally, $\boldsymbol{x}$ requires a description by n numbers. However, if we can solve $\mathcal{P}_0(\boldsymbol{D}, \boldsymbol{x}, \delta)$, for some $\delta \geq \epsilon$, then the resulting solution, denoted as $\boldsymbol{\alpha}_0^\delta$, affords an approximation $\hat{\boldsymbol{x}} = \boldsymbol{D}\boldsymbol{\alpha}_0^\delta$ to $\boldsymbol{x}$ using at most k_0 scalars, with an approximation error at most δ. Notice that if $\boldsymbol{x}$ was generated by $\mathcal{M}(\boldsymbol{D}, k_0, \tau, \epsilon)$, it is guaranteed that there exists a $\boldsymbol{x}_0 = \boldsymbol{D}\boldsymbol{\alpha}_0$, where $\|\boldsymbol{\alpha}_0\|_0^0 = k_0$, and such that $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \leq \varepsilon \leq \delta$. By increasing δ we obtain a deeper compression with fewer non-zeros, and a larger approximation error. This way (and ignoring quantization issues), we obtain a rate-distortion curve for a compression mechanism.
- **Denoising and Linear Inverse Problems**: Consider a noisy indirect measurement of $\boldsymbol{x}$, i.e., $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{v}$, where $\boldsymbol{v}$ is an additive noise known to obey $\|\boldsymbol{v}\|_2 \leq \delta$. The operator $\boldsymbol{H}$ could be the identity (in which case the problem reduces to denoising), it could represent a blur, a tomographic projection, the masking of some pixels (which leads to inpainting), down-sampling (which leads to super-resolution), a random set of projections (which leads to compressed sensing; see [10] and references therein), or any other kind of linear degradation. By solving $\mathcal{P}_0(\boldsymbol{H}\boldsymbol{D}, \boldsymbol{y}, \epsilon + \delta)$, the resulting solution $\boldsymbol{\alpha}_0^{\delta+\epsilon}$ will have at most k_0 nonzeros. A theoretical analysis establishes that if k_0 is small enough, then $\boldsymbol{\alpha}_0^{\epsilon+\delta}$ is at most $\mathcal{O}((\epsilon+\delta)k_0/n)$ away from the original representation $\boldsymbol{\alpha}$, implying a very effective reconstruction [9]. Thus, we can expect to identify directly the sparse components of the underlying signal $\boldsymbol{x}$ and obtain an approximation/estimate $\hat{\boldsymbol{x}} = \boldsymbol{D}\boldsymbol{\alpha}_0^{\delta+\epsilon}$.
- **Morphological Component Analysis (MCA)**: Suppose that the observed signal is a superposition of two different signals, i.e., $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, where $\boldsymbol{x}_1$ is sparsely generated using model $\mathcal{M}(\boldsymbol{D}_1, k_1, \tau, \epsilon_1)$ and $\boldsymbol{x}_2$ is sparsely generated using model $\mathcal{M}(\boldsymbol{D}_2, k_2, \tau, \epsilon_2)$. Can we separate the two superimposed signals? For example, can we separate the texture and cartoon contents of an image, given that both are sparsely generated by two quite different dictionaries? Observe that $\boldsymbol{x}$ can be modeled has having been generated by $\mathcal{M}([\boldsymbol{D}_1, \boldsymbol{D}_2], k_1 + k_2, \tau, \epsilon_1 + \epsilon_2)$, where $[\boldsymbol{D}_1, \boldsymbol{D}_2]$ denotes the concatenation of the two dictionaries. Thus it makes sense to solve the problem $\mathcal{P}_0([\boldsymbol{D}_1, \boldsymbol{D}_2], \boldsymbol{x}, \epsilon_1 + \epsilon_2)$; the solution $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ allows generating a plausible answer to the separation problem: $\hat{\boldsymbol{x}}_1 = \boldsymbol{D}_1\boldsymbol{\alpha}_1$ and $\hat{\boldsymbol{x}}_2 = \boldsymbol{D}_2\boldsymbol{\alpha}_2$ [26].

A wide range of other applications can also be envisioned. All these applications call for the solution of $\mathcal{P}_0(\boldsymbol{D}, \boldsymbol{x}, \delta)$, or variants of it and, as mentioned above, these can be approximated reliably by practical algorithms.

C. Geometric Interpretation and Relation to Other Models

In order to gain a geometric insight into the proposed model and some alternatives to it, we return to the beginning of this section, considering a large corpus of examples of images (or image patches) $\mathcal{X} = \{\boldsymbol{x}_j, \ j = 0, 1, 2, ...\} \subset \mathbb{R}^n$. We concentrate on one arbitrary image $\boldsymbol{x}_0 \in \mathcal{X}$ and its δ-neighborhood, and aim to study the behavior of this neighborhood in the n-th dimensional space $\mathbb{R}^n$.

