Dictionaries for Sparse Representation Modeling

Ron Rubinstein, *Student Member, IEEE*, Alfred M. Bruckstein, *Member, IEEE*,
and Michael Elad, *Senior Member, IEEE*

***Abstract*—Sparse and redundant representation modeling of data assumes an ability to describe signals as linear combinations of a few atoms from a pre-specified dictionary. As such, the choice of the dictionary that sparsifies the signals is crucial for the success of this model. In general, the choice of a proper dictionary can be done using one of two ways: (i) building a sparsifying dictionary based on a mathematical model of the data, or (ii) learning a dictionary to perform best on a training set. In this paper we describe the evolution of these two paradigms. As manifestations of the first approach, we cover topics such as wavelets, wavelet packets, contourlets, and curvelets, all aiming to exploit 1-D and 2-D mathematical models for constructing effective dictionaries for signals and images. Dictionary learning takes a different route, attaching the dictionary to a set of examples it is supposed to serve. From the seminal work of Field and Olshausen, through the MOD, the K-SVD, the Generalized PCA and others, this paper surveys the various options such training has to offer, up to the most recent contributions and structures.**

***Index Terms*—Dictionary learning, harmonic analysis, signal representation, signal approximation, sparse coding, sparse representation.**

I. Introduction

THE process of digitally sampling a natural signal leads to its representation as the sum of Delta functions in space or time. This representation, while convenient for the purposes of display or playback, is mostly inefficient for analysis tasks. Signal processing techniques commonly require more meaningful representations which capture the useful characteristics of the signal — for recognition, the representation should highlight salient features; for denoising, the representation should efficiently separate signal and noise; and for compression, the representation should capture a large part of the signal with only a few coefficients. Interestingly, in many cases these seemingly different goals align, sharing a core desire for *simplification*.

Representing a signal involves the choice of a *dictionary*, which is the set of elementary signals – or *atoms* – used to decompose the signal. When the dictionary forms a basis, every signal is uniquely represented as the linear combination of the dictionary atoms. In the simplest case the dictionary is orthogonal, and the representation coefficients can be computed as inner products of the signal and the atoms; in the non-orthogonal case, the coefficients are the inner products of the signal and the dictionary inverse, also referred to as the bi-orthogonal dictionary.

R. Rubinstein, A.M. Bruckstein and M. Elad are with the Department of Computer Science, The Technion – Israel Institute of Technology, Haifa 32000, Israel, http://www.cs.technion.ac.il

This research was partly supported by the European Communitys FP7-FET program, SMALL project, under grant agreement no. 225913, and by the ISF grant number 599/08

For years, orthogonal and bi-orthogonal dictionaries were dominant due to their mathematical simplicity. However, the weakness of these dictionaries — namely their limited expressiveness — eventually outweighed their simplicity. This led to the development of newer *overcomplete* dictionaries, having more atoms than the dimension of the signal, and which promised to represent a wider range of signal phenomena.

The move to overcomplete dictionaries was done cautiously, in an attempt to minimize the loss of favorable properties offered by orthogonal transforms. Many proposed dictionaries formed *tight frames*, which ensured that the representation of the signal as a linear combination of the atoms could still be identified with the inner products of the signal and the dictionary. Another approach, manifested by the *Best Basis* algorithm, utilized a specific dictionary structure which essentially allowed it to serve as a pool of atoms from which an *orthogonal* sub-dictionary could be efficiently selected.

Research on *general* overcomplete dictionaries mostly commenced over the past decade, and is still intensely ongoing. Such dictionaries introduce an intriguing ambiguity in the definition of a signal representation. We consider the dictionary $\mathbf{D} = [\mathbf{d}_1\, \mathbf{d}_2 \ldots \mathbf{d}_L] \in \mathbb{R}^{N \times L}$, where the columns constitute the dictionary atoms, and $L \geq N$. Representing a signal $\mathbf{x} \in \mathbb{R}^N$ using this dictionary can take one of two paths — either the *analysis* path, where the signal is represented via its inner products with the atoms,

$$\boldsymbol{\gamma}_a = \mathbf{D}^T \mathbf{x} \;, \tag{1}$$

or the *synthesis* path, where it is represented as a linear combination of the dictionary atoms,

$$\mathbf{x} = \mathbf{D}\boldsymbol{\gamma}_s \;. \tag{2}$$

The two definitions coincide in the complete case ($L = N$), when the analysis and synthesis dictionaries are bi-orthogonal. In the general case, however, the two may dramatically differ.

The synthesis approach poses yet another interesting question: when $\mathbf{D}$ is overcomplete, the family of representations $\boldsymbol{\gamma}_s$ satisfying (2) is actually *infinitely large*, with the degrees of freedom identified with the null-space of $\mathbf{D}$. This allows us to seek the most informative representation of the signal with respect to some cost function $C(\boldsymbol{\gamma})$:

$$\boldsymbol{\gamma}_s = \operatorname*{Argmin}_{\boldsymbol{\gamma}} C(\boldsymbol{\gamma}) \quad \text{Subject To} \quad \mathbf{x} = \mathbf{D}\boldsymbol{\gamma} \;. \tag{3}$$

Practical choices of $C(\boldsymbol{\gamma})$ promote the *sparsity* of the representation, meaning that we want the sorted coefficients to decay quickly. Solving (3) is thus commonly referred to as *sparse coding*. We can achieve sparsity by choosing $C(\boldsymbol{\gamma})$ as some robust penalty function, which we loosely define as a function that is tolerant to large coefficients but aggressively

penalizes small non-zero coefficients. Examples include the Huber function [1] as well as the various ℓ^p cost functions with $0 \leq p \leq 1$.

The two options (1) and (2), and specifically the problem (3), have been extensively studied over the past few years. This in turn has led to the development of new signal processing algorithms which utilize general overcomplete transforms. However, in going from theory to practice, the challenge of *choosing* the proper dictionary for a given task must be addressed. Earlier works made use of traditional dictionaries, such as the Fourier and wavelet dictionaries, which are simple to use and perform adequately for 1-dimensional signals. However, these dictionaries are not well equipped for representing more complex natural and high-dimensional signal data, and new and improved dictionary structures were sought.

A variety of dictionaries were developed in response to the rising need. The newly developed dictionaries emerged from one of two sources — either a *mathematical model* of the data, or a *set of realizations* of the data. Dictionaries of the first type are characterized by an analytic formulation and a fast implicit implementation, while dictionaries of the second type deliver increased flexibility and the ability to adapt to specific signal data. Most recently, there is a growing interest in dictionaries which can mediate between the two types, and offer the advantages of both worlds. Such structures are just beginning to emerge, and research is still ongoing.

In this paper we present the fundamental concepts guiding modern dictionary design, and outline the various contributions in the field. In Section 2 we take a historical viewpoint, and trace the evolution of dictionary design methodology from the early 1960's to the late 1990's, focusing on the conceptual advancements. In Sections 3 and 4 we overview the state-of-the art techniques in both analytic and trained dictionaries. We summarize and conclude in Section 5.

II. A History of Transform Design

A. Signal Transforms: The Linear Era

Signal transforms have been around for as long as signal processing has been conducted. In the 1960's, early signal processing researchers gave significant attention to linear time-invariant operators, which were simple and intuitive processes for manipulating analog and digital signals. In this scenery, the Fourier transform naturally emerged as the basis which diagonalizes these operators, and it immediately became a central tool for analyzing and designing such operators. The transform gained tremendous popularity with the introduction of the Fast Fourier Transform (FFT) in 1965 by Cooley and Tukey [2], which provided its numerical appeal.

