1

Introduction to the World of Sparsity

We first explore recent developments in multiresolution analysis. Essential terminology is introduced in the scope of our general overview, which includes the coverage of sparsity and sampling, best dictionaries, overcomplete representation and redundancy, compressed sensing and sparse representation, and morphological diversity.

Then we describe a range of applications of visualization, filtering, feature detection, and image grading. Applications range over Earth observation and astronomy, medicine, civil engineering and materials science, and image databases generally.

1.1 SPARSE REPRESENTATION

1.1.1 Introduction

In the last decade, sparsity has emerged as one of the leading concepts in a wide range of signal-processing applications (restoration, feature extraction, source separation, and compression, to name only a few applications). Sparsity has long been an attractive theoretical and practical signal property in many areas of applied mathematics (such as computational harmonic analysis, statistical estimation, and theoretical signal processing).

Recently, researchers spanning a wide range of viewpoints have advocated the use of overcomplete signal representations. Such representations differ from the more traditional representations because they offer a wider range of generating elements (called *atoms*). Indeed, the attractiveness of redundant signal representations relies on their ability to *economically* (or compactly) represent a large class of signals. Potentially, this wider range allows more flexibility in signal representation and adaptivity to its *morphological* content and entails more effectiveness in many signal-processing tasks (restoration, separation, compression, and estimation). Neuroscience also underlined the role of overcompleteness. Indeed, the mammalian visual system has been shown to be likely in need of overcomplete representation (Field 1999; Hyvärinen and Hoyer 2001; Olshausen and Field 1996a; Simoncelli and

Olshausen 2001). In that setting, overcomplete *sparse coding* may lead to more effective (sparser) codes.

The interest in sparsity has arisen owing to the new sampling theory, *compressed sensing* (also called *compressive sensing* or *compressive sampling*), which provides an alternative to the well-known Shannon sampling theory (Candès and Tao 2006; Donoho 2006a; Candès et al. 2006b). Compressed sensing uses the prior knowledge that signals are sparse, whereas Shannon theory was designed for frequency band–limited signals. By establishing a direct link between sampling and sparsity, compressed sensing has had a huge impact in many scientific fields such as coding and information theory, signal and image acquisition and processing, medical imaging, and geophysical and astronomical data analysis. Compressed sensing acts today as wavelets did two decades ago, linking researchers from different fields. Further contributing to the success of compressed sensing is that some traditional inverse problems, such as tomographic image reconstruction, can be understood as compressed sensing problems (Candès et al. 2006b; Lustig et al. 2007). Such ill-posed problems need to be regularized, and many different approaches to regularization have been proposed in the last 30 years (Tikhonov regularization, Markov random fields, total variation, wavelets, etc.). But compressed sensing gives strong theoretical support for methods that seek a sparse solution because such a solution may be (under certain conditions) the exact one. Similar results have not been demonstrated with any other regularization method. These reasons explain why, just a few years after seminal compressed sensing papers were published, many hundreds of papers have already appeared in this field (see, e.g., the compressed sensing resources Web site http://www.compressedsensing.com).

By emphasizing so rigorously the importance of sparsity, compressed sensing has also cast light on all work related to sparse data representation (wavelet transform, curvelet transform, etc.). Indeed, a signal is generally not sparse in direct space (i.e., pixel space), but it can be very sparse after being decomposed on a specific set of functions.

1.1.2 What Is Sparsity?

1.1.2.1 Strictly Sparse Signals/Images

A signal x, considered as a vector in a finite-dimensional subspace of $\mathbb{R}^N$, $x = [x[1], \ldots, x[N]]$, is strictly or exactly sparse if most of its entries are equal to zero, that is, if its support $\Lambda(x) = \{1 \leq i \leq N \mid x[i] \neq 0\}$ is of cardinality $k \ll N$. A k-sparse signal is a signal for which exactly k samples have a nonzero value.

If a signal is not sparse, it may be *sparsified* in an appropriate transform domain. For instance, if x is a sine, it is clearly not sparse, but its Fourier transform is extremely sparse (actually, 1-sparse). Another example is a piecewise constant image away from edges of finite length that has a sparse gradient.