For a small enough value of δ, moving δ away from $\boldsymbol{x}_0$ along directions $\boldsymbol{e} = \boldsymbol{x} - \boldsymbol{x}_0$, where $\boldsymbol{x} \in \mathcal{X}$, represent small permissible perturbation directions that lead to feasible signals. The question we pose is: do those fill the entire $\delta-$ball in $\mathbb{R}^n$ space? We denote this set of permissible directions as $\Omega_{\boldsymbol{x}_0}^\delta$. Gathering all vectors $\boldsymbol{e} \in \Omega_{\boldsymbol{x}_0}^\delta$ into an $n \times |\Omega_{\boldsymbol{x}_0}^\delta|$ matrix $\boldsymbol{E}_{\boldsymbol{x}_0}$, we aim to study the behavior of its singular values. We will be interested in families of structured images, for which the effective rank of such matrices is $k_{\boldsymbol{x}_0} \ll n$, for any $\boldsymbol{x}_0 \in \mathcal{X}$. This is equivalent to the statement that the $k_{\boldsymbol{x}_0}+1$-th singular-value and beyond in these matrices tend to be near-zero and are therefore negligible, and the effective directions in the obtained subspace can be accumulated as an orthogonal set of $k_{\boldsymbol{x}_0}$ columns in a $n \times k_{\boldsymbol{x}_0}$ matrix $\boldsymbol{Q}_{\boldsymbol{x}_0}$.

Signals satisfying the above local low-dimensionality assumption essentially exhibit a local behavior that is approximately a linear subspace of dimension $k_{\boldsymbol{x}_0}$, shifted around $\boldsymbol{x}_0$. While this subspace and its dimension may vary from one point to another, all of them are characterized as being far smaller than the ambient n dimensions [34]. Experimental studies show that most informative signals we work with, such as images, audio, and more, follow such a structure, and this can be exploited in processing them [54], [37], [38].

Based on the above description, it is tempting to build a model of a signal source by holding many instances of

example signals $\mathcal{X}$, and using these directly to characterize these local subspaces; this is known as a local PCA modeling [54]. In this approach, the columns of $\boldsymbol{Q}_{\boldsymbol{x}_0}$ are the principal directions of the signal for the location $\boldsymbol{x}_0$, and those are permitted to vary freely as a function of $\boldsymbol{x}_0$. In general, a complete model of this sort requires either a storage of $\mathcal{X}$ and extraction of $\boldsymbol{Q}_{\boldsymbol{x}_0}$ per need, or a storage of all the possible unitary bases, $\boldsymbol{Q}_{\boldsymbol{x}}$, gathered off-line for every $\boldsymbol{x} \in \mathcal{X}$. Both options are prohibitive in most cases.

If we can further assume that a small number of such matrices, $\{\boldsymbol{Q}_p, \quad p = 1, ..., P\}$ (say, a few hundreds or thousands), covers all possible cases, this model can be made more efficient. This calls for a (fuzzy-) clustering of the signals $\boldsymbol{x} \in \mathcal{X}$ to subgroups that correspond each to a different matrix $\boldsymbol{Q}_p$. As the range of each matrix $\boldsymbol{Q}_p$ is a low-dimensional subspace, we essentially use an arrangement of many low-dimensional subspaces to approximate the distribution of all signals $\mathcal{X}$. Such a model has been proposed in recent years as a generalization to principal component analysis [54], and has shown to be very effective in grasping the behavior of image content, by working on small patches [34], [37], [38].

The sparse and redundant representation model is one step further in an attempt to make such local model more concise. Assuming for simplicity that all these local subspaces are of the same dimensionality, k_0, using sparse representations of cardinality k_0 over a dictionary $\boldsymbol{D}$ with m columns, we have at our disposal $\binom{m}{k_0}$ subspaces, just by allowing k_0-sparse representations with all the possible supports. This way, our dictionary holds very efficiently all the key directions required to represent the signal anywhere in $\mathbb{R}^n$, while enabling a very large number of possibilities. Clearly, though, this comes at the price of further restricting the structure of the model by assuming that the atoms in the dictionary are principal directions shared by many sub-spaces.

For completeness of this presentation, we mention that rather than describe the model of the signal by constructing parameters that characterize its local behavior, one could use the local neighbors directly. This is known as a direct example-based modeling, and its use for texture synthesis [20], inpainting, denoising [7], and other applications show promising results. Such a direct technique is very appealing, but it works well only if the sampling of the signal source is dense enough. This approach could be considered as an extreme sparse representation modeling of the signal $\boldsymbol{x}$, such that the δ-local neighborhood serves as its local (and varying) dictionary. A related approach models image patches as being probably close to a low dimensional manifold [47].

IV. Image Processing Applications

We now turn to discuss specific applications where sparse and redundant representation modeling is shown to be highly effective. We should note that in migrating from the above core ideas to applications, there is much room for imagination and creativity, as a direct deployment of the outlined ideas will not necessarily operate well. Thus, as we show next, each of the applications described below take a different route in using the model proposed, choosing the dictionary to work with, adopting the numerical scheme to approximate sparse representation, and more.

A. Image Denoising

We obtain an observed image $\boldsymbol{y}$, a noisy version of an unknown underlying clean image $\boldsymbol{x}$, i.e., $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v}$, and our goal is to recover $\boldsymbol{x}$. A popular noise model is to consider $\boldsymbol{v} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. In line with the discussion in Section III-B, denoising could be performed by solving $\mathcal{P}_0(\boldsymbol{D}, \boldsymbol{y}, \delta)$ for an appropriate choice of the dictionary $\boldsymbol{D}$ and the scalar δ. In the following we describe several such options, as practiced in the past decade. Note that we can replace $\mathcal{P}_0(\boldsymbol{D}, \boldsymbol{y}, \delta)$ and $\mathcal{P}_1(\boldsymbol{D}, \boldsymbol{y}, \delta)$ with the equivalent (Lagrangean) form

$$\mathcal{G}_p(\boldsymbol{D}, \boldsymbol{y}, \lambda): \quad \min_{\boldsymbol{\alpha}} \quad \lambda \|\boldsymbol{\alpha}\|_p^p + \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2, \tag{4}$$

that replaces the constraint by a penalty. We shall use these two modes interchangeably, and often times we shall prefer to work with $p > 0$ for obtaining better denoising performance.

