Applied Mathematics Meets Signal Processing

Stéphane Mallat

1991 Mathematics Subject Classification: Primary 11E16; Secondary 11D09, 11E04, 15A63.
Keywords and Phrases: Image compression, Markov random fields, noise removal, non linear approximations, sparse representations, wavelets.

1 Beyond Fourier

The Fourier transform has long ruled over signal processing, leaving little space for new challenging mathematics. Until the 70's, signals were mostly speech and other sounds, which were modeled as realizations of Gaussian processes. As a result, linear algorithms were considered optimal over all procedures. With a hypothesis of stationarity, we end-up restricting ourselves to the exclusive class of convolution operators that are diagonalized by the Fourier transform.

The situation has completely changed with the development of image processing in the 1980's. Images are poorly modeled by Gaussian processes, and transient structures such as edges are often more important than stationary properties. Non-linear algorithms were suddenly unavoidable, opening signal processing to modern mathematics. Beyond classical applications to transmission, coding and restoration, signal processing also entered the field of information analysis, whose main branches are speech understanding and computer vision. This interface with perception raised a rich body of new mathematical problems.

The construction of sparse representations for signals (functions), processes, and operators is at the root of many signal and information processing problems. A sparse representation characterizes an approximation with few parameters, that may be obtained from an expansion in a basis or in a more redundant "dictionary". Complex non-linear processings can often be reduced to simpler and faster algorithms over such representations. Sparse representations are also powerful tools which radiate in many branches of mathematics. At ICM'90, Coifman and Meyer gave a harmonic analysis point view, followed at ICM'94 by Daubechies and Donoho who explained the impact of wavelet bases in numerical analysis and statistics. Signal processing is now a driving force that has regrouped a community of mathematicians and engineers sharing representation techniques. Applications to signal compression, noise removal, and stochastic modeling lead us through recent developments in approximation theory, harmonic analysis, operator theory, probability, and statistics.

2 Sparse Representations

Sparse representations have direct applications to data compression, but are also necessary to reduce the complexity of classification and identification problems for large size signals. This section begins with an approximation theory point of view, and progresses towards signal compression.

2.1 Image Models

A Bayesian view of the world interprets a signal $f(x)$ as a realization of a process $F(x)$ and the error of a processing is measured in expected value with respect to the probability distribution of F. Natural images are realizations of non-Gaussian processes, and there is yet no stochastic model that incorporates the diversity of complex scenes with edges and textures, such as the Image 1(a). This motivates the use of poorer but more realistic deterministic models that consider signals as functions $f(x)$ in a subset $\mathcal{S}$ of $\mathbf{L^2}[0,1]^d$, with no prior information on their probability distribution in this set. For a particular processing, one then tries to minimize the maximum error for signals in $\mathcal{S}$, which is the *minimax* framework. The discretization of a signal f with N samples raises no difficulty since it is equivalent to a projection in a subspace of dimension N.

Large class of images, including Image 1(a), have bounded total variation. Over $[0,1]$ the total variation of $f(x)$ measures the sum of the amplitudes of its oscillations

$$\|f\|_{TV} = \int |f'(x)|\, dx < +\infty \ .$$

The total variation of an image over $[0,1]^2$ is defined by

$$\|f\|_{TV} = \int\!\!\int |\vec{\nabla} f(x)|\, dx \leq C \ .$$

This norm has a simple geometrical interpretation based on the level-sets

$$\Omega_t = \{(x,y) \in \mathbb{R}^2 \quad : \quad f(x,y) > t\} \ .$$

If $H^1(\partial\Omega_t)$ is the one-dimensional Hausdorff measure of the boundary of Ω_t then

$$\|f\|_{TV} = \int_{-\infty}^{+\infty} H^1(\partial\Omega_t)\, dt. \tag{1}$$

A bounded variation model for images also incorporates a bounded amplitude

$$\mathcal{S}_{\mathbf{BV}} = \{f \ : \ \|f\|_{TV} = \int\!\!\int |\vec{\nabla} f(x)|\, dx \leq C \ , \ \|f\|_\infty = \sup_{x\in[0,1]^2} |f(x)| \leq C\} \ . \tag{2}$$

Such images typically have level sets and thus “contours” of finite length. Although simple, this model is sufficient to illustrate the central ideas and difficulties of signal representations. More restricted classes of images, such as homogeneous textures, are better represented by Markov random fields over sparse representations, introduced in Section 4.

2.2 Representations are Approximations

A sparse representation of $f \in \mathbf{L^2}[0,1]^d$ can be obtained by truncating its decomposition in an orthonormal basis $\mathcal{B} = \{g_m\}_{m\in\mathbb{N}}$

$$f = \sum_{m=0}^{+\infty} \langle f, g_m\rangle \, g_m \, .$$

Understanding the performance of sparse representations in a basis is a central topic of approximation theory. A quick overview motivates the use of non-linear representations, but a more complete tutorial is found in [13].

A linear approximation of f from M inner products $\langle f, g_m\rangle$ is an orthogonal projection on a space $\mathbf{V}_M$ generated from M vectors of $\mathcal{B}$, say the first M

$$f_M = P_{\mathbf{V}_M} f = \sum_{m=0}^{M-1} \langle f, g_m\rangle \, g_m \, .$$

The maximum approximation error over a signal set $\mathcal{S}$ is

$$\epsilon_l(\mathcal{S}, M) = \sup_{f\in\mathcal{S}} \|f - f_M\|^2 = \sup_{f\in\mathcal{S}} \sum_{m=M}^{+\infty} |\langle f, g_m\rangle|^2.$$

Such a representation is efficient if $\epsilon_l(\mathcal{S}, M)$ has fast decay as M decreases, and hence if $|\langle f, g_m\rangle|$ has fast decay as m increases. This depends upon the choice of $\mathcal{B}$ relative to $\mathcal{S}$. For example, uniformly regular functions are well approximated by M low-frequency vectors of a Fourier basis $\{\mathrm{e}^{i2\pi mx}\}_{m\in\mathbb{Z}}$ of $\mathbf{L^2}[0,1]$. If $\mathcal{S}$ is included in a ball of a Sobolev space $\mathbf{W}^s[0,1]$ of functions of period 1 then the decay of Fourier coefficients at high frequencies implies that $\epsilon_l(\mathcal{S}, M) = O(M^{-2s})$ [13]. Bounded variation functions may have discontinuities, and are thus not well approximated in a Fourier basis. Using the concept of *M-width* introduced by Kolmogorov, one can prove that for a ball $\mathcal{S}_{\mathbf{BV}}$ of bounded variation functions, the most rapid error decay in a basis $\mathcal{B}$ is $\epsilon_l(\mathcal{S}_{\mathbf{BV}}, M) \sim M^{-1}$ [13].