The Fourier basis describes a signal in terms of its global frequency content, as a combination of orthogonal waveforms

$$\mathcal{F} = \left\{ \phi_n(x) = e^{inx} \right\}_{n \in \mathbb{Z}} \ .$$

A signal is approximated in this basis by projecting it onto the K lowest frequency atoms, which has a strong smoothing and noise-reducing effect. The Fourier basis is thus efficient at describing uniformly *smooth* signals. However, the lack of localization makes it difficult to represent *discontinuities*, which generate large coefficients over all frequencies. Therefore, the Fourier transform typically produces oversmooth results in practical applications. For finite signals, the Fourier transform implicitly assumes a periodic extension of the signal, which introduces a discontinuity at the boundary. The Discrete Cosine Transform (DCT) is the result of assuming an anti-symmetric extension of the signal, which results in continuous boundaries, and hence in a more efficient approximation. Since the DCT has the added advantage of producing non-complex coefficients, it is typically preferred in applications; see Figure 1 for some 2-D DCT atoms.

Signal approximation in the Fourier basis was soon categorized as a specific instance of *linear approximation*: given a basis $\{\boldsymbol{\phi}_n\}_{n=0}^{N-1}$ of $\mathbb{R}^N$, a signal $\mathbf{x} \in \mathbb{R}^N$ is linearly approximated by projecting it onto a *fixed* subset of $K < N$ basis elements

$$\mathbf{x} \approx \sum_{n \in I_K} (\boldsymbol{\psi}_n^T \mathbf{x}) \boldsymbol{\phi}_n \ , \tag{4}$$

where $\{\boldsymbol{\psi}_n\}_{n=0}^{N-1}$ is in general the bi-orthogonal basis ($\boldsymbol{\psi}_n = \boldsymbol{\phi}_n$ in the orthonormal case). This process is an under-complete linear transform of $\mathbf{x}$, and, with the right choice of basis, can achieve *compaction* — the ability to capture a significant part of the signal with only a few coefficients. Indeed, this concept of *compaction* is later replaced with *sparsity*, though the two are closely related [3].

Optimizing compaction was a major driving force for the continued development of more efficient representations. During the 1970's and 1980's, a new and very appealing source of compaction was brought to light: *the data itself*. The focus was on a set of statistical tools developed during the first half of the century, known as the Karhunen-Loève Transform (KLT) [4], [5], or Principal Component Analysis (PCA) [6]. The KLT is a linear transform which can be adapted to represent signals coming from a certain known distribution. The adaptation process fits a low-dimensional subspace to the data which minimizes the ℓ^2 approximation error. Specifically, given the data covariance matrix $\boldsymbol{\Sigma}$ (either known or empirical), the KLT atoms are the first K eigenvectors of the eigenvalue decomposition of $\boldsymbol{\Sigma}$,

$$\boldsymbol{\Sigma} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T \ .$$

From a statistical point of view, this process models the data as coming from a low-dimensional Gaussian distribution, and thus is most effective for Gaussian data. Figure 1 shows an example of the KLT basis trained from a set of image patches. The DCT basis shown in the same figure, is regarded as a good approximation of the KLT for natural image patches when a non-adaptive transform is required.

Compared to the Fourier transform, the KLT is superior (by construction) in terms of representation efficiency. However, this advantage comes at the cost of a non-structured and substantially more complex transform. As we will see, this tradeoff between *efficiency* and *adaptivity* continues to play a major role in modern dictionary design as well.

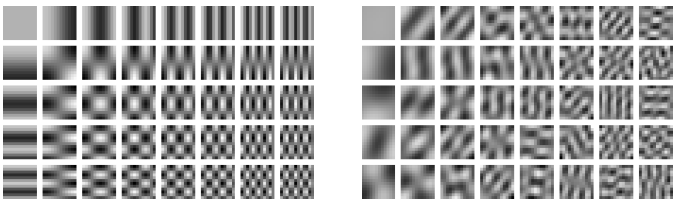

Fig. 1. Left: A few 12×12 DCT atoms. Right: The first 40 KLT atoms, trained using 12×12 image patches from *Lena*.

B. Non-Linear Revolution and Elements of Modern Dictionary Design

In statistics research, the 1980's saw the rise of a new powerful approach known as *robust statistics*. Robust statistics advocates sparsity as a key for a wide range of recovery and analysis tasks. The idea has its roots in classical Physics, and more recently in Information Theory, and promotes simplicity and conciseness in guiding phenomena descriptions. Motivated by these ideas, the 1980's and 1990's were characterized by a search for sparser representations and more efficient transforms.

Increasing sparsity required departure from the linear model, towards a more flexible *non-linear* formulation. In the non-linear case, each signal is allowed to use a different set of atoms from the dictionary in order to achieve the best approximation. Thus, the approximation process becomes

$$\mathbf{x} \approx \sum_{n \in I_K(\mathbf{x})} c_n \boldsymbol{\phi}_n \ , \tag{5}$$

where $I_K(\mathbf{x})$ is an index set adapted to each signal individually (we refer the reader to [5], [7] for more information on this wide topic).

The non-linear view paved the way to the design of newer, more efficient transforms. In the process, many of the fundamental concepts guiding modern dictionary design were formed. Following the historic time line, we trace the emergence of the most important modern dictionary design concepts, which are mostly formed during the last two decades of the 20th century.

Localization: To achieve sparsity, transforms required better localization. Atoms with concentrated supports allow more flexible representations based on the local signal characteristics, and limit the effects of irregularities, which are observed to be the main source of large coefficients. In this spirit, one of the first structures to be used was the Short Time Fourier Transform (STFT) [8], which emerges as a natural extension to the Fourier transform. In the STFT, the Fourier transform is applied locally to (possibly overlapping) portions of the signal, revealing a *time-frequency* (or space-frequency) description of the signal. An example of the STFT is the JPEG image compression algorithm [9], which is based on this concept.

During the 1980's and 1990's, the STFT was extensively researched and generalized, becoming more known as the *Gabor* transform, named in homage of Dennis Gabor, who first suggested the time-frequency decomposition back in 1946 [10]. Gabor's work was independently rediscovered in 1980 by Bastiaans [11] and Janssen [12], who studied the fundamental properties of the expansion.

A basic 1-D Gabor dictionary consists of windowed waveforms

$$\mathcal{G} = \left\{ \phi_{n,m}(x) = w(x - \beta m) e^{i 2\pi \alpha n x} \right\}_{n,m \in \mathbb{Z}} \ ,$$

where $w(\cdot)$ is a low-pass window function localized at 0 (typically a Gaussian), and α and β control the time and frequency resolution of the transform. Much of the mathematical foundations of this transform were laid out during the late 1980's by Daubechies, Grossman and Meyer [13], [14] who studied the transform from the angle of frame theory, and by Feichtinger and Gröchenig [15]–[17] who employed a generalized group-theoretic point of view. Study of the discrete version of the transform and its numerical implementation followed in the early 1990's, with notable contributions by Wexler and Raz [18] and by Qian and Chen [19].

In higher dimensions, more complex Gabor structures were developed which add *directionality*, by varying the orientation of the sinusoidal waves. This structure gained substantial support from the work of Daugman [20], [21], who discovered oriented Gabor-like patterns in simple-cell receptive fields in the visual cortex. These results motivated the deployment of the transform to image processing tasks, led by works such as Daugman [22] and Porat and Zeevi [23]. Today, practical uses of the Gabor transform are mainly in analysis and detection tasks, as a collection of directional filters. Figure 2 shows some examples of 2-D Gabor atoms of various orientations and sizes.

Multi-Resolution: One of the most significant conceptual advancements achieved in the 1980's was the rise of *multi-scale* analysis. It was realized that natural signals, and images specifically, exhibited meaningful structures over many scales, and could be analyzed and described particularly efficiently by multi-scale constructions. One of the simplest and best known such structures is the *Laplacian pyramid*, introduced in 1984 by Burt and Adelson [24]. The Laplacian pyramid represents an image as a series of difference images, where each one corresponds to a different scale and roughly a different frequency band.