More generally, we can model a signal x as the linear combination of T elementary waveforms, also called *signal atoms*, such that

$$x = \mathbf{\Phi}\alpha = \sum_{i=1}^{T} \alpha[i]\varphi_i, \tag{1.1}$$

where $\alpha[i]$ are called the representation coefficients of x in the *dictionary* $\mathbf{\Phi} = [\varphi_1, \ldots, \varphi_T]$ (the $N \times T$ matrix whose columns are the atoms φ_i, in general normalized to a unit ℓ_2 norm, i.e., $\forall i \in \{1, \ldots, T\}$, $\|\varphi_i\|^2 = \sum_{n=1}^{N} |\varphi_i[n]|^2 = 1$).

Signals or images x that are sparse in $\mathbf{\Phi}$ are those that can be written *exactly* as a superposition of a small fraction of the atoms in the family $(\varphi_i)_i$.

1.1.2.2 Compressible Signals/Images

Signals and images of practical interest are not, in general, strictly sparse. Instead, they may be *compressible* or *weakly sparse* in the sense that the sorted magnitudes $|\alpha_{(i)}|$ of the representation coefficients $\alpha = \mathbf{\Phi}^{\mathrm{T}} x$ decay quickly according to the power law

$$\left|\alpha_{(i)}\right| \leq C i^{-1/s}, \; i = 1, \ldots, T,$$

and the nonlinear approximation error of x from its k-largest coefficients (denoted x_k) decays as

$$\|x - x_k\| \leq C(2/s - 1)^{-1/2} k^{1/2-1/s}, \; s < 2.$$

In other words, one can neglect all but perhaps a small fraction of the coefficients without much loss. Thus x can be well approximated as k-sparse.

Smooth signals and piecewise smooth signals exhibit this property in the wavelet domain (Mallat 2008). Owing to recent advances in harmonic analysis, many redundant systems, such as the undecimated wavelet transform, curvelet, and contourlet, have been shown to be very effective in sparsely representing images. As popular examples, one may think of wavelets for smooth images with isotropic singularities (Mallat 1989, 2008), bandlets (Le Pennec and Mallat 2005; Peyré and Mallat 2007; Mallat and Peyré 2008), grouplets (Mallat 2009) or curvelets for representing piecewise smooth C^2 images away from C^2 contours (Candès and Donoho 2001; Candès et al. 2006a), wave atoms or local discrete cosine transforms to represent locally oscillating textures (Demanet and Ying 2007; Mallat 2008), and so on. Compressibility of signals and images forms the foundation of transform coding, which is the backbone of popular compression standards in audio (MP3, AAC), imaging (JPEG, JPEG-2000), and video (MPEG).

Figure 1.1 shows the histogram of an image in both the original domain (i.e., $\mathbf{\Phi} = \mathbf{I}$, where $\mathbf{I}$ is the identity operator, hence $\alpha = x$) and the curvelet domain. We can see immediately that these two histograms are very different. The second histogram presents a typical sparse behavior (unimodal, sharply peaked with heavy tails), where most of the coefficients are close to zero and few are in the tail of the distribution.

Throughout the book, with a slight abuse of terminology, we may call signals and images sparse, both those that are strictly sparse and those that are compressible.

1.1.3 Sparsity Terminology

1.1.3.1 Atom

As explained in the previous section, an atom is an elementary signal-representing template. Examples include sinusoids, monomials, wavelets, and Gaussians. Using a