To improve this result, a more adaptive representation is constructed by projecting f over M basis vectors selected depending upon f

$$f_M = \sum_{m\in I_M} \langle f, g_m\rangle \, g_m \, . \tag{3}$$

Since $\|f - f_M\|^2 = \sum_{m\notin I_M} |\langle f, g_m\rangle|^2$, the best approximation is obtained by selecting in I_M the M vectors which yield coefficients $|\langle f, g_m\rangle|$ of maximum amplitude. This approximation depends upon $2M$ parameters, the M indexes in I_M and the values $\{\langle f, g_m\rangle\}_{m\in I_M}$. Let us sort the inner products of f in decreasing order. We denote $c_k = \langle f, g_{m_k}\rangle$ such that $|c_k| \geq |c_{k+1}|$ for $k \geq 1$. The non-linear approximation error is

$$\|f - f_M\|^2 = \sum_{k=M+1}^{+\infty} |c_k|^2 \quad\text{and}\quad \epsilon_n(\mathcal{S}, M) = \sup_{f\in\mathcal{S}} \|f - f_M\|^2 \, .$$

It depends upon the decay rate of the sorted amplitudes $|c_k|$. In the basis $\mathcal{B}$, a $w\mathbf{l^p}$ ball of radius C is defined by

$$\mathcal{S}_{wl^p} = \{f \ : \ |c_k| = |\langle f, g_{m_k}\rangle| \leq C\, k^{-1/p}\} \ . \tag{4}$$

We easily verify that $\mathcal{S} \subset \mathcal{S}_{wl^p}$ for some $C > 0$ and $p < 2$ if and only if $\epsilon_n(\mathcal{S}, M) = O(M^{1-\frac{2}{p}})$. The main difficulty of non-linear approximations is to find the minimum p and a corresponding basis $\mathcal{B}$ such that $\mathcal{S} \subset \mathcal{S}_{wl^p}$. Such a basis is said to be *optimal* for the non-linear approximation of $\mathcal{S}$. Unconditional bases are examples of optimal bases.

An orthonormal basis $\mathcal{B}$ is an *unconditional* basis of a Banach subspace $\mathbf{B} \subset \mathbf{L^2}[0,1]^d$ if there exists A such that for any sign sequence $s_m \in \{-1, 1\}$ and $f \in \mathbf{B}$

$$\left\| \sum_{m=0}^{+\infty} s_m \, \langle f, g_m\rangle \, g_m \right\|_{\mathbf{B}} \leq A \, \left\| \sum_{m=0}^{+\infty} \langle f, g_m\rangle \, g_m \right\|_{\mathbf{B}} \ .$$

The fact that $\|f\|_{\mathbf{B}} < +\infty$ can thus be characterized from the amplitudes $|\langle f, g_m\rangle|$, and related to a decay condition of the sorted coefficients. One can prove [13] that if $\mathcal{B}$ is an unconditional basis of $\mathbf{B}$ then it is an optimal basis for the non-linear approximation of a ball $\mathcal{S} = \{f \ : \ \|f\|_{\mathbf{B}} \leq C\}$ of $\mathbf{B}$.

2.3 Wavelet Adaptive Grid

Wavelet bases have important applications in mathematics and signal processing because of their ability to build sparse representations for large classes of functions. The first orthonormal wavelet bases of $\mathbf{L^2}(\mathbb{R})$ were introduced by Strömberg and Meyer [25]. A multiresolution interpretation of wavelet bases gives a general framework for constructing nearly all wavelets that generate a wavelet basis of $\mathbf{L^2}(\mathbb{R})$ [19]. It also leads to a fast discrete algorithm that requires $O(N)$ calculations to compute N wavelet coefficients [22]. Daubechies [9] discovered wavelets with compact support, and the resulting bases have been adapted to $\mathbf{L^2}[0,1]^d$. Her presentation at ICM'94 [10] introduces the main results, that we quickly summarize.

An orthonormal wavelet basis of $\mathbf{L^2}[0,1]$ is a family of functions

$$\mathcal{B} = \Big(\{\phi_{l,n}\}_{0 \leq n < 2^l} \ \cup \ \{\psi_{j,n}\}_{j \geq l, 0 \leq n < 2^j} \Big) \ .$$

At resolution 2^l, the *scaling functions* $\{\phi_{l,n}\}_{0 \leq n < 2^l}$ generate a space $\mathbf{V}_l$ which includes all polynomials of degree q, for some $q \geq 0$. The wavelets $\psi_{j,n}$ at higher resolutions $2^j > 2^l$ are thus orthogonal to all polynomials of degree q. Wavelets $\psi_{j,n}$ whose support lie inside $(0,1)$ are obtained by dilating and translating a single "mother" wavelet ψ

$$\psi_{j,n}(t) = \sqrt{2^j} \, \psi(2^j t - n) \ .$$

Boundary wavelets are modified to keep the support inside $[0,1]$.

A linear approximation of f from $M = 2^J > 2^l$ wavelets and scaling functions is calculated by keeping all coefficients at resolutions $2^j < 2^J$:

$$f_M = \sum_{n=0}^{2^l} \langle f, \phi_{l,n}\rangle \, \phi_{l,n} + \sum_{j=l}^{J-1} \sum_{n=0}^{2^j} \langle f, \psi_{j,n}\rangle \, \psi_{j,n}.$$

The first sum provides a coarse approximation of f at resolution 2^l, and each partial sum $\sum_{n=0}^{2^j} \langle f, \psi_{j,n} \rangle \, \psi_{j,n}$ brings "details" that improve this approximation from resolution 2^j to resolution 2^{j+1}. If f is continuous, this linear approximation at resolution 2^J is essentially equivalent to a uniform grid approximation calculated by interpolating the samples $\{f(2^{-J}n)\}_{0 \leq n < 2^J}$. Like a linear Fourier approximation, this uniform grid approximation is efficient only if f is uniformly regular. It provides poor approximations of functions with singularities, such as bounded variation functions.

A non-linear wavelet approximation keeps the M wavelet coefficients of largest amplitude. The amplitude of $|\langle f, \psi_{j,n} \rangle|$ depends upon the local regularity of f. Suppose that the mother wavelet ψ is $\mathbf{C}^{q+1}$ and orthogonal to polynomials of degree q. One can prove [25] that f is uniformly Lipschitz $\alpha < q+1$ over an interval $[a,b]$ if and only if there exists $A > 0$ such that for all $\psi_{j,n}$ whose support are included in $[a,b]$ (modulo boundary issues)

$$|\langle f, \psi_{j,n} \rangle| \leq A \, 2^{-(\alpha+1/2)j} \ .$$

In the domains where the Lipschitz regularity α is large, $|\langle f, \psi_{j,n} \rangle|$ decays quickly as the resolution 2^j increases. At high resolution 2^j, large coefficients appear in the neighborhood of singularities, where $0 \leq \alpha < 1$. More wavelet coefficients are kept in the neighborhood of singularities, so a non-linear wavelet approximation is equivalent to an adaptive grid whose resolution is refined in the neighborhood of singularities.