In the second half of the 1980's, though, the signal processing community was particularly excited about the development of a new very powerful tool, known as *wavelet analysis* [5], [25], [26]. In a pioneering work from 1984, Grossman and Morlet [27] proposed a signal expansion over a series of translated and dilated versions of a single elementary function, taking the form

$$\mathcal{W} = \left\{ \phi_{n,m}(x) = \alpha^{n/2} f(\alpha^n x - \beta m) \right\}_{n,m \in \mathbb{Z}} \ .$$

This simple idea captivated the signal processing and harmonic analysis communities, and in a series of influential works by Meyer, Daubechies, Mallat and others [13], [14], [28]–[33], an extensive wavelet theory was formalized. The theory was formulated for both the continuous and discrete domains, and a complete mathematical framework relating the two was put forth. A significant breakthrough came from Meyer's work in 1985 [28], who found that unlike the Gabor transform (and

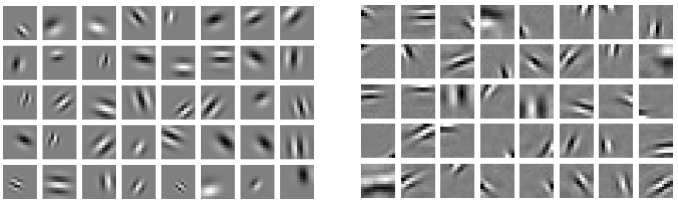

Fig. 2. Left: A few 12×12 Gabor atoms at different scales and orientations. Right: A few atoms trained by Olshausen and Field (extracted from [34]).

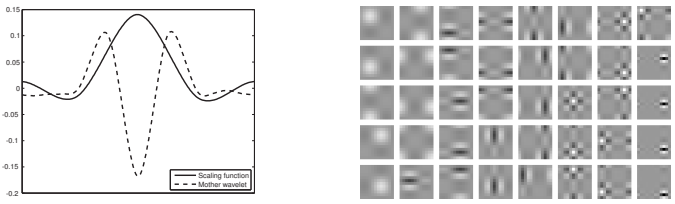

Fig. 3. Left: Coiflet 1-D scaling function (solid) and mother wavelet (dashed). Right: Some 2-D separable Coiflet atoms.

contrary to common belief) the wavelet transform could be designed to be *orthogonal* while maintaining stability — an extremely appealing property to which much of the initial success of the wavelets can be attributed to.

Specifically of interest to the signal processing community was the work of Mallat and his colleagues [31]–[33] which established the wavelet decomposition as a multi-resolution expansion and put forth efficient algorithms for computing it. In Mallat's description, a multi-scale wavelet basis is constructed from a pair of localized functions referred to as the *scaling function* and the *mother wavelet*, see Figure 3. The scaling function is a low frequency signal, and along with its translations, spans the coarse approximation of the signal. The mother wavelet is a high frequency signal, and with its various scales and translations spans the signal detail. In the orthogonal case, the wavelet basis functions at each scale are critically sampled, spanning precisely the new detail introduced by the finer level.

Non-linear approximation in the wavelet basis was shown to be optimal for piecewise-smooth 1-D signals with a finite number of discontinuities, see e.g. [32]. This was a striking finding at the time, realizing that this is achieved without prior detection of the discontinuity locations. Unfortunately, in higher dimensions the wavelet transform loses its optimality; the multi-dimensional transform is a simple separable extension of the 1-D transform, with atoms supported over rectangular regions of different sizes (see Figure 3). This separability makes the transform simple to apply, however the resulting dictionary is only effective for signals with *point* singularities, while most natural signals exhibit elongated *edge* singularities. The JPEG2000 image compression standard, based on the wavelet transform, is indeed known for its *ringing* (smoothing) artifacts near edges.

Adaptivity: Going to the 1990's, the desire to push sparsity even further, and describe increasingly complex phenomena, was gradually revealing the limits of approximation in orthogonal bases. The weakness was mostly associated with the small and fixed number of atoms in the dictionary — dictated by the orthogonality — from which the optimal representation could be constructed. Thus, one option to obtain further sparsity was to adapt *the transform atoms themselves* to the signal content.

One of the first such structures to be proposed was the *wavelet packet* transform, introduced by Coifman, Meyer and Wickerhauser in 1992 [35]. The transform is built upon the success of the wavelet transform, adding adaptivity to allow finer tuning to the specific signal properties. The main observation of Coifman *et al.* was that the wavelet transform enforced a very specific time-frequency structure, with high frequency atoms having small supports and low frequency atoms having large supports. Indeed, this choice has deep connections to the behavior of real natural signals; however, for specific signals, better partitionings may be possible. The wavelet packet dictionary essentially unifies all dyadic time-frequency atoms which can be derived from a specific pair of scaling function and mother wavelet, so atoms of different frequencies can come in an array of time supports. Out of this large collection, the wavelet packet transform allows to efficiently select an optimized *orthogonal* sub-dictionary for any given signal, with the standard wavelet basis being just one of an exponential number of options. The process was thus named by the authors a *Best Basis* search. The wavelet packet transform is, by definition, at least as good as wavelets in terms of coding efficiency. However, we note that the multi-dimensional wavelet packet transform remains a separable and non-oriented transform, and thus does not generally provide a substantial improvement over wavelets for images.

Geometric Invariance and Overcompleteness: In 1992, Simoncelli *et al.* [36] published a thorough work advocating a dictionary property they termed *shiftability*, which describes the invariance of the dictionary under certain geometric deformations, e.g. translation, rotation or scaling. Indeed, the main weakness of the wavelet transform is its strong translation-sensitivity, as well as rotation-sensitivity in higher dimensions. The authors concluded that achieving these properties required abandoning orthogonality in favor of *overcompleteness*, since the critical number of atoms in an orthogonal transform was simply insufficient. In the same work, the authors developed an overcomplete *oriented* wavelet transform — the *steerable wavelet transform* — which was based on their previous work on steerable filters and consisted of localized 2-D wavelet atoms in many orientations, translations and scales.

For the basic 1-D wavelet transform, translation-invariance can be achieved by increasing the sampling density of the atoms. The *stationary wavelet transform*, also known as the undecimated or non-subsampled wavelet transform, is obtained from the orthogonal transform by eliminating the sub-sampling and collecting *all* translations of the atoms over the signal domain. The algorithmic foundation for this was laid by Beylkin in 1992 [37], with the development of an efficient algorithm for computing the undecimated transform. The stationary wavelet transform was indeed found to substantially improve signal recovery compared to orthogonal wavelets, and its benefits were independently demonstrated in 1995 by Nason and Silverman [38] and Coifman and Donoho [39].

C. From Transforms to Dictionaries

By the second half of the 1990's, most of the concepts for designing effective transforms were laid out. At the same time, a conceptual change of a different sort was gradually taking place. In a seminal work from 1993, Mallat and Zhang [40] proposed a novel sparse signal expansion scheme based on the selection of a small subset of functions from a general overcomplete *dictionary* of functions. Shortly after, Chen, Donoho and Saunders published their influential paper on the *Basis Pursuit* [41], and the two works signalled the beginning of a fundamental move from *transforms* to *dictionaries* for sparse signal representation. An array of works since has formed a wide mathematical and algorithmic foundation of this new field, and established it as a central tool in modern signal processing (see [42]).

This seemingly minor terminological change from transforms to dictionaries enclosed the idea that a signal was allowed to have *more than one description* in the representation domain, and that selecting the best one depended on the task. Moreover, it de-coupled the processes of *designing* the dictionary and *coding* the signal: indeed, given the dictionary — the collection of elemental signals — different cost functions could be proposed in (3), and different coding methods could be applied.

The first dictionaries to be used in this way were the existing transforms — such as the Fourier, wavelet, STFT, and Gabor transforms, see e.g. [40], [41]. As an immediate consequence, the move to a dictionary-based formalism provided the benefit of constructing *dictionary mergers*, which are the unions of several simpler dictionaries; these were proposed by Chen, Donoho and Saunders in [41], and provide a simple way to increase the variety of features representable by the dictionary.