1) Orthogonal Wavelet Denoising: Wavelet-based models had a strong impact on the field of image processing, especially in coding and denoising. Their success is due to the tendency of images to become sparse in the wavelet transform domain [41], [43], [49]. This implies that image approximations based on a small subset of wavelets are typically very accurate, which is a key to wavelet-based denoising and compression. Interestingly, it has been found that the human visual system exploits this sparseness property by using wavelet-like representations in the visual cortex [35], [46].

As already mentioned in Section II, the discrete wavelet transform (DWT) of an image $\boldsymbol{x}$ is represented by the multiplication $\boldsymbol{\alpha} = \boldsymbol{T}\boldsymbol{x}$, where $\boldsymbol{T}$ is an orthogonal matrix, i.e., $\boldsymbol{x} = \boldsymbol{T}^T\boldsymbol{\alpha}$. Therefore, the dictionary to be used within $\mathcal{G}_p(\boldsymbol{D}, \boldsymbol{y}, \lambda)$ is $\boldsymbol{D} = \boldsymbol{T}^T$. Due to the unitarity of $\boldsymbol{T}$, we have that $\|\boldsymbol{D}\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 = \|\boldsymbol{T}^T\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 = \|\boldsymbol{\alpha} - \boldsymbol{T}\boldsymbol{y}\|_2^2$, and thus our denoising process translates into

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \lambda \|\boldsymbol{\alpha}\|_p^p + \frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2^2, \tag{5}$$

where $\boldsymbol{\beta} = \boldsymbol{T}\boldsymbol{y}$ is the DWT of the noisy image; that is, we have problem $\mathcal{G}_p(\boldsymbol{I}, \boldsymbol{\beta}, \lambda)$. Of course, the final image estimate is obtained as $\hat{\boldsymbol{x}} = \boldsymbol{T}^T\hat{\boldsymbol{\alpha}}$. As both terms, $\|\boldsymbol{\alpha}\|_p^p$ and $\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2^2$, are separable, the optimization decouples into a set of independent scalar problems of the form $\hat{x}_i = \arg\min_x 0.5(x-y)^2 + \lambda|x|^p$, which have particularly simple closed form solutions in the two notable cases $p = 0$ and $p = 1$, called hard- and soft-thresholding, respectively. The corresponding functions hard(y, a) and soft(y, a) are shown in Figure 1; soft thresholding should use $a = \lambda$, whereas hard thresholding should use $\sqrt{2\lambda}$.

The solution of (5) is thus obtained by applying this soft/hard thresholding function in a component-wise fashion to the observed coefficients $\boldsymbol{T}\boldsymbol{y}$ and this way obtain $\hat{\boldsymbol{\alpha}}$. Then, an inverse DWT leads to the denoised image, $\hat{\boldsymbol{x}} = \boldsymbol{T}^T\hat{\boldsymbol{\alpha}}$. This signal/image denoising method just described is the well-known transform–shrink–inverse-transform approach, which sparked the explosion of interest in wavelet-based signal/image

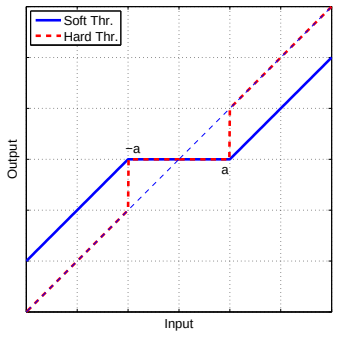

Fig. 1. The soft and hard shrinkage curves.

processing that took place in the early 90's [17], [18], [42], [43], [53].

It should be noted that another important contribution to this explosion of interest in wavelet was the development of fast wavelet transforms [40]; instead of the quadratic cost $O(n^2)$ of a direct implementation of matrix-vector products by the orthogonal matrix $\boldsymbol{T}$ and its inverse $\boldsymbol{T}^T$, it is possible to compute these products with linear cost $O(n)$, by exploiting the multi-scale structure of the wavelet basis. Notice that this is even faster than the $O(n \log n)$ cost of the fast Fourier transform (FFT).

2) Denoising With Wavelet Frames: It is well known that the dyadic image partitioning underlying the orthogonal DWT causes blocky artifacts in the processed images. In denoising problems, translation-invariant approaches have been shown to significantly reduce these artifacts and produce a substantial improvement in the denoising outcome [13], [27], [36].

Let $\boldsymbol{x}$ be represented with a redundant dictionary $\boldsymbol{D}$ that has more columns than rows, as advocated in Section III, i.e. $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}$. Assume that $\boldsymbol{D}$ is associated with some transform matrix $\boldsymbol{T} = \boldsymbol{D}^+$. This allows computing $\boldsymbol{\alpha}$ from $\boldsymbol{x}$ as $\boldsymbol{\alpha} = \boldsymbol{T}\boldsymbol{x}$, via the well-known perfect reconstruction property $\boldsymbol{D}\boldsymbol{T} = \mathbf{I}$ [52]. When $\boldsymbol{D}$ is a tight frame, we simply have $\boldsymbol{T} = \boldsymbol{D}^T$, thus $\boldsymbol{D}\boldsymbol{D}^T = \mathbf{I}$; this is the case, for example, for any orthogonal basis. In the case of an overcomplete tight frame, $\boldsymbol{D}\boldsymbol{D}^T = \mathbf{I}$ holds, but, unlike in an orthogonal basis, $\boldsymbol{D}^T\boldsymbol{D} \neq \mathbf{I}$.