The impact of wavelet bases in functional analysis comes from the fact that they are unconditional bases of a large family of smoothness spaces (Besov spaces) [25], and are thus optimal for non-linear approximations in balls of these spaces. Although the space $\mathbf{BV}$ of bounded variation functions admits no unconditional basis, it can be embedded in two Besov spaces. This allows one to prove that wavelet bases are optimal to approximate a ball $\mathcal{S}_{\mathbf{BV}}$ of bounded variation functions. A ball $\mathcal{S}_{\mathbf{BV}}$ is included in a $w\mathbf{l^p}$ ball (4) for $p = 2/3$ but not for $p < 2/3$ [12]. Hence $\epsilon_n(\mathcal{S}_{\mathbf{BV}}, M) = O(M^{1-2/p}) = O(M^{-2})$. When M increases, the asymptotic decay of $\epsilon_n(\mathcal{S}_{\mathbf{BV}}, M)$ is thus faster than any linear approximation using M parameters, which decays at most like M^{-1}.

In two dimensions, wavelet bases are constructed with three "mother" wavelets ψ^k for $1 \leq k \leq 3$, which are dilated and translated

$$\psi^k_{j,n}(x_1, x_2) = \psi^k_{j,n}(x) = 2^j \, \psi^k(2^j x_1 - n_1, 2^j x_2 - n_2) \ .$$

Appropriate modifications are made at the boundary so that supports stay in $[0,1]^2$. A wavelet $\psi^k_{j,n}$ has a square support of size proportional to 2^{-j}, and centered near $2^{-j}n = (2^{-j}n_1, 2^{-j}n_2)$. An orthonormal wavelet basis of $\mathbf{L^2}[0,1]^2$ is obtained by adding orthonormal scaling functions that define a lower resolution space

$$\mathcal{B} = \Big(\{\phi_{l,n}\}_{2^{-l}n \in [0,1)^2} \ \cup \ \{\psi^k_{j,n}\}_{j \geq l \, , \, 2^{-j}n \in [0,1)^2 \, , \, 1 \leq k \leq 3} \Big) \ . \tag{5}$$

A discrete image is a square array of N^2 points (pixels), with $N = 512$ in Image 1(a). The wavelet basis (5) can be discretized to define an orthonormal basis of images of N^2 pixels. The wavelet coefficients of the image 1(a) are shown in 1(b). Each sub-image gives the values of $\{|\langle f, \psi^k_{j,n}\rangle|\}_{2^{-j}n\in[0,1)^2}$ for a fixed j and a fixed k. The number of wavelet coefficients in each sub-image is 2^{2j}. White and black points correspond respectively to nearly zero or large coefficients $|\langle f, \psi^k_{j,n}\rangle|$. These sub-images go by triplets corresponding to the index $1 \leq k \leq 3$. The wavelets for $k = 1, 2, 3$ are sensitive to image variations along different orientations. Most points are white, meaning that the majority of wavelet coefficients are nearly zero. The few large ones are located in the domains where the image intensity has a sharp variation due to an "edge" or a "texture".

(a) (b)

Figure 1: (a): Original image f. (b): Amplitude of coefficients $|\langle f, \psi^k_{j,n}\rangle|$ in a wavelet orthonormal basis. Each sub-image corresponds to a different resolution 2^j and different orientation k (see text).

A linear approximation from $M = 2^{2J}$ wavelets is calculated by keeping all coefficients at resolutions $2^j < 2^J$. This uniform grid approximation is particularly ineffective for images including discontinuities. For a ball of bounded variation images (2), one can prove that $\epsilon_l(\mathcal{S}_{\mathbf{BV}}, M) = A > 0$. The maximum approximation error does not decay to zero as M increases.

Non-linear approximations are much more effective because they keep wavelet coefficients near the singularities and (1) indicates that the lengths of "edges" remains finite. More formally, one can prove that $\mathcal{S}_{\mathbf{BV}}$ is included in a $w\mathbf{l}^1$ ball [4] and as a consequence $\epsilon_n(\mathcal{S}_{\mathbf{BV}}, M) = O(M^{-1})$. The wavelet adaptive grid gives much better image approximation than a uniform grid, and no other orthonormal basis can improve the approximation rate of an orthonormal wavelet basis.

2.4 Signal Compression

Economic storage and fast transmission of large signals through channels of limited bandwidth (such as Internet) are major applications of signal compression. Coding efficiently a signal with as few bits as possible requires to build a sparse representation. Signal processing engineers did not wait for a mathematical analysis of non-linear approximations in order to develop compressed audio or image codes in orthonormal bases. The first wavelet image coder was implemented in 1986 [34], before wavelet orthonormal bases had truly been studied in mathematics. The fast orthogonal wavelet transform is indeed computed with a "filter bank" algorithm, which was initially introduced in signal processing to *multiplex* signals (aggregate several signals into one) [7]. A discrete filter bank theory has been developed in signal processing [33], but only later the connection with wavelet orthonormal bases was established [19]. Although the mathematics came late, analyzing the performance of image coders requires use of recent approximation theory results, and these open directions for potential improvements.

The signals in $\mathcal{S}$ are now discretized and approximated at resolution N, which means that they belong to a space of dimension N. A *transform code* decomposes f in an orthonormal basis $\mathcal{B} = \{g_m\}_{0\leq m<N}$

$$f = \sum_{m=0}^{N-1} \langle f, g_m \rangle \, g_m \, ,$$

and approximates each coefficient $\langle f, g_m \rangle$ with a *quantized value*, which is coded with as few bits as possible. A uniform quantizer with *bin size* Δ approximates $x \in \mathbb{R}$ by $Q(x) = k\Delta$ with $k \in \mathbb{Z}$ and $|x - Q(x)| \leq \Delta/2$. The resulting quantized signal is

$$\tilde{f} = \sum_{m=0}^{N-1} Q(\langle f, g_m \rangle) \, g_m \, .$$

The problem is to minimize the maximum *distortion* $d(\mathcal{S}, R) = \sup_{f\in\mathcal{S}} \|f - \tilde{f}\|^2$ for a maximum number of bits R allocated to code $\tilde{f}$.

(a) (b)

Figure 2: (a): Image coded with 0.25 bits/pixel, by quantizing the wavelet coefficients of the original image displayed in Figure 1. (b): Image coded with 0.125 bits/pixel.