D. Higher Dimensional Signals

The variety of dictionaries developed through the mid-1990's served one-dimensional signals relatively well. However, the dictionaries for multi-dimensional signal representation were still unsatisfying. Particularly frustrating, for instance, was the common knowledge that 2-D piecewise-smooth signals could be described much more efficiently using a simple piecewise-linear approximation over an adaptive triangle grid, than using any existing dictionary [5], [43].

In 1998, Donoho developed the *wedgelet* dictionary for 2-D signal representation [44], which bears some resemblance to the adaptive triangulation structure. The wedgelet dictionary consists of constant-valued, axis-aligned squares, bisected by straight lines, and spanning many sizes and locations. Donoho showed that this dictionary is optimal for piecewise-constant images with regular edge discontinuities, and provided a quick (though non-optimal) approximation technique. The elegant wedgelet construction, though too simplistic for many tasks, was adopted and generalized by several researchers, leading to such structures as wavelet-wedgelets hybrids (*wedgeprints*) [45], piecewise-linear wedgelets (*platelets*) [46], and higher-dimensional wedgelets (*surflets*) [47].

In parallel to the wedgelet transform, Candès and Donoho introduced the *ridgelet* transform as a multi-dimensional extension of the wavelet transform [48]. A ridgelet atom is a translated and dilated wavelet in one direction, and fixed in the orthogonal directions (similar to a plane wave). The transform is proven to be optimal for piecewise-smooth functions with plane discontinuities. Indeed, the basic ridgelet dictionary is unsuitable for natural signals due its lack of localization. However, with proper localization and multi-scale extension, the dictionary forms the core of the much more powerful *curvelet* transform [43], [49], introduced by the authors soon after, and which provides a comprehensive framework for representing multi-dimensional signals. Similar efforts led to the development of the *contourlet* and *bandelet* transforms, which are described in more detail in the next section.

E. Analytic versus Trained Dictionaries

The dictionaries described so far all roughly fall under the umbrella of *Harmonic Analysis*, which suggests modeling interesting signal data by a more simple class of *mathematical functions*, and designing an efficient representation around this model. For example, the Fourier dictionary is designed around smooth functions, while the wavelet dictionary is designed around piecewise-smooth functions with point singularities. The dictionaries of this sort are characterized by an analytic formulation, and are usually supported by a set of optimality proofs and error rate bounds. An important advantage of this approach is that the resulting dictionary usually features a fast implicit implementation which does not involve multiplication by the dictionary matrix. On the other hand, the dictionary can only be *as successful as its underlying model*, and indeed, these models are typically over-simplistic compared to the complexity of natural phenomena.

Through the 1980's and 1990's, *Machine Learning* techniques were rapidly gaining interest, and promised to confront this exact difficulty. The basic assumption behind the learning approach is that the structure of complex natural phenomena can be more accurately extracted *directly from the data* than by using a mathematical description. One direct benefit of this is that a finer adaptation to specific instances of the data becomes possible, replacing the use of generic models.

A key contribution to the area of dictionary learning was provided by Olshausen and Field in 1996 [34]. In their widely celebrated paper, the authors trained a dictionary for sparse representation of small image patches collected from a number of natural images. With relatively simple algorithmic machinery, the authors were able to show a remarkable result — the trained atoms they obtained were incredibly similar to the simple-cell receptive fields, which until then were only weakly explained via Gabor filters. The finding was highly motivating to the sparse representation community, as it demonstrated that the single assumption of sparsity could account for a fundamental biological visual behavior. Also, the results demonstrated the potential in example-based methods to uncover elementary structures in complex signal data.

The experiments of Olshausen and Field inspired a series of subsequent works aimed at improving the example-based training process. Towards the end of the 1990's, these works mostly focused on statistical training methods, which model

the examples as random independent variables originating from a sparse noisy source. With $\mathbf{X} = [\mathbf{x}_1\, \mathbf{x}_2 \ldots \mathbf{x}_n]$ denoting the data matrix, the statistical approach suggests seeking for the dictionary which either maximizes the likelihood of the data $P(\mathbf{X}|\mathbf{D})$ (*Maximum Likelihood* estimation), e.g. [50], or maximizes the posterior probability of the dictionary $P(\mathbf{D}|\mathbf{X})$ (*Maximum A-Posterior* estimation), e.g. [51]. The resulting optimization problems in these works are typically solved in an Expectation-Maximization (EM) fashion, alternating estimation of the sparse representations and the dictionary; earlier works employ gradient descent or similar methods for both tasks, while later ones employ more powerful sparse-coding techniques for the estimation of the sparse representations.

III. Analytic Dictionaries — State-of-the-Art

Recent advances in analytic dictionary design have mostly focused on the move to two and higher dimensions. Multi-dimensional signals are significantly more complex than one-dimensional ones due to the addition of *orientation*. Also, the elementary singularities become *curves* — or *manifolds* in general — rather than points, and thus have a much more complex geometry to trace. In order to handle these complex signals, new transforms that are both localized and oriented have been developed.

Analytic dictionaries are typically formulated as *tight frames*, meaning that $\mathbf{D}\mathbf{D}^T\mathbf{x} = \mathbf{x}$ for all $\mathbf{x}$, and therefore the dictionary transpose can be used to obtain a representation over the dictionary. The analytic approach then proceeds by analyzing the behavior of the filter-set $\mathbf{D}^T\mathbf{x}$, and establishes decay rates and error bounds.

The tight frame approach has several advantages. Analyzing the behavior of $\mathbf{D}^T$ as an analysis operator seems easier than deriving sparsity bounds in a synthesis framework, and indeed, results obtained for the analysis formulation also induce upper bounds for the synthesis formulation. Another benefit is that — when formulated carefully — the algorithms for both analysis and synthesis operators become nearly reversals, simplifying algorithm design. Finally, the tight frame approach is beneficial in that it simultaneously produces a useful structure for both the analysis and synthesis frameworks, and has a meaningful interpretation in both.

Sparse-coding in this case is typically done by computing the analysis coefficients $\mathbf{D}^T\mathbf{x}$, and passing them through a non-linear shrinking operator. This method has the advantage of providing a simple and efficient way to achieve sparse representations over the dictionary, though it is worth noting that from a pure synthesis point of view, this process is sub-optimal, and one might benefit from employing a more advanced sparse-coding technique, e.g. an *iterated shrinkage* technique [52], directly to the expansion coefficients.

A. Curvelets

The curvelet transform was introduced by Candès and Donoho in 1999 [43], and was later refined into its present form in 2003 [53]. When published, the transform astonished the harmonic analysis community by achieving what was then believed to be only possible with adaptive representations: it could represent 2-D piecewise-smooth functions with smooth curve discontinuities at an (essentially) optimal rate.

The curvelet transform is formulated as a *continuous* transform, with discretized versions developed for both formulations [49], [53], [54]. Each curvelet atom is associated with a specific location, orientation and scale. In the 2-D case, a curvelet atom is roughly supported over an elongated elliptical region, and is oscillatory along its width and smooth along its length, see Figure 4. The curvelet atoms are characterized by their specific anisotropic support, which obeys a parabolic scaling law $width \sim length^2$. As it turns out, this property is useful for the efficient representation of smooth curves [55], and indeed several subsequent transforms follow this path. In higher dimensions, the curvelet atoms become flattened ellipsoids, oscillatory along their short direction and smooth along the other directions [53], [54], [56].

B. Contourlets

The curvelet transform offers an impressively solid continuous construction and exhibits several useful mathematical properties. However, its discretization turns out to be challenging, and the resulting algorithms are relatively complicated. Also, current discretizations have relatively high redundancies, which makes them more costly to use and less applicable for tasks like compression.

With this in mind, Do and Vetterli proposed the *contourlet* transform in 2002 [57], [58] as an alternative to the 2-D curvelet transform. The transform was later refined in 2006 by Lu and Do [59], and a multi-dimensional version, named *surfacelets*, was also recently introduced [60].