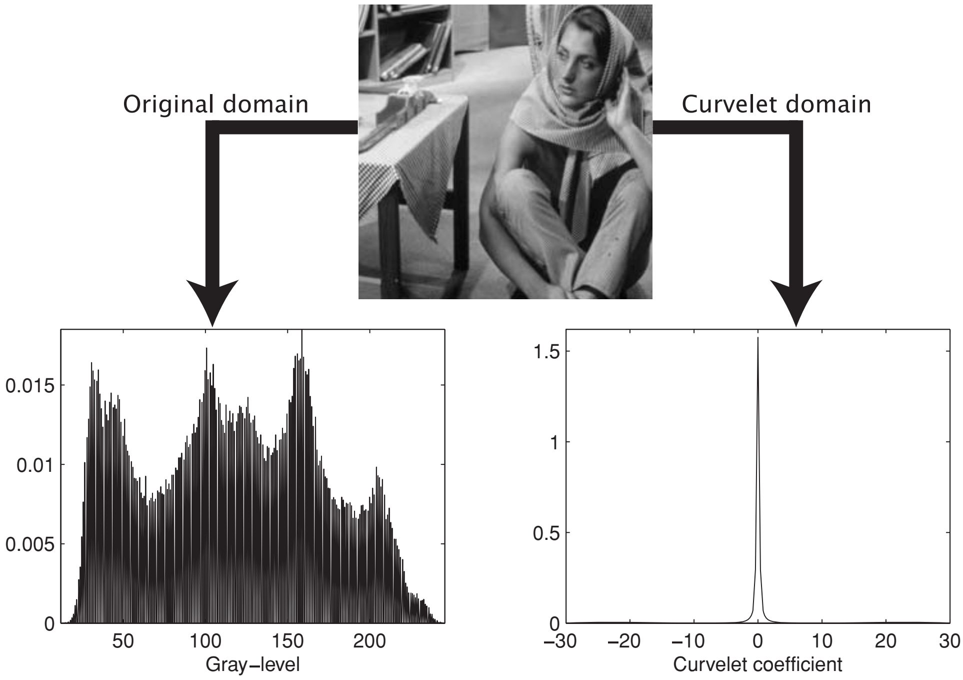

Figure 1.1. Histogram of an image in (left) the original (pixel) domain and (right) the curvelet domain.

collection of atoms as building blocks, one can construct more complex waveforms by linear superposition.

1.1.3.2 Dictionary

A dictionary $\mathbf{\Phi}$ is an indexed collection of atoms $(\varphi_\gamma)_{\gamma\in\Gamma}$, where Γ is a countable set; that is, its cardinality $|\Gamma| = T$. The interpretation of the index γ depends on the dictionary: frequency for the Fourier dictionary (i.e., sinusoids), position for the Dirac dictionary (also known as *standard unit vector basis* or *Kronecker basis*), position scale for the wavelet dictionary, translation-duration-frequency for cosine packets, and position-scale-orientation for the curvelet dictionary in two dimensions. In discrete-time, finite-length signal processing, a dictionary is viewed as an $N \times T$ matrix whose columns are the atoms, and the atoms are considered as column vectors. When the dictionary has more columns than rows, $T > N$, it is called *overcomplete* or *redundant*. The overcomplete case is the setting in which $x = \mathbf{\Phi}\alpha$ amounts to an underdetermined system of linear equations.

1.1.3.3 Analysis and Synthesis

Given a dictionary, one has to distinguish between analysis and synthesis operations. *Analysis* is the operation that associates with each signal x a vector of coefficients α attached to an atom: $\alpha = \mathbf{\Phi}^{\mathrm{T}} x$.[1] *Synthesis* is the operation of reconstructing x by

[1] The dictionary is supposed to be real. For a complex dictionary, $\mathbf{\Phi}^{\mathrm{T}}$ is to be replaced by the conjugate transpose (adjoint) $\mathbf{\Phi}^*$.

superposing atoms: $x = \mathbf{\Phi}\alpha$. Analysis and synthesis are different linear operations. In the overcomplete case, $\mathbf{\Phi}$ is not invertible, and the reconstruction is not unique (see also Section 8.2 for further details).