The distortion of a transform code is first related to a non-linear approximation. Let M be the number of coefficients above $\Delta/2$ and f_M the non-linear approximation of f from these M largest coefficients

$$f_M = \sum_{|\langle f, g_m \rangle| > \Delta/2} \langle f, g_m \rangle \, g_m \, .$$

Since $Q(x) = 0$ when $|x| < \Delta/2$, and $|x - Q(x)| \leq \Delta/2$, the distortion is

$$d(f,R) = \|f - \tilde{f}\|^2 \leq \|f - f_M\|^2 + M\,\frac{\Delta^2}{4}\ . \tag{6}$$

This connects us with non-linear approximations. Suppose that $\mathcal{S}$ is in a $w\mathbf{l^p}$ ball $\mathcal{S}_{wl^p}$ (4) of radius C, with $p < 2$. Denote by $M_0 = C^p (\Delta/2)^{-p}$. Since $|c_k| = |\langle f, g_{m_k}\rangle| \leq C\,k^{-1/p}$, necessarily $M \leq M_0$. We also verify that

$$d(\mathcal{S},R) = \sup_{f\in\mathcal{S}} d(f,R) \leq \sup_{f\in\mathcal{S}} \|f - f_{M_0}\|^2 + M_0\,\frac{\Delta^2}{4}\ = O(M_0^{-2/p+1})\ . \tag{7}$$

The total distortion is thus driven by the non-linear approximation error.

To optimize the transform code, we must minimize the maximum number of bits R required to code the N values $\{Q(\langle f, g_m\rangle)\}_{0\leq m<N}$ for $f \in \mathcal{S}$. For high compression rates $N \gg M_0 \geq M$, in which case a large proportion $\frac{N-M}{N}$ of coefficients quantized to zero. An *entropy code* takes advantage of this, by allocating fewer bits to code coefficients that occur more frequently than others. Knowing that $\mathcal{S} \subset \mathcal{S}_{wl^p}$, one can construct an arithmetic code which requires a maximum number of bits $R \sim M_0 \log_2 \frac{M_0}{N}$ [22]. We thus derive from (7) that $d(\mathcal{S},R) = O(R^{1-2/q})$ for any $q > p$.

The decay rate of $d(\mathcal{S},R)$ is maximized in a basis $\mathcal{B}$ which is optimal for non-linear approximations in $\mathcal{S}$, because it minimizes the exponent p such that $\mathcal{S} \subset \mathcal{S}_{wl^p}$ In particular, wavelet bases are optimal for bounded variation images and the minimum is $p = 1$. The Figures 2(a,b) are compressed images $\tilde{f}$ calculated by quantizing the wavelet coefficients in Figure 1(b). They are coded respectively with $\frac{R}{N} = 0.25$ bits/pixel and 0.125 bits/pixel, with an optimized coder for zero coefficients [30]. The original image 1(a) is coded with 8 bits/pixel, so this corresponds to compression factors of 32 and 64. For 0.25 bits/pixel, the distortions are hardly visible but become apparent for 0.125 bits/pixel.

Let us emphasize that the choice of basis depends entirely on the nature of the signals in $\mathcal{S}$. For sounds, totally different bases must be chosen in order to approximate efficiently complex oscillatory waveforms of varying durations. Figure 3 shows the recording of the word "greasy". Current compression audio standard for Compact Disk quality, such as the AC-system of Dolby, are calculated in bases that are similar to a local cosine basis. Such a basis is constructed with an even function $w(t)$, called a *window*, which has a support $[-2l, 2l]$ and is translated to cover the real axis uniformly:

$$\sum_{p=-\infty}^{+\infty} |w(t - p\,l)|^2 = 1.$$

Malvar [23], Coifman and Meyer [5] proved that if further symmetry properties are imposed on $w(t)$ then multiplications by cosine functions yield an orthonormal basis of $\mathbf{L^2}(\mathbb{R})$

$$\left\{ g_{p,k}(t) = \frac{1}{\sqrt{l}}\, w(t - p\,l)\, \cos\!\Big(\pi k\, (l^{-1} t - p)\Big) \right\}_{k\in\mathbb{N}, p\in\mathbb{Z}} . \tag{8}$$

As in the image case, the performance of an audio code in this basis depends on being able to approximate the recorded sound with few local cosine vectors. However, the most relevant audio distortions measures are not $\mathbf{L^2}$ norms. Sophisticated masking techniques are used by engineers to introduce quantization errors which are below our hearing sensitivity threshold [22], and above our mathematical understanding.

Figure 3: Speech recording of the word "greasy" sampled at 16kHz.

2.5 Geometry and More Adaptivity

Wavelet bases are optimal for representing general bounded variation images, but better approximations can be obtained by taking advantage of the geometrical regularity of most images. The total variation formula (1) shows that the level sets of bounded variation images typically have a finite length. However, this imposes no condition on the regularity of these level sets. In the Image 1(a), the "contours" are mostly piecewise regular geometrical curves in the image plane, with small curvature at most locations. Understanding how to take advantage of this regularity is fundamental for image processing. This has motivated the use of non-linear partial differential equations to modify the curvature of level sets in images [1, 28, 31]. This important new branch of mathematical image processing leads to interesting applications for noise removal and image segmentation. Yet, we shall not follow this line of thought, which is not based on explicit sparse representations.

To understand the importance of geometrical regularity, let us consider a simple "image" $f = \mathbf{1}_\Omega$, which is the indicator function of a set Ω. The boundary $\partial\Omega$ of Ω is a differentiable curve of finite length with bounded curvature. If the square support of $\psi^k_{j,n}$ does not intersect $\partial\Omega$ then $\langle f, \psi^k_{j,n}\rangle = 0$. The wavelets $\psi^k_{j,n}$ are translated on a square grid with step sizes 2^{-j} and have square support proportional to 2^{-j}, as illustrated in Figure 4(a). At resolution 2^j, there are $O(2^j)$ wavelets $\psi^k_{j,n}$ whose supports intersect $\partial\Omega$. The M larger amplitude wavelet coefficients selected by a non-linear approximation are at resolutions $2^j \leq 2^J \sim M$ and the non-selected wavelets produce an error $\|f - f_M\|^2 \sim M^{-2}$, like for any bounded variation image.

A better piecewise linear approximation is calculated with an adaptive triangulation of $[0,1]^2$ having M triangles [16]. Since the curvature of $\partial\Omega$ is bounded, this boundary can be covered by $M/2$ triangles, which have a narrow width proportional to M^{-2} along the normal to $\partial\Omega$, and which are elongated along the tangent to $\partial\Omega$. The interior and exterior of Ω are covered by $M/2$ larger triangles, as illustrated in Figure 4(b). A function f_M which is linear on each triangle can

(a) (b)

Figure 4: (a): Wavelets $\psi_{j,n}$ are translated on a square grid of interval 2^{-j}, and have a square support proportional to 2^{-j}. For $f = \mathbf{1}_\Omega$, the darker points locate the wavelets $\psi_{j,n}$ such that $\langle f, \psi_{j,n} \rangle \neq 0$. (b): A piecewise linear approximation of $f = \mathbf{1}_\Omega$ is optimized by choosing narrow triangles that are elongated along the boundary where f is discontinuous.

approximate $f = \mathbf{1}_\Omega$ with $\|f - f_M\|^2 = O(M^{-4})$. The approximation error is concentrated on the triangles along the border and the small width of these triangles yields a smaller error than with wavelets of square support. The error is reduced because the triangles are adapted to the geometry of $\partial\Omega$.