A standard example of a tight wavelet frame, known as a shift-invariant representation, contains all possible shifts of each element of an orthogonal wavelet basis. The size of $\boldsymbol{D}$ in such a case is $n \times (n \log n)$, in the case of a full wavelet decomposition (all the available scales). In recent years, other wavelet-type frames have been introduced, such as the *curvelets*, to address the problem of finding optimally sparse representations for images with discontinuities along piecewise smooth edges [8].

Returning to our denoising goal, we adopt again the formulation of $\mathcal{G}_p(\boldsymbol{D}, \boldsymbol{y}, \lambda)$ in Equation (4) as a way to derive the denoising algorithm. Thus, we aim to solve

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \lambda \|\boldsymbol{\alpha}\|_p^p + \frac{1}{2}\|\boldsymbol{D}\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2, \tag{6}$$

with $\boldsymbol{D}$ as described above.

Unlike (5), this problem does not have a simple closed form solutions, not even in the two notable cases $p = 0$ and $p = 1$ mentioned above, because the presence of $\boldsymbol{D}$ destroys the separability which was present in (5). The denoising problem posed in (6) can be solved (at least approximately) by greedy techniques, if the dimensions involved are relatively low ($n \leq 1000$) [6]. Considering a high dimensional case (e.g., $n = 10^6$, representing an image of size 1000×1000), such greedy methods are no longer relevant. In recent years, *iterative shrinkage/thresholding* (IST) algorithms, tailored for objective functions of the form (6), were independently proposed by several authors [45], [28], [16], [29], [3], [22], [24], [50], [51]. Recently, the work reported in [14] brought important contributions (including strong convergence results) to the study of a class of *forward-backward splitting* algorithms, which includes IST as a particular member. The IST algorithm for (6) has the form

$$\hat{\boldsymbol{\alpha}}_{t+1} = \arg\min_{\boldsymbol{\alpha}} \lambda \|\boldsymbol{\alpha}\|_p^p + \frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{\beta}_t\|_2^2, \tag{7}$$

where $\boldsymbol{\beta}_t = \hat{\boldsymbol{\alpha}}_t - \nu\,\boldsymbol{D}^T(\boldsymbol{D}\hat{\boldsymbol{\alpha}}_t - \boldsymbol{y})$. Each step of (7) is a pure denoising problem of the form (5), which has a simple closed form solution. This way, the overall denoising process requires a sequence of simple steps of multiplication by $\boldsymbol{D}$ and its adjoint, and a scalar shrinkage step. Assuming that the multiplication by the dictionary (and its adjoint) has a fast $O(n \log n)$ algorithm, the overall process is very fast and effective. We should also mention that fast IST-like algorithms were recently proposed by several authors [4], [21], [55].

3) Denoising With Learned Dictionaries: To complete our discussion on image denoising, we now turn to a more recent development that employs a learned dictionary $\boldsymbol{D}$. Rather than working on the image as a whole, we now adopt the sparse and redundant representation model on small image patches of size $\sqrt{n} \times \sqrt{n}$ (a typical choice is in the range $5 \leq \sqrt{n} \leq 20$) [23]. We assume that each and every patch in the given image (with overlaps) is expected to have a sparse representation with respect to $\boldsymbol{D}$. Thus, the dictionary is a small matrix of size $n \times m$, where for example, $n \approx 100$ and $m \approx 200$.

Embarking from the formulation of $\mathcal{P}_0(\boldsymbol{D}, \boldsymbol{y}, \delta)$ in Equation (2), we define the following optimality criterion for denoising the complete image based on the patch model,

$$\begin{aligned} \min_{\boldsymbol{x}, \{\boldsymbol{\alpha}_i\}_{i\in\Omega}} \quad & \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \sum_{i\in\Omega} \|\boldsymbol{\alpha}_i\|_0^0 \\ \text{subject to} \quad & \|\boldsymbol{R}_i\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\alpha}_i\|_2 \leq \delta, \ \forall i \in \Omega. \end{aligned} \tag{8}$$

In this formulation, the domain of the image is defined as Ω, and we index locations in it by $i \in \Omega$. The operator $\boldsymbol{R}_i$ extracts a patch of size $\sqrt{n} \times \sqrt{n}$ from location i. For each patch we construct a representation $\boldsymbol{\alpha}_i$ that should be both sparse and represent $\boldsymbol{R}_i\boldsymbol{x}$ to within a pre-specified error.