The contourlet transform shares many of the characteristics of the curvelet transform, including localization, orientation, and parabolic scaling. However, as opposed to curvelets, the contourlets are defined *directly in the discrete domain*, and thus have a native and simple construction for discrete signals. Also, the standard contourlet transform has much lower redundancy, approximately in the range $[1.3, 2.3]$ for the second-generation implementation [59], compared to $[2.8, 7.2]$ for second-generation curvelets [53].

The contourlet transform implementation is based on a pyramidal band-pass decomposition of the image followed by a directional filtering stage. The resulting oriented atoms are elongated and oscillatory along their width, with some visual resemblance to the curvelet atoms (see Figure 4). The main appeal of the transform is due to its simple discrete formulation, its low complexity and reduced redundancy. It should be noted, though, that while the transform is well suited for tasks such as compression, its aggressive sub-sampling has been noted to lead to artifacts in signal reconstruction, in which case a translation-invariant version of the transform is preferred [61], [62]; indeed, this option significantly increases redundancy and complexity, though the simpler structure of the transform remains.

C. Bandelets

The bandelet transform was proposed in 2005 by Le Pennec and Mallat [63], with a second version introduced soon after

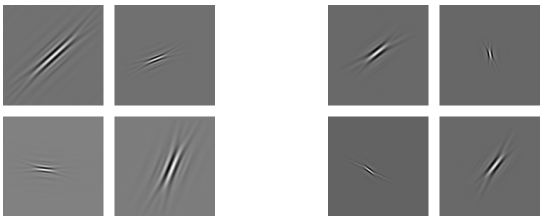

Fig. 4. Some curvelet atoms (left) and contourlet atoms (right). Both represent the second version of the corresponding transform.

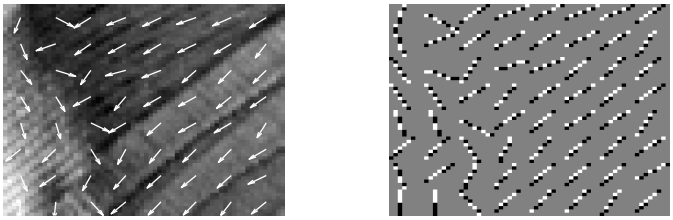

Fig. 5. Left: the flow in a specific image region. Right: some bandelet atoms adapted to the region. Note how the 1-D wavelets are skewed to follow edges.

by Peyré and Mallat [64]. The bandelet transform represents one of the most recent contributions in the area of *signal-adaptive transforms*, and as such it differs fundamentally from the non-adaptive curvelet and contourlet transforms.

The idea behind the bandelet construction is to exploit geometric regularity in the image — specifically edges and directional phenomena — in order to fit a specifically optimized set of atoms for the image. The original bandelet construction operates in the spatial domain, and is based on an adaptive subdivision of the image to dyadic regions according to the local complexity; in each region, a set of skewed wavelets is matched to the image flow, in such a way that the wavelet atoms essentially "wrap-around" the edges rather than cross them. This process significantly reduces the number of large wavelet coefficients, as these typically emerge from the interaction of a wavelet atom and a discontinuity.

The resulting set of atoms forms a (slightly) overcomplete set, which is specifically tailored for representing the given image. In the second bandelet construction, which is formulated in the wavelet domain, the transform is further refined to produce an *orthogonal* set. In terms of dictionaries, the bandelet transform selects a set of atoms from a nearly infinite set, and in fact discretization is the main source for limiting the size of this set. This is as opposed to the wavelet packet transform, for instance, where the complete set of atoms is not much larger than the signal dimension. See Figure 5 for an example of bandelets.

D. Other Analytic Dictionaries

Many additional analytic transforms have been developed during the past decade, some of which we mention briefly. The *complex wavelet transform* [65], [66] is an oriented and near-translation-invariant high-dimensional extension of the wavelet transform, achieved through the utilization of *two* mother wavelets satisfying a specific relationship between them. Similar to the original wavelet transform, the complex wavelet transform is efficient and simple to implement, and the added phase information delivers orientation sensitivity and other favorable properties. The *shearlet* transform [67]–[69] is a recently proposed alternative to curvelets, which utilizes structured shear operations rather than rotations to control orientation. Similar to curvelets, the shearlet transform is based on a comprehensive continuous mathematical construction, and it shares many of the properties of the curvelet transform while providing some attractive new features. See Figure 6 for some examples of complex wavelet and shearlet atoms.

Recent adaptive dictionaries include the *directionlet* transform [70], which is a discrete transform which constructs oriented and anisotropic wavelets based on local image directionality, utilizing a specialized directional grouping of the grid points for its numerical implementation. Finally, the *grouplet* transform [71] is a multi-scale adaptive transform which essentially generalizes Haar wavelets to arbitrary supports, based on image content regularity; when applied in the wavelet domain, the transform bears some resemblance to the second-generation bandelet transform, and thus is referred to as *grouped bandelets*.

IV. Dictionary Training — State-of-the-Art

Dictionary training is a much more recent approach to dictionary design, and as such, has been strongly influenced by the latest advances in sparse representation theory and algorithms. The most recent training methods focus on ℓ^0 and ℓ^1 sparsity measures, which lead to simple formulations and enable the use of recently developed efficient sparse-coding techniques [41], [42], [72]–[75].

The main advantage of trained dictionaries is that they lead to state-of-the-art results in many practical signal processing applications. The cost — as in the case of the KLT — is a dictionary with no known inner structure or fast implementation. Thus, the most recent contributions to the field employ *parametric* models in the training process, which produce structured dictionaries, and offer several advantages. A different development which we mention here (though we do not discuss the topic further) is the recent advancement in *online* dictionary learning [76], [77], which allows training dictionaries from very large sets of examples, and is found to accelerate convergence and improve the trained result.

A. Method of Optimal Directions

The Method of Optimal Directions (MOD) was introduced by Engan *et al.* in 1999 [78], [79], and was one of the first methods to implement what is known today as a *sparsification process*. Given a set of examples $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_n]$, the goal of the MOD is to find a dictionary $\mathbf{D}$ and a sparse matrix $\mathbf{\Gamma}$ which minimize the representation error,

$$\underset{\mathbf{D},\mathbf{\Gamma}}{\text{Argmin}} \|\mathbf{X} - \mathbf{D}\mathbf{\Gamma}\|_F^2 \quad \text{Subject To} \quad \|\boldsymbol{\gamma}_i\|_0^0 \leq T \quad \forall i\ , \qquad (6)$$

where $\boldsymbol{\gamma}_i$ represent the columns of $\mathbf{\Gamma}$, and the ℓ^0 sparsity measure $\|\cdot\|_0^0$ counts the number of non-zeros in the representation. The resulting optimization problem is combinatorial and highly non-convex, and thus we can only hope for a local

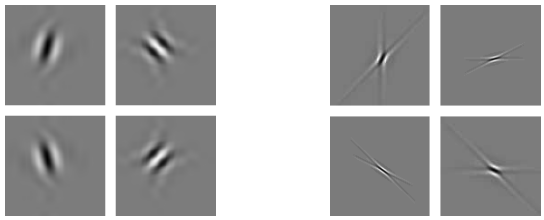

Fig. 6. Left: a few complex wavelet atoms (real part). Right: a few shearlets (extracted from [69]).

minimum at best. Similar to other training methods, the MOD alternates between sparse-coding and dictionary update steps. The sparse-coding is performed for each signal individually using any standard technique. For the dictionary update, (6) is solved via the analytic solution of the quadratic problem, given by $\mathbf{D} = \mathbf{X}\mathbf{\Gamma}^{+}$ with $\mathbf{\Gamma}^{+}$ denoting the Moore-Penrose pseudo-inverse.

The MOD typically requires only a few iterations to converge, and is overall a very effective method. The method suffers, though, from the relatively high complexity of the matrix inversion. Several subsequent works have thus focused on reducing this complexity, leading to more efficient methods.