Building a bridge between geometrical constraints and adaptive approximations is a fundamental issue for image processing. The human visual system takes great notice of geometrical "features" such as "corners" or the regularity of "edges" [24, 26]. The Kanizsa illusion shown in Figure 5 illustrates this fact. We perceive a triangular "edge" although the image has no grey level variation in the center. Such illusions are explained by imposing geometrical constraints on the interpretation (models) of images. It is also known [11] that *simple cells* in the visual cortex perform an image decomposition over a family of functions that have close similarities to wavelets, but which is more redundant that a basis and thus offers more flexibility. This indicates that our brain constantly crosses this bridge between functional analysis and geometry.

(a) (b)

Figure 5: The illusory edges of a straight and of a curved triangle are perceived in domains where the images are uniformly white.

Adapting to geometry in images can be interpreted as a particular instance of a more general adaptive approximation problem. A basis is a complete family in our functional space, but it is often too small to fully utilize all of the structures included in complex signals. More precise approximations are obtained with M vectors selected from a much larger dictionary $\mathcal{D} = \{g_\gamma\}_{\gamma\in\Gamma}$, that may include an infinite number of bases. This follows the same idea that motivates someone to enlarge his vocabulary to build more concise and precise sentences. For recognition, is also often important to construct representations that have invariant properties, with respect to translation or affine transformations. This imposes some further conditions on the dictionary [20]. A dictionary for images can be constructed with wavelets whose supports have a parameterized elongation and an arbitrary orientation. Like the elongated triangles in Figure 4(b), the chosen wavelets can be adapted to the geometry of the level sets in the image. Audio signals are also more efficiently approximated with a dictionary of local cosine vectors such as (8), but where the window length l may be freely adapted to the duration of waveforms produced by attacks, harmonics or other transient events.

An adaptive representation is constructed from a dictionary $\mathcal{D}$ by selecting M vectors $\{g_{\gamma_k}\}_{1\leq k\leq M}$ to approximate f with a partial sum

$$f_M = \sum_{k=1}^{M} \alpha_k \, g_{\gamma_k} \ .$$

In the absence of orthogonality, finding the M vectors that minimize $\|f - f_M\|$ leads to a combinatorial explosion. Greedy pursuit algorithms have been developed to avoid this explosion [20], by selecting the vectors g_{γ_k} one by one from the dictionary, but their approximation performance is far from optimal [3, 13]. In structured dictionaries composed of orthonormal bases embedded in a tree, Coifman and Wickerhauser [6] have introduced dynamical programming algorithm that selects M vectors which define a "reasonable" but non optimal approximation. There is yet no approximation theory that can analyze the performance of these highly non-linear approximations and improve their performance.

Let us finally mention that enlarging the dictionary has a cost. In a larger dictionary, more parameters are needed to characterize the index γ_k of each selected vector. For a fixed approximation error, making the dictionary too large can increase the total number of parameters that characterize the signal approximation f_M. Finding dictionaries of optimal size is thus another open issue.

3 Noise Removal by Thresholding

The removal of noise, added when measuring the signal or during its transmission, is an important problem where sparse representations play a crucial role. In a basis that transforms the signal into few large amplitude values plus a small remainder, most of the noise is easily suppressed by a thresholding which sets to zero the smallest coefficients. A similar version of this simple idea has been used to remove noise from television images since the 1960's. However, it is only recently that Donoho and Johnstone [14] were able to develop the mathematics proving that thresholding estimators are nearly optimal in sparse representations, which opened new signal processing applications.

A discrete approximation of $f(x)$ defined over $[0,1]^d$ is characterized by N coefficients, denoted $f[n]$, for $0 \leq n < N$. The measured noisy data are

$$D[n] = f[n] + W[n] \ , \tag{9}$$

where the noise values $W[n]$ are modeled by independent Gaussian random variables, and thus define a *white noise*. Figure 6(a) gives an example. An estimator F of f is calculated by applying an operator L on the data, $F = LD$. The risk of this estimation is

$$r(L,f) = \mathsf{E}\{\|f - LD\|^2\}.$$

We want to minimize the maximum risk over a signal set $\mathcal{S}$

$$r(L,\mathcal{S}) = \sup_{f \in \mathcal{S}} r(L,f) \ .$$

The goal is to find an operator L which approaches the optimal minimax risk

$$r_o(\mathcal{S}) = \inf_{All\, L} r(L,\mathcal{S}).$$

There is a considerable body of literature in mathematical statistics for evaluation of minimax risk [15].

A new approach to minimax estimation is to separate the representation from the estimation problem. The first step is to construct an appropriate representation by decomposing $D = f + W$ in an orthogonal basis $\mathcal{B} = \{g_m\}_{0 \leq m < N}$:

$$\langle D, g_m \rangle \ = \ \langle f, g_m \rangle + \langle W, g_m \rangle.$$

A thresholding estimator is then simply defined by

$$F = L_t D = \sum_{m=0}^{N-1} \theta_T(\langle D, g_m \rangle)\, g_m, \tag{10}$$

where $\theta_T(x) = x\, \mathbf{1}_{|x|>T}$. It sets to zero all coefficients below T and keeps the others. The threshold T is chosen to be just above $\max_{0 \leq m < N} |\langle W, g_m \rangle|$, with a high probability, so that $\theta_T(\langle D, g_m \rangle) = 0$ if $\langle f, g_m \rangle \approx 0$.

Since W is a Gaussian white noise of variance σ^2, in any basis $\mathcal{B}$, the noise coefficients $\langle W, g_m \rangle$ are independent Gaussian random variables of same variance σ^2. Let M be the number of coefficients such that $|\langle f, g_m \rangle| \geq \sigma$, and f_M be the non-linear approximation (3) of f from these M largest vectors. If $T = \sigma\, \sqrt{2 \log_e N}$ then Donoho and Johnstone proved [14] that

$$r(L_t, f) \leq (2 \log_e N + 1) \left(\|f - f_M\|^2 + (M+1)\,\sigma^2 \right).$$

The right part of the upper bound is similar to the distortion (6) of a transform code. The risk is thus reduced by choosing a basis where there is a small number

(a) (b)

Figure 6: (a): Image contaminated by an additive Gaussian white noise. (b): Thresholding estimation calculated in a wavelet basis.

M of large amplitude coefficients above σ, which yield a small approximation error $\|f - f_M\|$. Once more we face the problem of finding a sparse but precise representation. Figure 6(b) is an estimation calculated by thresholding the wavelet coefficients of the noisy image shown in (a).