Our goal is to find both the set of representations, and a clean image estimate $\hat{\boldsymbol{x}}$. A block-coordinate relaxation technique can be used, where we fix $\boldsymbol{x} = \boldsymbol{y}$ and find the representations $\{\boldsymbol{\alpha}_i\}_{i\in\Omega}$ first. These are found by solving a set of problems of the form $\mathcal{P}_0(\boldsymbol{D}, \boldsymbol{R}_i\boldsymbol{y}, \delta)$. Since these problems are of low-dimension, a greedy approach, such as the OMP, can be used effectively. Once found, $\hat{\boldsymbol{x}}$ can be computed by

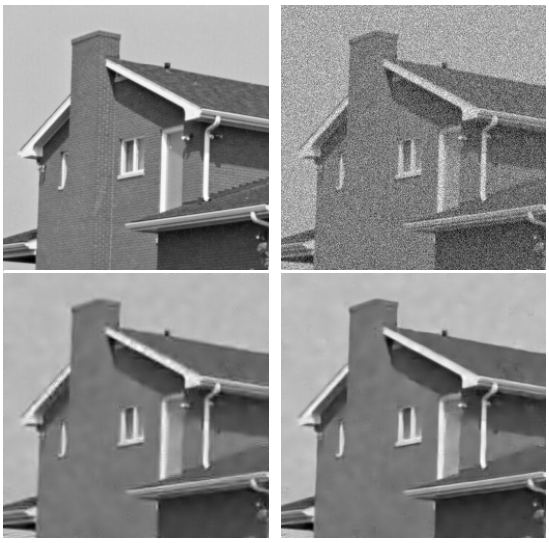

Fig. 2. Denoising results using the local-patch processing algorithm [23]. Top left: the original image; Top right: The noisy image with $\sigma = 25$ (PSNR=20.19dB); Bottom left: Denoised using a redundant DCT dictionary (PSNR=31.00dB); and Bottom right: Denoised using an adaptive dictionary (PSNR=32.25dB).

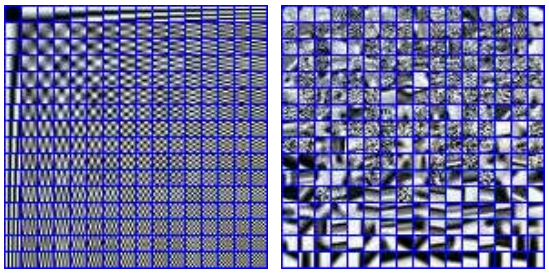

Fig. 3. The atoms of the (redundant) DCT and the adaptive dictionaries. In both cases the atoms are of size 8×8 pixels, and there are $m = 256$ of them.

fixing these representations and solving

$$\min_{\boldsymbol{x}} \ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \lambda \sum_{i \in \Omega} \ \|\boldsymbol{R}_i\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\alpha}_i\|_2^2. \tag{9}$$

Notice that the constraints have been turned into a penalty for this expression to be easy to handle. Since this is a purely quadratic expression, its minimization is easily obtained as a linear set of equations. As it so happens, this set of equations leads to a simple averaging of the cleaned patches, with a small portion the noisy image [23].

As for the identity of $\boldsymbol{D}$, one could use a pre-specified dictionary, such as a redundant DCT, or gather many patch examples and learn a dictionary that sparsifies them, with the hope that this property generalizes well to the image to be cleaned [1]. Alternatively, somewhat better denoising can be obtained if the learning procedure is employed on patches extracted from the noisy image $\boldsymbol{y}$ itself, adapting to the specific image content to be handled [23]. Figure 2 presents the denoising results for the DCT and the adaptive dictionaries, and 3 shows the dictionaries themeselves.

B. Image Deblurring

Image deblurring is one of the earliest and most classical linear inverse problems in imaging, dating back to the 1960's [2]. In image deblurring a noisy blurred version of $\boldsymbol{x}$ is observed, i.e., $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}+\boldsymbol{v}$, where $\boldsymbol{H}$ represents a blur operation, which in practice may result from physical mechanisms such as relative motion between the camera and the subject (motion blur), lack of focus (defocusing blur), or a number of other mechanisms which are well modeled by a convolution.

There are (too) many image deblurring algorithms, and some of these exploit sparse representation modeling in various ways. We present one such approach here, to illustrate the power of this model, and the simplicity with which one can get state-of-the-art results for this application by a simple adoption of the paradigm presented.

As in the denoising case discussed above, image deblurring can be approached using the form $\mathcal{G}_p(\boldsymbol{H}\boldsymbol{D}, \boldsymbol{y}, \lambda)$ as defined in Equation (4), where we have adopted the idea presented in Section III-B of adding the blur operation into the effective dictionary. Thus the problem that needs to be solved is

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \lambda\|\boldsymbol{\alpha}\|_p^p + \frac{1}{2}\|\boldsymbol{H}\boldsymbol{D}\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2. \tag{10}$$

Assuming that we operate on the whole image using a tight frame of some sort (e.g., redundant Haar as in [27]), as described in Section IV-A2, we can apply the IST algorithm to this problem, leading to update equations of the same form as (7), but with $\boldsymbol{H}\boldsymbol{D}$ replacing $\boldsymbol{D}$. Figure 4 exemplifies wavelet-based deblurring, where $\boldsymbol{H}$ represents a 9×9 uniform blur and and $\boldsymbol{D}$ is a shift-invariant Haar frame.

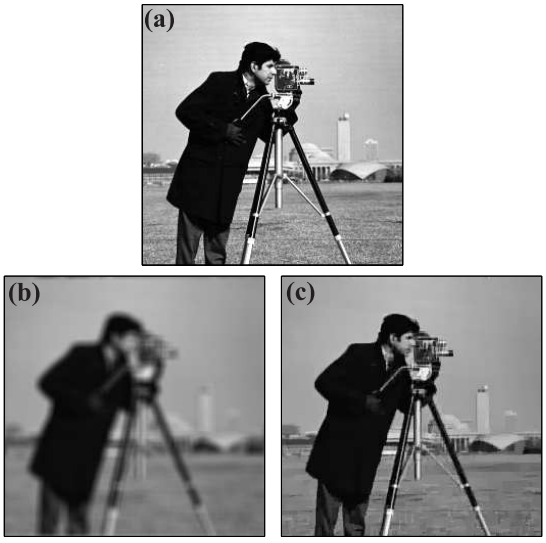

Fig. 4. (a) Original image, (b) blurred image (uniform 9×9), and (c) deblurred image, using the IST algorithm.