B. Union of Orthobases

Training a union-of-orthobases dictionary was proposed in 2005 by Lesage *et al.* [80] as a means of designing a dictionary with reduced complexity and which could be more efficiently trained. The process also represents one of the first attempts at training a *structured* overcomplete dictionary — a tight frame in this case. The model suggests training a dictionary which is the concatenation of k orthogonal bases, so $\mathbf{D} = [\mathbf{D}_1\,\mathbf{D}_2\ldots\mathbf{D}_k]$ with the $\{\mathbf{D}_i\}$ unitary matrices. Sparse-coding over this dictionary can be performed efficiently through a Block Coordinate Relaxation (BCR) technique [81].

A drawback of this approach is that the proposed model itself is relatively restrictive, and in practice it does not perform as well as more flexible structures. Interestingly, there is a close connection between this structure and the more powerful *Generalized PCA* model, described next. As the GPCA model deviates from the classical sparse representation paradigm, identifying such relations could prove valuable in allowing the merge of the two forces.

C. Generalized PCA

Generalized PCA, introduced in 2005 by Vidal, Ma and Sastry [82], offers a different and very interesting approach to overcomplete dictionary design. The GPCA view is basically an extension of the original PCA formulation, which approximates a set of examples by a low-dimensional subspace. In the GPCA setting, the set of examples is modeled as the union of *several* low-dimensional subspaces — perhaps of unknown number and variable dimensionality — and the algebraic-geometric GPCA algorithm determines these subspaces and fits orthogonal bases for them.

The GPCA viewpoint differs from the sparsity model described in (2), as each example in the GPCA setting is represented using the atoms corresponding to only one of the subspaces; thus, atoms from different subspaces cannot jointly represent a signal. This property has the advantage of limiting over-expressiveness of the dictionary, which characterizes other overcomplete dictionaries; on the other hand, the dictionary structure may be too restrictive for more complex natural signals.

A unique property of the GPCA is that as opposed to other training methods, it can detect the *number* of atoms in the dictionary in certain settings. Unfortunately, the algorithm may become very costly this way, especially when the amount and dimension of the subspaces increases. Indeed, intriguing models arise by merging the GPCA viewpoint with the classical sparse representation viewpoint: for instance, one could easily envision a model generalizing (6) where several distinct dictionaries are allowed to co-exists, and every signal is assumed to be sparse over exactly one of these dictionaries.

D. The K-SVD Algorithm

The desire to efficiently train a generic dictionary for sparse signal representation led Aharon, Elad and Bruckstein to develop the K-SVD algorithm in 2005 [83]. The algorithm aims at the same sparsification problem as the MOD (6), and employs a similar block-relaxation approach. The main contribution of the K-SVD is that the dictionary update, rather than using a matrix inversion, is performed atom-by-atom in a simple and efficient process. Further acceleration is provided by updating both the current atom and its associated sparse coefficients simultaneously. The result is a fast and efficient algorithm which is notably less demanding than the MOD.

The K-SVD algorithm takes its name from the Singular-Value-Decomposition (SVD) process that forms the core of the atom update step, and which is repeated K times, as the number of atoms. For a given atom k, the quadratic term in (6) is rewritten as

$$\|\mathbf{X} - \sum_{j\neq k}\mathbf{d}_j\boldsymbol{\gamma}_j^T - \mathbf{d}_k\boldsymbol{\gamma}_k^T\|_F^2 = \|\mathbf{E}_k - \mathbf{d}_k\boldsymbol{\gamma}_k^T\|_F^2\ , \qquad (7)$$

where $\boldsymbol{\gamma}_j^T$ are the *rows* of $\mathbf{\Gamma}$, and $\mathbf{E}_k$ is the residual matrix. The atom update is obtained by minimizing (7) for $\mathbf{d}_k$ and $\boldsymbol{\gamma}_k^T$ via a simple rank-1 approximation of $\mathbf{E}_k$. To avoid introduction of new non-zeros in $\mathbf{\Gamma}$, the update process is performed using only the examples whose current representations use the atom $\mathbf{d}_k$. Figure 7 shows an example of a K-SVD trained dictionary for 2-D image patch representation.

The K-SVD, as well as the MOD, suffer from a few common weaknesses. The high non-convexity of the problem means that the two methods will get caught in local minima or even saddle points. Also, the result of the training is a non-structured dictionary which is relatively costly to apply, and therefore these methods are suitable for signals of relatively small size, such as image patches. In turn, in recent years several *parametric* dictionary training methods have begun to appear, and aim to address these issues by importing the strengths of analytic dictionaries to the world of example-based methods.

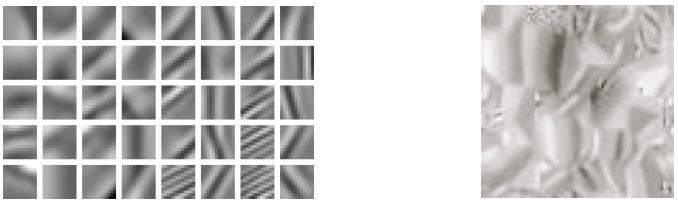

Fig. 7. Left: Atoms from a K-SVD dictionary trained on 12×12 image patches from *Lena*. Right: A signature dictionary, trained on the same image.

E. Parametric Training Methods

There are several motivations for training a parameterized dictionary. By reducing the number of free parameters and imposing various desirable properties on the dictionary, we can accelerate convergence, reduce the density of local minima, and assist in converging to a better solution. A smaller number of parameters also improves generalization of the learning process and reduces the number of examples needed. Another advantage of the parameterization is that the dictionary will typically have a more compact representation, and may lend itself to a more efficient implementation. Finally, with the proper structure, a parameterized dictionary may be designed to represent infinite or arbitrary-sized signals. Several parametric dictionary structures have been recently proposed, and we mention a few examples.

Translation-Invariant Dictionaries: Given a dictionary for a fixed-size signal patch, a dictionary for an arbitrary-sized signal can be constructed by collecting all the translations of the trained atoms over the signal domain and forming a large translation-invariant dictionary. Several training methods for such structures have been proposed in recent years. Blumensath and Davies [84] employed statistical training methodology to design dictionaries for time series representation; Jost *et al.* [85] developed a learning process based on a unique sequential computation of the dictionary atoms, and employed it to signals and images; finally, Engan *et al.* [86] extended the original MOD method to translation-invariant and optionally linearly-constrained dictionary training, which they successfully applied to electrocardiogram (ECG) recordings.

A different and unique approach to translation-invariance was recently proposed by Aharon and Elad in [87]. In the 2-D case, their proposed *signature dictionary* is a small image in which each $N \times N$ block constitutes an atom (see Figure 7). Thus, assuming a periodic extension, an $M \times M$ signature dictionary stores M^2 atoms in a compact structure. Compared to the previous methods, this approach does not aim to produce a dictionary for arbitrary-sized signals, and instead, describes an interesting form of invariance at the block level. A possible extension of this model could allow extraction of variable-sized atoms from the signature image, though this option remains for future research.

Multiscale Dictionaries: Training dictionaries with multi-scale structures is an exciting and challenging option which has been only partially explored. Sallee and Olshausen [88] proposed a pyramidal wavelet-like signal expansion, generated from the dilations and translations of a set of elementary small trained patches. The training method learns the elementary patches as well as a statistical model of the coefficients. In simulations, the structure was found to compete favorably with other pyramidal-based transforms. While the results of this method seem slightly constrained by the small number of elementary functions trained, it is likely to substantially benefit from increasing the overcompleteness and employing some more advanced sparse-coding machinery.

Another interesting contribution in this direction is the semi-multiscale extension of the K-SVD introduced in 2008 by Mairal, Sapiro and Elad [89]. The semi-multiscale structure is obtained by arranging several fixed-sized learned dictionaries of different scales over a dyadic grid. The resulting structure was found to deliver a pronounced improvement over the single-scale K-SVD dictionary in applications such as denoising and inpainting, producing nearly state-of-the-art denoising performance. The main significance of this work, though, is the potential it demonstrates in going to multi-scale learned structures. Such results are highly encouraging, and motivate further research into multi-scale training models.