1.1.4 Best Dictionary

Obviously, the best dictionary is the one that leads to the sparsest representation. Hence we could imagine having a huge dictionary (i.e., $T \gg N$), but we would be faced with a prohibitive computation time cost for calculating the α coefficients. Therefore there is a trade-off between the complexity of our analysis (i.e., the size of the dictionary) and computation time. Some specific dictionaries have the advantage of having fast operators and are very good candidates for analyzing the data. The Fourier dictionary is certainly the most well known, but many others have been proposed in the literature such as wavelets (Mallat 2008), ridgelets (Candès and Donoho 1999), curvelets (Candès and Donoho 2002; Candès et al. 2006a; Starck et al. 2002), bandlets (Le Pennec and Mallat 2005), and contourlets (Do and Vetterli 2005), to name but a few candidates. We will present some of these in the chapters to follow and show how to use them for many inverse problems such as denoising or deconvolution.

1.2 FROM FOURIER TO WAVELETS

The Fourier transform is well suited only to the study of stationary signals, in which all frequencies have an infinite coherence time, or, otherwise expressed, the signal's statistical properties do not change over time. Fourier analysis is based on global information that is not adequate for the study of compact or local patterns.

As is well known, Fourier analysis uses basis functions consisting of sine and cosine functions. Their frequency content is time-independent. Hence the description of the signal provided by Fourier analysis is purely in the frequency domain. Music or the voice, however, imparts information in both the time and the frequency domains. The windowed Fourier transform and the wavelet transform aim at an analysis of both time and frequency. A short, informal introduction to these different methods can be found in the work of Bentley and McDonnell (1994), and further material is covered by Chui (1992), Cohen (2003), and Mallat (2008).

For nonstationary analysis, a windowed Fourier transform (short-time Fourier transform, STFT) can be used. Gabor (1946) introduced a local Fourier analysis, taking into account a sliding Gaussian window. Such approaches provide tools for investigating time and frequency. Stationarity is assumed within the window. The smaller the window size, the better is the time resolution; however, the smaller the window size, also, the more the number of discrete frequencies that can be represented in the frequency domain will be reduced, and therefore the more weakened will be the discrimination potential among frequencies. The choice of window thus leads to an uncertainty trade-off.

The STFT transform, for a continuous-time signal $s(t)$, a window g around time point τ, and frequency ω, is

$$\mathrm{STFT}(\tau, \omega) = \int_{-\infty}^{+\infty} s(t) g(t - \tau) e^{-j\omega t} dt. \tag{1.2}$$

Considering

$$k_{\tau,\omega}(t) = g(t-\tau)e^{-j\omega t} \tag{1.3}$$

as a new basis, and rewriting this with window size a, inversely proportional to the frequency ω, and with positional parameter b replacing τ, as

$$k_{b,a}(t) = \frac{1}{\sqrt{a}}\psi^*\left(\frac{t-b}{a}\right), \tag{1.4}$$

yields the continuous wavelet transform (CWT), where ψ^* is the complex conjugate of ψ. In the STFT, the basis functions are windowed sinusoids, whereas in the CWT, they are scaled versions of a so-called mother function ψ.

In the early 1980s, the wavelet transform was studied theoretically in geophysics and mathematics by Morlet, Grossman, and Meyer. In the late 1980s, links with digital signal processing were pursued by Daubechies and Mallat, thereby putting wavelets firmly into the application domain.

A wavelet mother function can take many forms, subject to some admissibility constraints. The best choice of mother function for a particular application is not given a priori.

From the basic wavelet formulation, one can distinguish (Mallat 2008) between (1) the CWT, described earlier, and (2) the discrete wavelet transform, which discretizes the continuous transform but does not, in general, have an exact analytical reconstruction formula; and within discrete transforms, distinction can be made between (3) redundant versus nonredundant (e.g., pyramidal) transforms and (4) orthonormal versus other bases of wavelets. The wavelet transform provides a decomposition of the original data, allowing operations to be performed on the wavelet coefficients, and then the data are reconstituted.