C. Image Inpainting

Image inpainting refers to the desire to fill-in missing values in an image, based on their surrounding neighbors. Such a problem is often encountered in a faulty transmission of image content over unreliable channels, when missing blocks are to

Fig. 5. Two input images with missing pixels (in black) are given on the left (top and bottom). The corresponding MCA inpainting results from [26] are shown on the right.

be recovered. Another instance of the inpainting problem is obtained in cases of scratches in images (e.g. old film) that are to be fixed, or if one desires to manipulate an image content to remove portions from it. Ever since its introduction in the image processing literature, this problem attracted a considerable attention and many solutions have been proposed. One popular family of inpainting algorithms is based on partial differential equations that propagate the information from the borders of the holes in the image.

More recently, sparse representation modeling found its way into this field, with promising performance. At its core, the inpainting problem can be formulated as the problem $\mathcal{G}_p(\boldsymbol{M}\boldsymbol{D}, \boldsymbol{y}, \lambda)$, where $\boldsymbol{m}$ is a diagonal mask matrix of size $n \times n$, with 1-s for existing pixels, and 0-es elsewhere. We further assume that in the missing pixels the image $\boldsymbol{y}$ is set to zero. Having posed the problem this way, all the above discussion on denosing and deblurring becomes relevant, and in fact leads to an effective solution of the problem.

There are several algorithms proposed for the inpainting problem along the above lines, some of which adopt the global approach of handling the image as a whole [26], and others that operate on patches [32], [33], [39]. Due to their close resemblance to deblurring algorithms, we shall not dwell further on this topic, and simply show typical examples. The results shown in Figure 5 are taken from [26]. This inpainting algorithm follow the global approach, and combines MCA (for separation of cartoon and texture content) with inpainting. The MCA dictionary used is composed of curvelets and local-DCT, and as can be seen, the quality of the filling-in is near-perfect.

D. Super Resolution

With high-definition TV and electronic displays becoming ever more popular, a very pressing task is how to convert all the old images and videos to a higher resolution, such that their sharpness and naturalness match those of true high-resolution images. Image super-resolution is arguably one of the most classical inverse problems in image processing and is, by nature, intrinsically under-determined. The problem can be simply stated as that of recovering a high-resolution image $\boldsymbol{x} \in \mathbb{R}^n$ from its low-resolution version $\boldsymbol{y} \in \mathbb{R}^k$ (with $q < n$). We model the relation between these two by

$$\boldsymbol{y} = \boldsymbol{S}\boldsymbol{H}\boldsymbol{x} = \boldsymbol{L}\boldsymbol{x}, \tag{11}$$

where $\boldsymbol{H}$ is a linear filter that models certain low-pass filtering (blurring, e.g., with a Gaussian kernel), $\boldsymbol{S}$ is a down-sampling operator, and $\boldsymbol{L} = \boldsymbol{S}\boldsymbol{H}$. The dimension of $\boldsymbol{y}$ is significantly smaller than that of $\boldsymbol{x}$, thus there are infinitely many possible vectors $\boldsymbol{x}$ that satisfy the above equation.

Obviously, to obtain a unique and "good" high-resolution image, proper regularization is needed by imposing certain priors on the solution. As discussed in Section III-C, one plausible model for natural images is based on local PCA. The model suggests that a small image patch is likely to have a very sparse representation with respect to a dictionary of densely sampled patches from natural images. Empirical evidence has suggested that a large set (of the order of a hundred thousand) of patches randomly sampled from natural images is in fact an excellent candidate for such a dictionary [57].

With respect to such a dictionary of high-resolution patches, denoted as $\mathbf{D}_h$, we may assume that any high-resolution image patch has a sparse representation,

$$\boldsymbol{x} = \mathbf{D}_h\, \boldsymbol{\alpha}, \tag{12}$$

for some $\boldsymbol{\alpha}$ with $\|\boldsymbol{\alpha}\|_0 \leq k$. Thus, the super-resolution problem becomes that of recovering $\boldsymbol{\alpha}$ from low-dimensional measurements

$$\boldsymbol{y} = \boldsymbol{L}\,\boldsymbol{x} = \boldsymbol{L}\,\boldsymbol{D}_h\,\boldsymbol{\alpha}. \tag{13}$$

Notice that $\boldsymbol{D}_l = \boldsymbol{L}\,\boldsymbol{D}_h$ can be seen as a dictionary of corresponding low-resolution image patches. We may thus attempt to recover α by solving $\mathcal{P}_p(\boldsymbol{D}_l, \boldsymbol{y}, \delta)$ or $\mathcal{G}_p(\boldsymbol{D}_l, \boldsymbol{y}, \lambda)$. This seemingly naive choice of dictionary has turned out to be extremely effective for generating high-quality high-resolution images. Such a sparsity-based approach for super resolution has been thoroughly explored in [57], [56]. Figure 6 shows a typical result in comparison with other state-of-the-art super-resolution methods in the literature. It is striking to see that such a simple scheme achieves qualitatively and quantitatively better results than other much more sophisticated learning or estimation schemes.