Sparse Dictionaries: One of the most recent contributions to the field of parametric dictionaries, specifically aimed at merging the advantages of trained and analytic dictionaries, was recently provided by Rubinstein, Zibulevsky and Elad [90]. Their proposed *sparse dictionary* takes the form $\mathbf{D} = \mathbf{BA}$, where $\mathbf{B}$ is some fixed analytic dictionary with a fast computation, and $\mathbf{A}$ is a sparse matrix. Thus, the dictionary is compactly expressed and has a fast implementation, while adaptivity is provided through the matrix $\mathbf{A}$. Also, the parameterization of the dictionary is shown to improve learning generalization and to reduce the training set size. All this enables the training method to learn significantly larger dictionaries than the MOD or K-SVD, such as for large image patches, or 3-D signal patches. Nonetheless, the sparse dictionary structure remains targeted at fixed-size signals, and indeed further work is required to design more general dictionary models which will truly capture the benefits of both analytic and example-based worlds.

Non-Adaptive Parameter Tuning: We conclude with a very recent contribution to parametric dictionary design due to Yaghoobi, Daudet and Davies [91]. This work, which represents a very different approach to the design problem, assumes a *pre-selected* family of dictionaries characterized by the choice of a small set of values — for instance, we have seen the Gabor and wavelet dictionary families which are controlled by the resolution parameters α and β. Yaghoobi *et al.* propose a *non-adaptive* method for selecting the "best" set of parameters for the given family, such that the resulting dictionary is as close as possible to a Grassmannian frame. In other words, the algorithm aims to minimize the correlation between the dictionary atoms — corresponding to the off-diagonals of $\mathbf{D}^T\mathbf{D}$ — which is a feature known to be favorable for sparse-coding techniques [42]. The main advantage of this method is that by applying it to existing analytic dictionaries, it can produce dictionaries with efficient implementations which are specifically optimized for sparse-coding tasks. Another interesting option may be to incorporate this machinery in a parametric *example-based* training method,

leading to an adaptive learning process which also promotes well-conditioned results.

V. Conclusions

Dictionary design has significantly evolved over the past decades, beginning with simple orthogonal transforms and leading to the complex overcomplete analytic and trained dictionaries now defining the state-of-the-art. Substantial conceptual advancement has been made in understanding the elements of an efficient dictionary design — most notably adaptivity, multi-scale, geometric invariance, and overcompleteness. However, with a wealth of tools already developed, much work remains to be done; indeed, the various components have yet to be neatly merged into a single efficient construct. Many future research directions have been mentioned in the text, and demonstrate the viability and vividness of the field as well as the large number of challenges that still await. Of specific interest, we highlight the strong need for a multi-scale structured dictionary learning paradigm, as well as methods to use such dictionaries in applications, which will clearly be the focus of much research in the near future.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable and enlightening comments, which substantially enhanced the final result.

Images in this paper were generated using several software packages, and the authors would like to acknowledge the writers of these packages and thank them for their contribution and support of reproducible research. *In order of appearance*: Images of the curvelet transform were generated using the CurveLab toolbox curtesy of Candès, Demanet, Donoho and Ying (http://www.curvelet.org); images of the contourlet transform were generated using the SurfBox toolbox curtesy of Y. M. Lu (http://lcav.epfl.ch/~lu); images related to the bandelet transform were generated using the Bandelet Toolbox curtesy of G. Peyré (http://www.cmap.polytechnique.fr/~peyre/bandelets); and images of the complex wavelet transform were generated using the wavelet software curtesy of Cai, Li and Selesnick (http://taco.poly.edu/WaveletSoftware).

References

[1] P. J. Huber, *Robust statistics*. Wiley, New York, 1981.
[2] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, pp. 297–301, 1965.
[3] D. J. Field, "What is the goal of sensory coding?," *Neural Computation*, vol. 6, no. 4, pp. 559–601, 1994.
[4] A. K. Jain, *Fundamentals of digital image processing*. Prentice-Hall, 1989.
[5] S. Mallat, *A wavelet tour of signal processing, third ed.* Academic Press, 2009.
[6] I. T. Jolliffe, *Principal component analysis, second ed.* Springer, New York, 2002.
[7] R. A. DeVore, "Nonlinear approximation," *Acta Numerica*, vol. 7, pp. 51–150, 1998.
[8] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.
[9] W. B. Pennebaker and J. L. Mitchell, *JPEG still image data compression standard*. Springer, New York, 1993.
[10] D. Gabor, "Theory of communication," *J. Inst. Electr. Eng*, vol. 93, no. 26, pp. 429–457, 1946.
[11] M. J. Bastiaans, "Gabor's expansion of a signal into Gaussian elementary signals," *Proc. IEEE*, vol. 68, no. 4, pp. 538–539, 1980.
[12] A. Janssen, "Gabor representation of generalized functions," *J. Math. Anal. and Applic.*, vol. 83, no. 2, pp. 377–394, 1981.
[13] I. Daubechies, A. Grossmann, and Y. Meyer, "Painless nonorthogonal expansions," *J. Math. Phys.*, vol. 27, p. 1271, 1986.
[14] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 961–1005, 1990.
[15] H. G. Feichtinger and K. Gröchenig, "Banach spaces related to integrable group representations and their atomic decompositions, part I," *J. Funct. Anal.*, vol. 86, no. 2, pp. 307–340, 1989.
[16] H. G. Feichtinger and K. Gröchenig, "Gabor wavelets and the Heisenberg group: Gabor expansions and short time Fourier transform from the group theoretical point of view," *Wavelets: A Tutorial in Theory and Applications, C.K. Chiu (ed.)*, pp. 359–397, 1992.
[17] H. G. Feichtinger and K. Gröchenig, "Gabor frames and time-frequency analysis of distributions," *J. Funct. Anal.*, vol. 146, no. 2, pp. 464–495, 1997.
[18] J. Wexler and S. Raz, "Discrete gabor expansions," *Signal processing*, vol. 21, no. 3, pp. 207–221, 1990.
[19] S. Qian and D. Chen, "Discrete gabor transform," *IEEE Trans. Sigal Process.*, vol. 41, no. 7, pp. 2429–2438, 1993.
[20] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles.," *Vision research*, vol. 20, no. 10, pp. 847–856, 1980.
[21] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Am. A*, vol. 2, no. 7, pp. 1160–1169, 1985.
[22] J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," *IEEE Trans. Acoust. Speech Sigal Process.*, vol. 36, no. 7, pp. 1169–1179, 1988.
[23] M. Porat and Y. Y. Zeevi, "The generalized Gabor scheme of image representation in biological and machine vision," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 10, no. 4, pp. 452–468, 1988.
[24] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, 1983.
[25] I. Daubechies, *Ten lectures on wavelets*. Society for Industrial Mathematics, 1992.
[26] Y. Meyer and D. Salinger, *Wavelets and operators*. Cambridge University Press, 1995.
[27] J. Morlet and A. Grossman, "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM J. Math. Anal.*, vol. 15, pp. 723–736, 1984.
[28] Y. Meyer, "Principe dincertitude, bases hilbertiennes et algèbres dopérateurs," *Séminaire Bourbaki*, no. 662, 1985-86.
[29] P. G. Lemarie and Y. Meyer, "Ondelettes et bases hilbertiennes," *Rev. Mat. Iberoamericana*, vol. 2, no. 1-2, pp. 1–18, 1986.
[30] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Comm. Pure Appl. Math*, vol. 41, no. 7, pp. 909–996, 1988.
[31] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 11, no. 7, pp. 674–693, 1989.
[32] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. Inf. Theo.*, vol. 38, no. 2, pp. 617–643, 1992.
[33] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 14, no. 7, pp. 710–732, 1992.
[34] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.
[35] R. R. Coifman, Y. Meyer, and V. Wickerhauser, "Wavelet analysis and signal processing," *Wavelets and their Applications*, pp. 153–178, 1992.
[36] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *IEEE Trans. Inf. Theory*, vol. 38, no. 2 part 2, pp. 587–607, 1992.
[37] G. Beylkin, "On the representation of operators in bases of compactly supported wavelets," *SIAM Journal on Numerical Analysis*, pp. 1716–1740, 1992.
[38] G. P. Nason and B. W. Silverman, "The stationary wavelet transform and some statistical applications," *Lecture Notes in Statistics 103: Wavelets and Statistics (Ed. A. Antoniadis and G. Oppenheim)*, pp. 281–299, 1995.
[39] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising," *Lecture Notes in Statistics 103: Wavelets and Statistics (Edited by A. Antoniadis and G. Oppenheim)*, pp. 125–150, 1995.