1.3 FROM WAVELETS TO OVERCOMPLETE REPRESENTATIONS

1.3.1 The Blessing of Overcomplete Representations

As discussed earlier, different wavelet transform algorithms correspond to different wavelet dictionaries. When the dictionary is overcomplete, $T > N$, the number of coefficients is larger than the number of signal samples. Because of the redundancy, there is no unique way to reconstruct x from the coefficients α. For compression applications, we obviously prefer to avoid this redundancy, which would require us to encode a greater number of coefficients. But for other applications, such as image restoration, it will be shown that redundant wavelet transforms outperform orthogonal wavelets. Redundancy here is welcome, and as long as we have fast analysis and synthesis algorithms, we prefer to analyze the data with overcomplete representations.

If wavelets are well designed for representing isotropic features, ridgelets or curvelets lead to sparser representation for anisotropic structures. Both ridgelet and curvelet dictionaries are overcomplete. Hence, as we will see throughout this book, we can use different transforms, overcomplete or otherwise, to represent our data:

- the Fourier transform for stationary signals
- the windowed Fourier transform (or a local cosine transform) for locally stationary signals

- the isotropic undecimated wavelet transform for isotropic features; this wavelet transform is well adapted to the detection of isotropic features such as the clumpy structures to which we referred earlier
- the anisotropic biorthogonal wavelet transform; we expect the biorthogonal wavelet transform to be optimal for detecting mildly anisotropic features
- the ridgelet transform, developed to process images that include ridge elements and so to provide a good representation of perfectly straight edges
- the curvelet transform to approximate curved singularities with few coefficients and then provide a good representation of curvilinear structures

Therefore, when we choose one transform rather than another, we introduce, in fact, a prior on what is in the data. The analysis is optimal when the most appropriate decomposition to our data is chosen.

1.3.2 Toward Morphological Diversity

The morphological diversity concept was introduced to model a signal as a sum of a mixture, each component of the mixture being sparse in a given dictionary (Starck et al. 2004b; Elad et al. 2005; Starck et al. 2005a). The idea is that a single transformation may not always represent an image well, especially if the image contains structures with different spatial morphologies. For instance, if an image is composed of edges and texture, or alignments and Gaussians, we will show how we can analyze our data with a large dictionary and still have fast decomposition. We choose the dictionary as a combination of several subdictionaries, and each subdictionary has a fast transformation/reconstruction. Chapter 8 will describe the morphological diversity concept in full detail.

1.3.3 Compressed Sensing: The Link between Sparsity and Sampling

Compressed sensing is based on a nonlinear sampling theorem, showing that an N-sample signal x with exactly k nonzero components can be recovered perfectly from order $k \log N$ incoherent measurements. Therefore the number of measurements required for exact reconstruction is much smaller than the number of signal samples and is directly related to the sparsity level of x. In addition to the sparsity of the signal, compressed sensing requires that the measurements be incoherent. Incoherent measurements mean that the information contained in the signal is spread out in the domain in which it is acquired, just as a Dirac in the time domain is spread out in the frequency domain. Compressed sensing is a very active domain of research and applications. We will describe it in more detail in Chapter 11.

1.3.4 Applications of Sparse Representations

We briefly motivate the varied applications that will be discussed in the following chapters.

The human visual interpretation system does a good job at taking scales of a phenomenon or scene into account simultaneously. A wavelet or other multiscale transform may help us with visualizing image or other data. A decomposition into different resolution scales may open up, or lay bare, faint phenomena that are part of what is under investigation.

In capturing a view of multilayered reality in an image, we are also picking up noise at different levels. Therefore, in trying to specify what is noise in an image, we may find it effective to look for noise in a range of resolution levels. Such a strategy has proven quite successful in practice.

Noise, of course, is pivotal for the effective operation, or even selection, of analysis methods. Image deblurring, or deconvolution or restoration, would be trivially solved were it not for the difficulties posed by noise. Image compression would also be easy were it not for the presence of what is, by definition, noncompressible, that is, noise.

In all these areas, efficiency and effectiveness (or quality of the result) are important. Various application fields come immediately to mind: astronomy, remote sensing, medicine, industrial vision, and so on.

All told, there are many and varied applications for the methods described in this book. On the basis of the description of many applications, we aim to arm the reader for tackling other, similar applications. Clearly this objective holds, too, for tackling new and challenging applications.