The asymptotic performance of tresholding estimators is calculated as the resolution N of the measurements increases to $+\infty$. For a given set $\mathcal{S}_0$ of signals $f(x)$, we look for an orthonormal basis $\mathcal{B}_0$ which is optimal for non-linear approximations. Suppose that $\mathcal{S}_0$ is a ball of a space $\mathbf{B}$, then we can choose $\mathcal{B}_0$ to be an unconditional basis of $\mathbf{B}$. The set $\mathcal{S}$ of discretized signals is obtained with a projection in dimension N. These signals are decomposed in the basis $\mathcal{B}$ derived from $\mathcal{B}_0$ through the same projection. As N increases, one can prove [15] that the thresholding estimator is nearly optimal in the sense that

$$r(L_t, \mathcal{S}) \leq O(\log N)\, r_o(\mathcal{S}) \ . \tag{11}$$

This result applies to discretized signals from Besov spaces, decomposed in a discrete wavelet basis. It is also valid for a set $\mathcal{S}_{\mathbf{BV}}$ of bounded variation signals decomposed in a wavelet basis, because $\mathbf{BV}$ is embedded in two Besov spaces which are close enough. In this case, the tresholding risk has faster asymptotic decay than the risk of any linear estimator as N increases.

The efficiency of thresholding estimators depends crucially on the approximation performance of the representation. To take advantage of complex signal structures, such as the geometrical regularity of some images, the thresholding must be calculated in more adaptive representations, as explained in Section 2.5. However, the minimax optimality of these highly adaptive estimators remains to be understood.

4 Sparse Interaction Processes

In many classification problems, including speech recognition and visual texture discrimination, the observed signal is modeled as the realization of a process that we need to characterize. This is difficult because the underlying process is often non Gaussian or non-stationary, and a single realization provides little data to identify it. It is therefore necessary to characterize these processes with few parameters in an appropriate representation, that can be estimated and used for the classification. After studying non stationary Gaussian processes, we consider more general Markov random field models.

4.1 Non Stationary Gaussian Processes

Gaussians processes provide resonable models for large class of signals, including speech recordings. A zero-mean Gaussian process $X(t)$ for $t \in \mathbb{R}$ is entirely characterized by its covariance $k(t,s) = \mathsf{E}\{X(t)\,X(s)\}$, which is the kernel of the covariance operator K:

$$Kf(t) = \int_{-\infty}^{+\infty} k(t,s)\, f(s)\, ds. \tag{12}$$

To estimate this covariance from few realizations, it is necessary to reduce the number of coefficients describing the kernel. This can be done by finding an orthonormal basis $\mathcal{B} = \{g_m\}_{m\in\mathbb{Z}}$ in which the matrix coefficients

$$\langle Kg_m, g_n\rangle = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} k(t,s)\, g_m(s)\, g_n(t)\, ds\, dt \tag{13}$$

have fast off-diagonal decay. These matrix values are the decomposition coefficients of the kernel $k(t,s)$ in a separable orthonormal basis $\{g_n(t)\, g_m(s)\}_{(n,m)\in\mathbb{Z}^2}$ of $\mathbf{L^2}(\mathbb{R}^2)$. Finding a sparse matrix represention is thus equivalent to approximating $k(t,s)$ with few non-zero coefficients in a separable basis. If the matrix coefficients have a sufficiently fast off-diagonal decay, then K is closely approximated (with a sup or a Hilbert Schmidt norm) by a narrow band matrix $\tilde{K}$ in $\mathcal{B}$, which is the covariance of a Gaussian process $\tilde{X}$ that approximates X [21]. Since $\tilde{K}$ has a band-matrix representation, for each $m \in \mathbb{N}$ there exists a neighborhood $\mathcal{N}_m$ which is a finite set of integers such that if $n \notin \mathcal{N}_m$ then

$$\langle \tilde{K} g_m, g_n\rangle = \mathsf{E}\{\langle \tilde{X}, g_m\rangle\, \langle \tilde{X}, g_n\rangle\} = 0.$$

Since $\langle \tilde{X}, g_m\rangle$ and $\langle \tilde{X}, g_n\rangle$ are jointly Gaussian random variables, they are independent because uncorrelated. The model $\tilde{X}$ of X has therefore a representation in $\mathcal{B}$ with coefficients that are dependent only in small neighborhoods, which is a particular case of Markov random field.

Writing the covariance operator K as a pseudo-differential operators is a powerful approach to find bases where the matrix coefficients have fast off-diagonal decay [25]. Let $\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(s)\, \mathrm{e}^{-i\omega s}\, ds$ be the Fourier transform of f. The *symbol* of the operator K is

$$\beta(t,\omega) = p.v. \int_{-\infty}^{+\infty} k(t, t-s)\, \mathrm{e}^{-i\omega s}\, ds\,.$$

Applying the Parseval formula to (12) yields

$$Kf(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \beta(t,\omega)\, \hat{f}(\omega)\, \mathrm{e}^{i\omega t}\, d\omega.$$

For example, if $\beta(t,\omega) = \sum_{p=0}^{P} a_p(t)\,(i\omega)^p$ then $K = \sum_{p=0}^{P} a_p(t)\,(\frac{d}{dt})^p$ is a differential operator with time varying coefficients. The process X is stationary if $k(t,s) = k(t-s)$, in which case $\beta(t,\omega) = \beta(\omega)$ is the spectrum of K. The Fourier transform is therefore an ideal tool to characterize stationary Gaussian processes. For non-stationary processes, one needs to relate the properties of $X(t)$ to the properties of $\beta(t,\omega)$, and derive a basis where K is approximated by a narrow band matrix.

Locally stationary processes $X(t)$ appear in many physical systems in which the mechanisms that produce random fluctuations change slowly in time or space [29]. Over short time intervals l, such processes can be approximated by a stationary one. This is the case for many components of speech or audio signals. Over a sufficiently short time interval, the throat behaves like a steady resonator which is excited by a stationary noise source. A simple class of locally stationary processes is obtained by imposing that there exists $A > 0$ such that for all $k, j \geq 0$

$$|\partial_t^k \partial_\omega^j \beta(t,\omega)| \leq A\, l^{j-k}\ .$$

We derive [21] the existence of a local cosine basis (8) in which the operator K is closely approximated by a narrow band matrix. The size l of each window is adapted to the interval of stationarity. When the length $l(t)$ of the interval of stationarity varies strongly in time, which is the case of audio signals, the resulting covariance operator has more complex properties and often does not belong to a classical family of pseudo-differential operators. Depending upon the regularity of $l(t)$, adapted local cosine bases can still provide sparse representations of such operators [21].

Multifractals provide useful models for signals having some self-similarity properties [27]. Among the many examples, let us mention economic records like the Dow Jones industrial average, physiological data including heart records, electromagnetic fluctuations in galactic radiation noise, some image textures, variations of traffic flow... A fractional Brownian motion $X(t)$ of Hurst exponent H is a canonical example of fractal Gaussian processes, whose increments are stationary and which is self-similar in the sense that $s^{-H}X(st)$ has the same probability distribution as $X(t)$, for all $s > 0$. The symbol of the covariance K of X is $\beta(t,\omega) = \lambda\,|\omega|^{-2H-1}$. This corresponds to a Calderón-Zygmund operator of the first generation [25], which is known to have fast off-diagonal decay in a wavelet basis. In signal processing, fractional Brownian motions are often approximated by a process $\tilde{X}$ whose covariance $\tilde{K}$ is diagonal in a wavelet basis, which leads to fast synthesis algorithms [27]. General conditions on $\partial_t^k \partial_\omega^j \beta(t,\omega)$ can be established to guarantee that K has fast off-diagonal decay in a wavelet basis [2]. Multifractional Brownian motions are examples with Hurst exponents that vary in time: $\beta(t,\omega) = \beta_0(t)\,|\omega|^{-2H(t)-1}$. Accurate estimations of $\beta_0(t)$ and $H(t)$ are obtained in wavelet bases.

