Curvelets – A Surprisingly Effective Nonadaptive Representation For Objects with Edges

Emmanuel J. Candès and David L. Donoho

Abstract. It is widely believed that to efficiently represent an otherwise smooth object with discontinuities along edges, one must use an adaptive representation that in some sense 'tracks' the shape of the discontinuity set. This folk-belief — some would say folk-theorem — is incorrect. At the very least, the possible quantitative advantage of such adaptation is vastly smaller than commonly believed. We have recently constructed a tight frame of curvelets which provides stable, efficient, and near-optimal representation of otherwise smooth objects having discontinuities along smooth curves. By applying naive thresholding to the curvelet transform of such an object, one can form m-term approximations with rate of L^2 approximation rivaling the rate obtainable by complex adaptive schemes which attempt to 'track' the discontinuity set. In this article we explain the basic issues of efficient m-term approximation, the construction of efficient adaptive representation, the construction of the curvelet frame, and a crude analysis of the performance of curvelet schemes.

§1. Introduction

In many important imaging applications, images exhibit edges – discontinuities across curves. In traditional photographic imaging, for example, this occurs whenever one object occludes another, causing the luminance to undergo step discontinuities at boundaries. In biological imagery, this occurs whenever two different organs or tissue structures meet.

In image *synthesis* applications, such as CAD, there is no problem in dealing with such discontinuities, because one knows where they are and builds the discontinuities into the representation by specially adapting the representation — for example, inserting free knots, or adaptive refinement rules.

In image *analysis* applications, the situation is different. When working with real rather than synthetic data, one of course doesn't 'know' where these edges are; one only has a digitized pixel array, with potential imperfections caused by noise, by blurring, and of course by the unnatural pixelization of the underlying continuous scene. Hence the typical image analyst only

Saint-Malo Proceedings
XXX, XXX, and Larry L. Schumaker (eds.), pp. 1–10.

ISBN 1-xxxxx-xxx-x.

has recourse to representations which don't 'know' about the existence and geometry of the discontinuities in the image.

The success of discontinuity-adapting methods in CAD and related image synthesis fields creates a temptation for an image analyst – a temptation to spend a great deal of time and effort importing such ideas into image analysis. Almost everyone we know has yielded to this temptation in some form, which creates a possibility for surprise.

Oracles and Ideally-Adapted Representation

One could imagine an ideally-privileged image analyst who has recourse to an *oracle* able to reveal the positions of all the discontinuities underlying the image formation. It seems natural that this ideally-privileged analyst could do far better than the normally-endowed analyst who knows nothing about the position of the discontinuities in the image.

To elaborate this distinction, we introduce terminology borrowed from fluid dynamics, where 'edges' arise in the form of fronts or shock fronts.

A *Lagrangian representation* is constructed using full knowledge of the intrinsic structure of the object and adapting perfectly to that structure.

- In fluid dynamics this means that the fluid flow pattern is known, and one constructs a coordinate system which 'flows along with the particles', with coordinates mimicking the shape of the flow streamlines.
- In image representation this could mean that the edge curves are known, and one constructs an image representation adapted to the structure of the edge curves. For example, one might construct a basis with discontinuities exactly where the underlying object has discontinuities.

An *Eulerian representation* is fixed, constructed once and for all. It is nonadaptive – having nothing to do with the known or hypothesized details of the underlying object.

- In fluid dynamics, this would mean a usual euclidean coordinate system, one that does not depend in any way on the fluid motion.
- In image representation, this could mean that the representation is some fixed coordinate representation, such as wavelets or sinusoids, which does not change depending on the positions of edges in the image.

It is quite natural to suppose that the Lagrangian perspective, when it is available, is much more powerful that the Eulerian one. Having the privilege of 'inside information' about the position of important geometric characteristics of the solution seems *a priori* rather valuable. In fact, this position has rather a large following. Much recent work in computational harmonic analysis (CHA) attempts to find bases which are optimally adapted to the specific object in question [7,10,11]; in this sense much of the ongoing work in CHA is based on the presumption that the Lagrangian viewpoint is best.

In the setting of edges in images, there has, in fact, been considerable interest in the problem of developing representations which are adapted to the structure of discontinuities in the object being studied. The (equivalent)

concepts of probing and minimum entropy segmentation are old examples of this: wavelet systems which are specifically constructed to allow discontinuities in the basis elements at specific locations [8,9]. More recently, we are aware of much informal unpublished or preliminary work attempting to build $2D$ edge-adapted schemes; we give two examples.

- *Adaptive triangulation* aims to represent a smooth function by partitioning the plane into a sequence of triangular meshes, refining the meshes at one stage to create finer meshes at the next stage. One represents the underlying object using piecewise linear functions supported on individual triangles. It is easy to see how, in an *image synthesis* setting, one can in principle develop a triangulation where the triangles are arranged to track a discontinuity very faithfully, with the bulk of refinement steps allocated to refinements near the discontinuity, and one obtains very effective representation of the object. It is not easy to see how to do this in an *image analysis* setting, but one can easily be persuaded that the development of adaptive triangulation schemes for noisy, blurred data is an important and interesting project.
- In an *adaptively warped wavelet representation*, one deforms the underlying image so that the object being analyzed has all its discontinuities aligned purely horizontal or vertical. Then one analyzes the warped object in a basis of tensor-product wavelets where elements take the form $\psi_{j,k}(x_1) \cdot \psi_{j',k'}(x_2)$. This is very effective for objects which are smooth apart from purely horizontal and purely vertical discontinuities. Hence, the warping deforms the singularities to render the the tensor product scheme very effective. It is again not easy to see how adaptive warping could work in an *image analysis* setting, but one is easily persuaded that development of adaptively warped representations for noisy, blurred data is an important and interesting project.