1.4 NOVEL APPLICATIONS OF THE WAVELET AND CURVELET TRANSFORMS

To provide an overview of the potential of the methods to be discussed in later chapters, the remainder of the present chapter is an appetizer.

1.4.1 Edge Detection from Earth Observation Images

Our first application (Figs. 1.2 and 1.3) in this section relates to Earth observation. The European Remote Sensing Synthetic Aperture Radar (SAR) image of the Gulf of Oman contains several spiral features. The Sea-viewing Wide Field-of-view Sensor (SeaWiFS) image is coincident with this SAR image.

There is some nice correspondence between the two images. The spirals are visible in the SAR image as a result of biological matter on the surface, which forms into slicks when there are circulatory patterns set up due to eddies. The slicks show up against the normal sea surface background due to reduction in backscatter from the surface. The biological content of the slicks causes the sea surface to become less rough, hence providing less surface area to reflect back emitted radar from the SAR sensor. The benefit of SAR is its all-weather capability; that is, even when SeaWiFS is cloud covered, SAR will still give signals from the sea surface. Returns from the sea surface, however, are affected by wind speed over the surface, and this explains the large black patches. The patches result from a drop in the wind at these locations, leading to reduced roughness of the surface.

Motivation for us was to know how successful SeaWiFS feature (spiral) detection routines would be in highlighting the spirals in this type of image, bearing in mind the other features and artifacts. Multiresolution transforms could be employed in this context, as a form of reducing the background signal to highlight the spirals.

Figure 1.2 shows an original SAR image, followed by a superimposition of resolution-scale information on the original image. The right-hand image is given

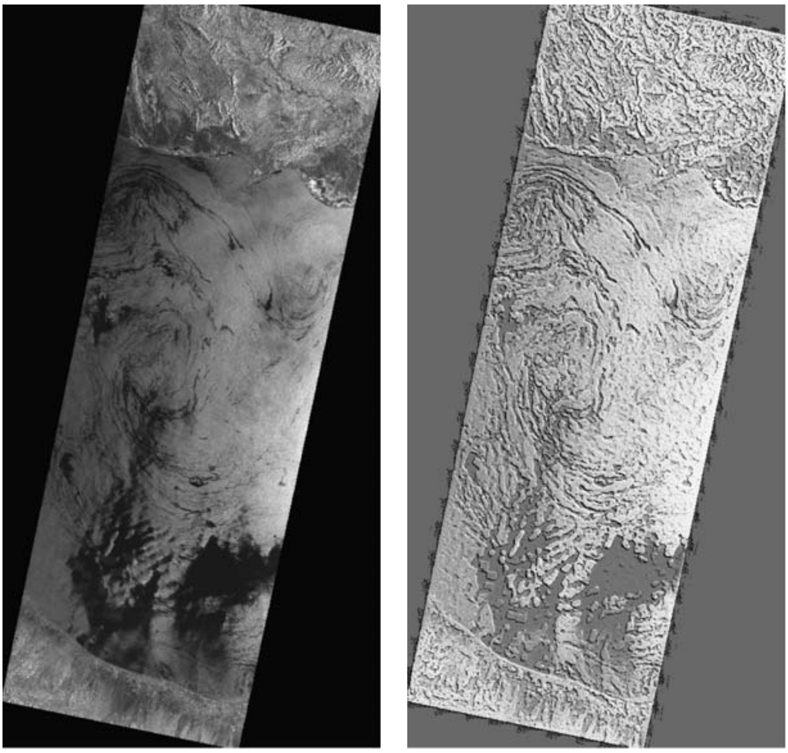

Figure 1.2. (left) SAR image of Gulf of Oman region and (right) resolution-scale information superimposed.

by the original image plus 100 times the resolution scale 3 image plus 20 times the resolution scale 4 image.

In Fig. 1.3, the corresponding SeaWiFS image is shown. The weighting used here for the right-hand image is the original image times 0.0005 plus the resolution scale 5 image.