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

[41] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *Technical Report – Statistics, Stanford*, 1995.

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

[43] E. J. Candès and D. L. Donoho, "Curvelets – a surprisingly effective nonadaptive representation for objects with edges," *Curves and Surfaces*, 1999.

[44] D. L. Donoho, "Wedgelets: nearly minimax estimation of edges," *Annals of statistics*, vol. 27, no. 3, pp. 859–897, 1999.

[45] M. B. Wakin, J. K. Romberg, H. Choi, and R. G. Baraniuk, "Wavelet-domain approximation and compression of piecewise smooth images," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1071–1087, 2006.

[46] R. M. Willett and R. D. Nowak, "Platelets: a multiscale approach for recovering edges and surfaces in photon-limited medical imaging," *IEEE Trans. Med. Imaging*, vol. 22, no. 3, pp. 332–350, 2003.

[47] V. Chandrasekaran, M. B. Wakin, D. Baron, and R. G. Baraniuk, "Representation and compression of multi-dimensional piecewise functions using surflets," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 374–400, 2009.

[48] E. J. Candès and D. L. Donoho, "Ridgelets: a key to higher-dimensional intermittency?," *Philosophical Transactions A: Mathematical, Physical and Engineering Sciences*, vol. 357, no. 1760, pp. 2495–2509, 1999.

[49] J. L. Starck, E. J. Candès, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Process.*, vol. 11, no. 6, pp. 670–684, 2002.

[50] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural computation*, vol. 12, no. 2, pp. 337–365, 2000.

[51] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T. W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, no. 2, pp. 349–396, 2003.

[52] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457.

[53] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying, "Fast discrete curvelet transforms," *Multiscale Modeling & Simulation*, vol. 5, pp. 861–899, 2006.

[54] L. Ying, L. Demanet, and E. J. Candès, "3D discrete curvelet transform," in *Proc. SPIE: Wavelets XI*, vol. 5914, pp. 351–361, 2005.

[55] E. J. Candès and D. L. Donoho, "Continuous curvelet transform: I. Resolution of the wavefront set," *Appl. Comput. Harmon. Anal*, vol. 19, no. 2, pp. 162–197, 2005.

[56] A. Woiselle, J. L. Starck, and M. J. Fadili, "New 3D data representations: applications in astrophysics," To appear.

[57] M. N. Do and M. Vetterli, "Contourlets: a new directional multiresolution image representation," in *Signals, Systems and Computers, 2002. Conference Record of the Thirty-Sixth Asilomar Conference on*, vol. 1, pp. 497–501, 2002.

[58] M. N. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2091–2106, 2005.

[59] Y. Lu and M. N. Do, "A new contourlet transform with sharp frequency localization," in *Proc. IEEE Int. Conf. on Image Proc*, pp. 1629–1632, 2006.

[60] Y. M. Lu and M. N. Do, "Multidimensional directional filter banks and surfacelets," *IEEE Trans. Image Process.*, vol. 16, no. 4, pp. 918–931, 2007.

[61] A. L. da Cunha, J. Zhou, and M. N. Do, "The nonsubsampled contourlet transform: theory, design, and applications.," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 3089–3101, 2006.

[62] R. Eslami and H. Radha, "Translation-invariant contourlet transform and its application to image denoising," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3362–3374, 2006.

[63] E. LePennec and S. Mallat, "Sparse geometric image representations with bandelets," *IEEE Trans. Image Process.*, vol. 14, no. 4, pp. 423–438, 2005.

[64] G. Peyré and S. Mallat, "Surface compression with geometric bandelets," in *ACM Transactions on Graphics (Proc. SIGGRAPH 05)*, vol. 24, pp. 601–608, 2005.

[65] N. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Appl. Comput. Harmon. Anal.*, vol. 10, no. 3, pp. 234–253, 2001.

[66] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Process Mag.*, vol. 22, no. 6, pp. 123–151, 2005.

[67] D. Labate, W. Lim, G. Kutyniok, and G. Weiss, "Sparse multidimensional representation using shearlets," in *Proc. SPIE: Wavelets XI*, vol. 5914, pp. 254–262, 2005.

[68] G. Kutyniok and D. Labate, "Resolution of the wavefront set using continuous shearlets," *Trans. Amer. Math. Soc.*, vol. 361, pp. 2719–2754, 2009.

[69] G. Easley, D. Labate, and W. Lim, "Sparse directional image representations using the discrete shearlet transform," *Appl. Comput. Harmon. Anal.*, vol. 25, pp. 25–46, 2008.

[70] V. Velisavljevic, B. Beferull-Lozano, M. Vetterli, and P. L. Dragotti, "Directionlets: anisotropic multidirectional representation with separable filtering," *IEEE Trans. Image Process.*, vol. 15, no. 7, pp. 1916–1933, 2006.

[71] S. Mallat, "Geometrical grouplets," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 2, pp. 161–180, 2009.

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

[73] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, 1997.

[74] D. L. Donoho, Y. Tsaig, I. Drori, and J. L. Starck, "Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit," 2007. Submitted.

[75] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2009.

[76] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.* To appear.

[77] K. Skretting and K. Engan, "Recursive least squares dictionary learning algorithm," *IEEE Trans. Signal Process.* To appear.

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

[79] K. Engan, B. D. Rao, and K. Kreutz-Delgado, "Frame design using FOCUSS with method of optimal directions (MOD)," *Proc. Norwegian Signal Processing Symposium*, pp. 65–69, 1999.

[80] S. Lesage, R. Gribonval, F. Bimbot, and L. Benaroya, "Learning unions of orthonormal bases with thresholded singular value decomposition," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, pp. 293–296, 2005.

[81] S. Sardy, A. G. Bruce, and P. Tseng, "Block coordinate relaxation methods for nonparametric wavelet denoising," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 361–379, 2000.

[82] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1945–1959, 2005.

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

[84] T. Blumensath and M. Davies, "Sparse and shift-invariant representations of music," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 1, p. 50, 2006.

[85] P. Jost, P. Vandergheynst, S. Lesage, and R. Gribonval, "MoTIF: An efficient algorithm for learning translation invariant dictionaries," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, 2006.

[86] K. Engan, K. Skretting, and J. H. Husøy, "Family of iterative LS-based dictionary learning algorithms, ILS-DLA, for sparse signal representation," *Digital Signal Processing*, vol. 17, no. 1, pp. 32–49, 2007.

[87] M. Aharon and M. Elad, "Sparse and redundant modeling of image content using an image-signature-dictionary," *SIAM Journal on Imaging Sciences*, vol. 1, no. 3, pp. 228–247, 2008.

[88] P. Sallee and B. A. Olshausen, "Learning sparse multiscale image representations," *Adv. Neural Inf. Process. Syst.*, vol. 15, pp. 1327–1334, 2003.

[89] 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, 2008.

[90] R. Rubinstein, M. Zibulevsky, and M. Elad, "Learning sparse dictionaries for sparse signal representation," *IEEE Trans. Signal Process.* To appear.

[91] M. Yaghoobi, L. Daudet, and M. E. Davies, "Parametric dictionary design for sparse coding," *IEEE Trans. Signal Process.* To appear.