When the process is uniformly locally stationary or multifractal, the basis which compresses the covariance matrix is known beforehand. For more complex non-stationary processes, this basis must also be estimated, given some prior information. This is an adaptive approximation problem, similar to the ones described in Section 2.5, although we approximate operators as opposed to functions. Best basis search algorithms have been introduced to perform such adaptive approximation of covariance operators [21], but these techniques are still in their infancy, and more work is needed to understand the properties of the resulting statistical estimators.

4.2 Markov Random Fields in Sparse Representations

The characterization and synthesis of visual textures is one of the most challenging low-level vision problem. Homogeneous visual textures such as images of woods, carpets or marbles, can be considered as stationary, but they are not Gaussian. Figure 7 gives two examples. It is necessary to model these processes with few parameters to hope identify them from a single realization. This is feasible since the human visual system can do it. The importance of this problem goes well beyond texture discrimination. Indeed, providing a general framework to model non-Gaussian processes is necessary to analyze the properties of various classes of signals such as financial time series or the velocity of turbulent fluids.

Markov random field models of textures have been proposed by Cross and Jain [8], but such models became computationally and mathematically attractive through the work of Geman and Geman [17], who introduced a stochastic relaxation algorithm for sampling Gibbs distributions. To simplify the presentation, we restrict ourselves to a random vector $X(n)$, where $n \in \mathbb{Z}^d$ varies over a grid $\mathcal{G}$ of size N. We define a neighborhood system $\mathcal{N} = \{\mathcal{N}_n\}_{n\in\mathcal{G}}$ such that $n \notin \mathcal{N}_n$ and $m \in \mathcal{N}_n$ if and only if $n \in \mathcal{N}_m$. For any $\mathcal{G}_0 \subset \mathcal{G}$, let $X(\mathcal{G}_0)$ denote the set of values taken by X over $\mathcal{G}_0$. We say that $p(X)$ is a *Markov random field* distribution with respect to $\mathcal{N}$ if

$$p\Big(X(n) \mid X(\mathcal{G} - \{n\})\Big) = p\Big(X(n) \mid X(\mathcal{N}_n)\Big).$$

A subset C of $\mathcal{G}$ is called a clique if every pair of elements in C are neighbors of each other. Let $\mathcal{C}$ be the set of all cliques. If X takes its values in a finite alphabet then the Hammersley-Clifford theorem proves that $p(X)$ is a Markov random field if and only if it can be written as a *Gibbs distribution* with respect to $\mathcal{N}$

$$p(X) = \frac{1}{Z} \exp\Big[-\sum_{C\in\mathcal{C}} \phi_C(X)\Big],$$

where Z is a normalization constant and ϕ_C is a *potential function* which depends only of the values of X in the clique C. Markov random field models have interesting applications to texture discrimination and image restoration, but limited success due to the difficulty to incorporate the long range interactions of image pixels. Several approaches have been introduced to circumvent this problem, including renormalization techniques [18].

Mumford and Zhu [35] introduced a different point of view by creating Markov random field models on a sparse representation of X, rather than on the sample values $X(n)$. Let $\mathcal{D} = \{g_\gamma\}_{\gamma\in\Gamma}$ be a dictionary of vectors, which can be an orthogonal basis or be more redundant. Let $X_\gamma = \langle X, g_\gamma\rangle$. A neighborhood system $\mathcal{N} = \{\mathcal{N}_\gamma\}_{\gamma\in\Gamma}$ is defined over Γ. For example, if $\mathcal{D} = \{\psi^k_{j,n}\}_{k,j,n}$ is a wavelet basis in two dimensions, the index $\gamma = (k, j, n)$ specifies the orientation k, the resolution 2^j and the position $2^{-j}n$ of the wavelet. The neighborhood $\mathcal{N}_{(k,j,n)}$ includes wavelets $\psi^{k'}_{j',n'}$ with $|j - j'| \leq 1$ and a position $2^{-j'}n'$ which is close to $2^{-j}n$. The multiresolution aspect of wavelet bases allows one to construct Markov random field models that incorporate short range and long range interactions.

To construct a Markov random field model X from observed signals $\{X^{obs}_p\}_{0\leq p<P}$, we compute average measurements over M cliques $\{C_m\}_{0\leq m<M}$ with potential functions ϕ_{C_m}

$$\mu^{obs}_{C_m} = \frac{1}{P}\sum_{p=1}^{P}\phi_{C_m}(X^{obs}_p) \ .$$

If X is stationary then a spatial averaging is done over all ϕ_{C_m} that perform identical calculations but at translated locations. These empirical averages are estimates of $\mathsf{E}\{\phi_{C_m}(X)\}$ for the model X that we construct. Most often, the cliques have at most two elements $C = \{\gamma, \gamma'\}$. Covariance measurements correspond to $\phi_C(X) = X_\gamma\, X_{\gamma'}$. However, different potential functions may be useful such as p^{th} order moments

$$\phi_C(X) = |X_\gamma|^p\, |X_{\gamma'}|^p \quad \text{for } p > 0 \ . \tag{14}$$

(a) (b) (c) (d)

Figure 7: (a): Observation of a uniform texture. (b): Realization of the wavelet Markov random field model calculated from (a). (c): The center shows an example of texture. (d): The center is identical to (c) whereas the periphery is a realization of a wavelet Markov random field model calculated from (c).

The *maximum entropy principle* suggests choosing $p(X)$ that achieves the maximum entropy

$$p(X) = \arg\max\{-\int p(X)\,\log p(X)\,dX\} \ .$$

under the constraints

$$\mathsf{E}\{\phi_{C_m}(X)\} = \int \phi_{C_m}(X)\,p(X)\,dX = \mu^{obs}_{C_m} \quad \text{for } 1 \leq m \leq M \ . \tag{15}$$

By maximizing the entropy, the resulting $p(X)$ is the "most uniform" distribution given the prior knowledge provided by the observation $\mu^{obs}_{C_m}$. It thus does not include more "information" than what is available. The solution is calculated with Lagrange multipliers

$$p(X, \Lambda) = \frac{1}{Z(\Lambda)} \exp\left(- \sum_{m=1}^{M} \lambda_m \, \phi_{C_m}(X) \right) . \tag{16}$$

The parameter vector $\Lambda = \{\lambda_m\}_{1 \leq m \leq M}$ is uniquely characterized by the constraints (15), if the potential functions satisfy a linear independence property.