V. Summary and Conclusions

In this paper, we have briefly reviewed sparse and redundant representations as a new model that harnesses the local low-dimensional structure of natural images. Overwhelming theoretical and empirical evidence suggest that this model provides a unified mathematical and computational framework for image processing. Within this framework, one can develop more efficient and effective solutions to many conventional image processing tasks, including but not limited to image compression, denoising, deblurring, inpainting, super resolution, segmentation, and more. Despite its success so far, many difficult and open problems remain regarding why these algorithms work so well and under what conditions. We hope this paper could inspire researchers to investigate these problems

Fig. 6. The girl image magnified by a factor of 4. Top left: low-resolution input, To right: ℓ^2-based back projection, Bottom left: learning-based method in [30], and Bottom right: the sparse representation method.

further and gain better insights about image modeling and processing.

ACKNOWLEDGMENT

ME thanks M. Aharon, D. Donoho, and J.-L. Starck, with whom he worked on image denoising and inpainting. His work is partly supported by the Israel Science Foundation grant no. 599/08.

MF thanks J. Bioucas-Dias, R. Nowak, and S. Wright, with whom he worked on image denoising and deblurring. His work was partially supported by *Fundação para a Ciência e Tecnologia (FCT)*, Portuguese Ministry of Science, Technology, and Higher Education.

YM thanks his former students and colleagues on the work of generalized PCA and image super resolution, H. Derksen, R. Fossum, W. Hong, K. Huang, T. Huang, S. Sastry, R. Vidal, J. Wright, and J. Yang. His work is partially supported by research grants from NSF and ONR of USA.

REFERENCES

[1] M. Aharon, M. Elad, and A.M. Bruckstein, "The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation", *IEEE Transactions On Signal Processing*, vol. 54, pp. 4311–4322, November 2006.

[2] H. Andrews and B. Hunt. *Digital Image Restoration*, Prentice Hall, Englewood Cliffs, NJ, 1977.

[3] J. Bect, L. Blanc-Féraud, G. Aubert, and A. Chambolle, "A ℓ^1-unified variational framework for image restoration", *European Conference on Computer Vision – ECCV'2004*, pp. 1–13, Springer-Verlag, 2004.

[4] J. Bioucas-Dias, M. Figueiredo, "A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration", *IEEE Transactions on Image Processing,* vol. 16, pp. 2992–3004, 2007.

[5] A. Blake and A. Zisserman, *Visual Reconstruction,* MIT Press, Cambridge, MA, 1987.

[6] A. Bruckstein, D. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modelling of signals and images", *SIAM Review*, vol. 51, pp. 34–81, 2009.

[7] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition – CVPR'05*, pp. 60-65, San Diego, 2005.

[8] E.J. Candès and D. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities", *Communications on Pure and Applied Mathematics,* vol. 57, pp. 219-266, 2004.

[9] E.J. Candès and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n", *Annals of Statistics*, vol. 35, pp. 23132351, 2005.

[10] E.J. Candès and M.B. Wakin, "An introduction to compressive sampling", *IEEE Signal Processing Magazine*, vol. 24, pp. 21–30, March 2008.

[11] T. Chan, S. Esedoglu, F. Park, and A. Yip, "Recent developments in total variation image restoration", in *Handbook of Mathematical Models in Computer Vision*, N. Paragios, Y. Chen, O. Faugeras (Editors), Springer Verlag, 2005.

[12] S. Chen, D. Donoho, and M. Saunders, "Atomic decompositions by basis pursuit", *SIAM Review*, vol. 43, pp. 129-159, 2001.

[13] R. Coifman and D. Donoho. "Translation invariant de-noising," in A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, LNS vol. 103, pp. 125–150, Springer-Verlag, New York, 1995.

[14] P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *SIAM Journal on Multiscale Modeling & Simulation*, vol. 4, pp. 1168–1200, 2005.

[15] S. Dai, M. Han, W. Xu, Y. Wu, and Y. Gong. "Soft edge smoothness prior for alpha channel super resolution," *Proceedings of International Conference on Computer Vision*, 2007.

[16] I. Daubechies, M. Defriese, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint", *Communications on Pure and Applied Mathematics*, vol. LVII, pp. 1413-1457, 2004.

[17] D. Donoho. "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, pp. 613–627, 1995.

[18] D. Donoho and I. Johnstone. "Ideal adaptation via wavelet shrinkage." *Biometrika*, vol. 81, pp. 425–455, 1994.

[19] D. Donoho and J. Tanner, "Neighborliness of randomly-projected simplices in high dimensions", *Proceedings of the National Academy of Sciences,* vol. 102, pp. 9452–9457, 2005.

[20] A. Efros and T. Leung, "Texture synthesis by non-parametric sampling", *Proceedings of the International Conference on Computer Vision – ICCV'99*, pp. 1033-1038, Corfu, Greece, 1999.

[21] M. Elad, B. Matalon, and M. Zibulevsky, "Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization", *Applied and Computational Harmonic Analysis,* vol. 23, pp. 346–367, 2007.

[22] M. Elad, "Why simple shrinkage is still relevant for redundant representations?", *IEEE Transactions on Information Theory*, vol. 52, pp. 5559-5569, 2006.

[23] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries", *IEEE Transactions on Image Processing*, vol. 15, pp. 3736–3745, December 2006.

[24] M. Elad, B. Matalon, and M. Zibulevsky, "Image denoising with shrinkage and redundant representations", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition – CVPR'2006*, New York, 2006.