In both cases, the analysis was based on the starlet transform, to be discussed in Section 3.5.

1.4.2 Wavelet Visualization of a Comet

Figure 1.4 shows periodic comet P/Swift-Tuttle observed with the 1.2 m telescope at Calar Alto Observatory in Spain in October and November 1992. Irregularity of the nucleus is indicative of the presence of jets in the coma (see resolution scales 4 and 5 of the wavelet transform, where these jets can be clearly seen). The starlet transform, or B_3 spline à trous wavelet transform, was used.

1.4.3 Filtering an Echocardiograph Image

Figure 1.5 shows an echocardiograph image. We see in this noninvasive ultrasound image a cross section of the heart, showing blood pools and tissue. The heavy speckle, typical of ultrasound images, makes interpretation difficult. For the filtered

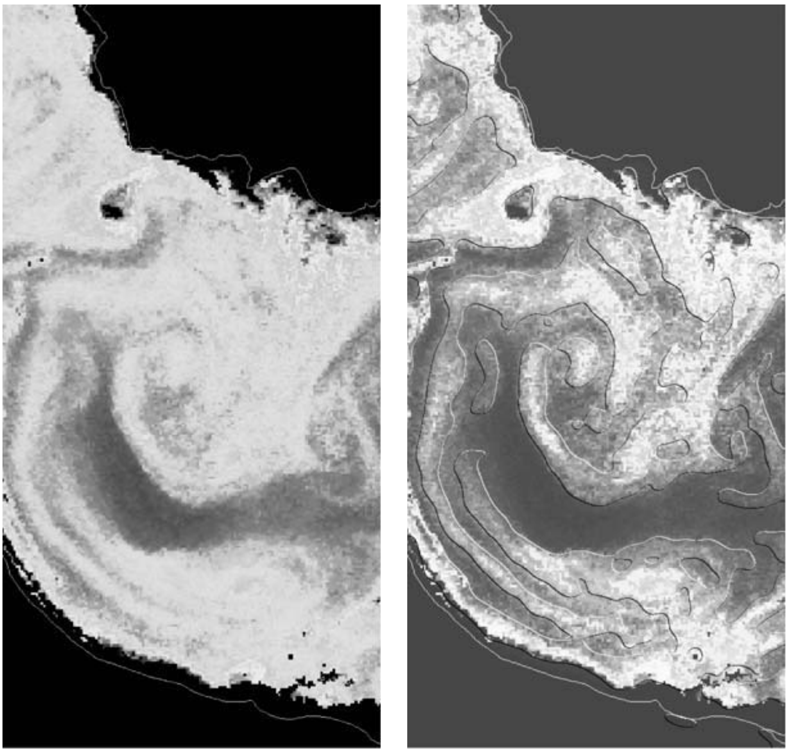

Figure 1.3. (left) SeaWiFS image of the Gulf of Oman region and (right) resolution-scale information superimposed. (*See color plates.*)

image in Fig. 1.5, wavelet scales 4 and 5 were retained, and here we see the sum of these two images. Again, the starlet transform was used.

In Fig. 1.6, a superimposition of the original image is shown with resolution-level information. This is done to show edge or boundary information and simultaneously to relate this to the original image values for validation purposes. In Fig. 1.6, the left image is the original image plus 500 times the second derivative of the fourth resolution–scale image resulting from the starlet transform algorithm. The right image in Fig. 1.6 is the original image plus 50,000 times the logarithm of the second derivative of the fourth resolution scale.

1.4.4 Curvelet Moments for Image Grading and Retrieval

1.4.4.1 Image Grading as a Content-Based Image Retrieval Problem

Physical sieves are used to classify crushed stone based on size and granularity. Then mixes of aggregate are used. We directly address the problem of classifying the mixtures, and we assess the algorithmic potential of this approach, which has considerable industrial importance.

The success of content-based image finding and retrieval is most marked when the user's requirements are very specific. An example of a specific application domain is the grading of engineering materials. Civil engineering construction *aggregate* sizing is carried out in the industrial context by passing the material over