If $\phi_{C_m}(X)$ are covariance measurements then (16) is the probability distribution of a Gaussian process, and if $\mathcal{D}$ is an orthonormal basis then Λ is calculated by inverting a band covariance matrix. The entropy maximization is a convex problem [17], but for general potential functions ϕ_{C_m} the vector Λ can not be calculated analytically. Numerical procedures compute Λ iteratively by estimating $\mathsf{E}_{p(X,\Lambda)}\{\phi_{C_m}(X)\}$, while updating Λ with a gradient descent to reach the conditions (15). Let us mention that the estimation of $\mathsf{E}_{p(X,\Lambda)}\{\phi_{C_m}(X)\}$ is performed with a Gibbs sampler or other Markov chain Monte Carlo methods, which are computationally expensive.

Mumford and Zhu [35], as well as Simoncelli and Portilla [32], use such Markov random fields to construct a model from a single observation of a texture. The Markov model of Simoncelli and Portilla is calculated in a wavelet basis, with constraints on covariance values and on moments (14) with $p = 1$. The textured image 7(a) is the only observation used to compute the parameters Λ of the model, with a stationarity assumption. The Figure 7(b) shows a realization of the resulting wavelet Markov model. It is remarkably close to the original texture, in the sense that visually it can not be distinguished preattentively, in less than 10^{-1} seconds. A similar wavelet Markov model is calculated from the "text" texture of Figure 7(c). The image 7(d) is obtained by adding a realization of this Markov model at the periphery, which is preattentively not discriminable from the center.

Markov random fields provide a general framework to construct processes with sparse interactions over appropriate representations. The validity of such models depends on the choice of representation and on the potential functions ϕ_C. Understanding how to optimize these two components and analyzing the properties of such Markov random fields over functional spaces is an open problem.

References

[1] L. Alvarez, F. Guichard, P.L. Lions and J.M. Morel. Axioms and fundamental equations of image processing. *Archieve for Rational Mechanic*, 123:199–257, 1993.

[2] A. Benassi, S. Jaffard, and D. Roux, Elliptic Gaussian random processes *Revista Mathematica Iberoamericana* 13:19–90, 1997.

[3] S. Chen and D. Donoho. Atomic decomposition by basis pursuit. In *SPIE International Conference on Wavelets*, San Diego, July 1995.

[4] A. Cohen, R. DeVore, P. Pertrushev, and H. Xu, Non-linear approximation and the space $BV(\mathbb{R}^2)$ *American J. of Math.*, to appear 1998.

[5] R. R. Coifman and Y. Meyer. Remarques sur l'analyse de Fourier a fenêtre. *C.R. Acad. Sci.*, pages 259–261, 1991.

[6] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Info. Theory*, 38(2):713–718, March 1992.

[7] A. Croisier, D. Esteban, and C. Galand. Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques. In *Int. Conf. on Info. Sciences and Systems*, pages 443–446, Patras, Greece, August 1976.

[8] G. Cross, and A. Jain. Markov random field texture models *IEEE Trans. on PAMI*, 5:25–39, 1983.

[9] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia, PA, 1992.

[10] I. Daubechies Wavelets and other phase space localization methods *Proc. of ICM 94*, Birkhäuser Verlag, Switzerland 1995.

[11] J. G. Daugmann. Two-dimensional spectral analysis of cortical receptive field profile. *Vision Research*, 20:847–856, 1980.

[12] R. DeVore, B. Jawerth, and V. Popov, Compression of wavelet decompositions *Amer. J. Math.*, 114:737–785, 1992.

[13] R. DeVore, Nonlinear approximation *Acta Numerica*, 51–150, Cambridge University Press, 1998.

[14] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, December 1994.

[15] D. Donoho, Abstract statistical estimation and modern harmonic analysis. *Proc. of ICM 94*, Birkhäuser Verlag, Switzerland 1995.

[16] N. Dyn and S. Rippa. Data-dependent triangulations for scattered data interpolation and finite element approximation. *Applied Num. Math.*, 12:89–105, 1993.

[17] S. Geman, and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images *IEEE Trans. on PAMI*, 6:721-741, November 1984.

[18] B. Gidas. A renormalization group approach to image processing. *IEEE Trans. on PAMI*, 11:164-180, 1989.

[19] S. Mallat. Multiresolution approximations and wavelet orthonormal bases of $\mathbf{L^2}(\mathbb{R})$. *Trans. Amer. Math. Soc.*, 315:69–87, September 1989.

[20] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, December 1993.

[21] S. Mallat, Z. Zhang, and G. Papanicolaou. Adaptive covariance estimation of locally stationary processes. *Annals of Stat.*, 26(1):1–47, 1997.

[22] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, Boston, MA, 1998.

[23] H. S. Malvar. *Signal Processing with Lapped Transforms.* Artech House, Norwood, MA, 1992.

[24] D. Marr. *Vision.* W.H. Freeman and Co., San Fransisco, 1982.

[25] Y. Meyer. *Wavelets and Operators.* Advanced Mathematics. Cambridge University Press, 1992.

[26] D. Mumford. Mathematical Theories of Shape: Do they model perception? *SPIE: Geometric Methods in Computer Vision*, 1570:2–10, 1991.

[27] J. F. Muzy, E. Bacry, and A. Arneodo. The multifractal formalism revisited with wavelets. *Int. J. of Bifurcation and chaos*, 4:245, 1994.

[28] S. Osher, and L. Rudin. Feature-oriented image enhancement using shock filters. *SIAM J. on Numerical Analysis*, 27:919–940, 1990.

[29] M. Priestley, Evolutionary spectra and non-stationary processes. *J. Roy. Stat. Soc. Ser. B*, 27:204–237, 1965.

[30] J. M. Shapiro. Embedded image coding using zero-trees of wavelet coefficients. *IEEE Trans. Signal Proc.*, 41(12):3445–3462, December 1993.

[31] G. Shapiro, and A. Tannenbaum. On invariant curve evolution and image analysis. *J. of Functional Analysis*, 119(1):79–120, 1993.

[32] E. Simoncelli., and J. Portilla. Texture characterization via second-order statistics of wavelet coefficient amplitudes. Proc. of 5^{th} IEEE Int. Conf. on Image Proc., Chicago, October 1998.

[33] M. Vetterli and J. Kovacevic. *Wavelets and Subband Coding.* Prentice-Hall, Englewood Cliffs, NJ, 1995.

[34] J. W. Woods and S. D. O'Neil. Sub-band coding of images. *IEEE Trans. Acoust., Speech, and Signal Proc.*, 34(5):1278–1288, May 1986.

[35] S. C. Zhu, and D. Mumford. Prior learning and Gibbs reaction-diffusion. *IEEE Trans. on PAMI*, 19:1236–1250, November 1997.

Stéphane Mallat
Dept. of Applied Mathematics
Ecole Polytechnique
91128 Palaiseau Cedex, France

Courant Institute
New York University
215 Mercer Street
New York, NY 10012, US