[25] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors", *Inverse Problems*, vol. 23, pp. 947–968, 2007.

[26] M. Elad, J-L. Starck, P. Querre, and D.L. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA), *Journal on Applied and Computational Harmonic Analysis*, vol. 19, pp. 340–358, November 2005.

[27] M. Figueiredo and R. Nowak. "Wavelet-based image estimation: an empirical Bayes approach using Jeffreys' noninformative prior," *IEEE Trans. on Image Proc.*, vol. 10, pp. 1322-1331, 2001.

[28] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. on Image Processing*, vol. 12, no. 8, pp. 906–916, 2003.

[29] M. Figueiredo and R. Nowak, "A bound optimization approach to wavelet-based image deconvolution," *IEEE Intern. Conf. on Image Processing – ICIP'05*, Genoa, Italy, 2005.

[30] W. T. Freeman, T. R. Jones, and E. C. Pasztor. "Example-based super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, issue 2, 2002.

[31] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 6, pp. 721–741, 1984.

[32] O.G. Guleryuz, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising - Part I: Theory", *IEEE Transactions On Image Processing*, vol. 15, pp. 539–554, March 2006.

[33] O.G. Guleryuz, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising - Part II: Adaptive algorithms", *IEEE Transactions On Image Processing*, vol. 15, pp. 555–571, March 2006.

[34] W. Hong, J. Wright, K. Huang, and Y. Ma, "Multi-scale hybrid linear models for lossy image representation", *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3655-3671, December 2006.

[35] A. Hyvärinen, "Sparse code shrinkage: Denoising of non-Gaussian data by maximum likelihood estimation", *Neural Computation*, vol. 11, pp. 1739–1768, 1999.

[36] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus and R. O. Wells. "Noise reduction using an undecimated discrete wavelet transform," *IEEE Signal Proc. Letters*, vol. 3, pp. 10–12, 1996.

[37] Y. Ma, A. Yang, H. Derksen, and R. Fossum. "Estimation of subspace arrangements with applications in modeling and segmenting mixed data", *SIAM Review*, vol. 50, no. 3, August 2008.

[38] Y. Ma, H. Derksen, W. Hong, and J. Wright. "Segmentation of multivariate mixed data via lossy coding and compression", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1546-1562, September 2007.

[39] J. Mairal, M. Elad, and G. Sapiro,"Sparse representation for color image restoration", *IEEE Transactions On Image Processing*, vol. 17, pp. 53–69, January 2008.

[40] S. Mallat, "A theory of multiresolution signal decomposition: the wavelet representation", *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 11, pp. 674–693, 1989.

[41] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, CA, 1998.

[42] M. Mihçak, I. Kozintsev, K. Ramchandran, and P. Moulin. "Low-complexity image denoising based on statistical modeling of wavelet coefficients." *IEEE Signal Proc. Letters*, vol. 6, pp. 300–303, 1999.

[43] P. Moulin and J. Liu. "Analysis of multiresolution image denoising schemes using generalized-Gaussian and complexity priors," *IEEE Transactions on Information Theory,* vol. 45, pp. 909–919, 1999.

[44] D. Mumford, "Empirical statistics and stochastic models for visual signals", in *New Directions in Statistical Signal Processing: From Systems to Brain,* S. Haykin, J. Principe, T. Sejnowski, and J. McWhirter, Editors, MIT Press, 2005.

[45] R. Nowak and M. Figueiredo, "Fast wavelet-based image deconvolution using the EM algorithm", *Proc. 35th Asilomar Conf. on Signals, Systems, and Computers*, vol. 1, pp. 371–375, 2001.

[46] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images", *Nature*, vol. 381, pp. 607–609, 1996.

[47] G. Peyré, "Manifold models for signals and images", *Computer Vision and Image Understanding*, vol. 113, pp. 249–260, 2009.

[48] L. Rudin, S. Osher, and E. Fatemi, Nonlinear total variation based noise removal algorithms, *Physica D*, vol. 60, pp. 259–268, 1992.

[49] E. Simoncelli and E. Adelson, "Noise removal via Bayesian wavelet coring", *Proceedings of the IEEE International Conference on Image Processing,* pp. 379–382, Lausanne, Switzerland, 1996.

[50] J.-L. Starck, E. Candès, and D. Donoho. "Astronomical image representation by the curvelet transform", *Astronomy and Astrophysics*, vol. 398, pp. 785–800, 2003.

[51] J.-L. Starck, M. Nguyen, and F. Murtagh. "Wavelets and curvelets for image deconvolution: a combined approach", *Signal Processing*, vol. 83, pp. 2279–2283, 2003.

[52] M. Vetterli, "A theory of multirate filter banks", *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, vol. 35, pp. 356–372, 1987.

[53] B. Vidakovic. "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors", *Jour. Amer. Statist. Assoc.*, vol. 93, pp. 173–179, 1998.

[54] R. Vidal, Y. Ma, and S. Sasry. "Generalized principal component analysis", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1945-1959, December 2005.

[55] S. Wright, R. Nowak, M. Figueiredo, "Sparse reconstruction by separable approximation, *IEEE Transactions on Signal Processing,* 2009 (to appear).

[56] J. Yang, H. Tang, Y. Ma, and T. Huang. "Face hallucination via sparse coding," *Proceedings of International Conference on Image Processing,* 2008.

[57] J. Yang, J. Wright, T. Huang, and Y. Ma. "Image super-resolution as sparse representation of raw image patches," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* June 2008.