Activity to build such adaptive representations is based on an article of faith: namely, that *Eulerian approaches are inferior, that oracle-driven Lagrangian approaches are ideal*, and that one should, in an image analysis setting, mimic Lagrangian approaches, attempting empirically to estimate from noisy, blurred data the information that an oracle would supply, and build an adaptive representation based on that information.

Quantifying Rates of Approximation

In order to get away from articles of faith, we now quantify performance, using an asymptotic viewpoint.

Suppose we have an object supported in $[0,1]^2$ which has a discontinuity across a nice curve Γ, and which is otherwise smooth. Then using a standard Fourier representation, and approximating with $\tilde{f}_m^F$ built from the best m nonzero Fourier terms, we have

$$\|f - \tilde{f}_m^F\|_2^2 \asymp m^{-1/2}, \qquad m \to \infty. \tag{1}$$

This rather slow rate of approximation is improved upon by wavelets. The approximant $\tilde{f}_m^W$ built from the best m nonzero wavelet terms satisfies

$$\|f - \tilde{f}_m^W\|_2^2 \asymp m^{-1}, \qquad m \to \infty. \tag{2}$$

This is better than the rate of Fourier approximation, and, until now, is the best published rate for a fixed non-adaptive method (i.e. best published result for an 'Eulerian viewpoint').

On the other hand, we will discuss below a method which is adapted to the object at hand, and which achieves a much better approximation rate than previously known 'nonadaptive' or 'Eulerian' approaches. This adaptive method selects terms from an overcomplete dictionary and is able to achieve

$$\|f - \tilde{f}_m^A\|_2^2 \asymp m^{-2}, \qquad m \to \infty. \tag{3}$$

Roughly speaking, the terms in this dictionary amount to triangular wedges, ideally fitted to approximate the shape of the discontinuity.

Owing to the apparent trend indicated by (1)-(3) and the prevalence of the puritanical belief that 'you can't get something for nothing', one might suppose that inevitably would follow the

Folk-Conjecture/[Folk-Theorem]. *The result (3) for adaptive representations far exceeds the rate of m-term approximation achievable by fixed nonadaptive representations.*

This conjecture appeals to a number of widespread beliefs:

- the belief that adaptation is very powerful,
- the belief that the way to represent discontinuities in image analysis is to mimic the approach in image synthesis
- the belief that wavelets give the best fixed nonadaptive representation.

In private discussions with many respected researchers we have many times heard expressed views equivalent to the purported Folk-Theorem.

The Surprise

It turns out that performance almost equivalent to (3) can be achieved by a *non*adaptive scheme. In other words, the Folk-Theorem is effectively false.

There is a tight frame, fixed once and for all nonadaptively, which we call a frame of curvelets, which competes surprisingly well with the ideal adaptive rate (3). A very simple m-term approximation – summing the m biggest terms in the curvelet frame expansion – can achieve

$$\|f - \tilde{f}_m^C\|_2^2 \le C \cdot m^{-2} (\log m)^3, \qquad m \to \infty, \tag{4}$$

which is nearly as good as (3) as regards asymptotic order.

In short, *in a problem of considerable applied relevance, where one would have thought that adaptive representation was essentially more powerful than fixed nonadaptive representation, it turns out that a new fixed nonadaptive representation is essentially as good as adaptive representation, from the point of view of asymptotic m-term approximation errors.* As one might expect, the new nonadaptive representation has several very subtle and distinctive features.

Contents

In this article, we would like to give the reader an idea of why (3) represents the ideal behavior of an adaptive representation, of how the curvelet frame is constructed, and of the key elements responsible for (4). We will also attempt to indicate why curvelets perform for singularities along curves the task that wavelets perform for singularities at points.

§2. A Precedent: Wavelets and Point Singularities

We mention an important precedent – a case where a nonadaptive scheme is roughly competitive with an ideal adaptive scheme.

Suppose we have a piecewise polynomial function f on the interval $[0, 1]$, with jump discontinuities at several points.

An obvious adaptive representation is to fit a piecewise polynomial with breakpoints at the discontinuities. If there are P pieces and each polynomial is of degree $\leq D$, then we need only keep $P \cdot (D + 1)$ coefficients and $P - 1$ breakpoints to exactly represent this function. Common sense tells us that this is the natural, and even, the ideal representation for such a function.

To build this representation, we need to know locations of the discontinuities. If the measurements are noisy or blurred, and if we don't have recourse to an oracle, then we can't necessarily build this representation.

A less obvious but much more robust representation is to take a nice wavelet transform of the object, and keep the few resulting nonzero wavelet coefficients. If we have an N-point digital signal $f(i/N)$, $1 \leq i \leq N$, and we use Daubechies wavelets of compact support, then there are no more than $C \cdot \log_2(N) \cdot P \cdot (D + 1)$ nonzero wavelet coefficients for the digital signal.

In short, the nonadaptive representation needs only to keep a factor $C \log_2(N)$ more data to give an equally faithful representation.

We claim that this phenomenon is at least partially responsible for the widespread success of wavelet methods in data compression settings. One can build a single fast transform and deal with a wide range of different f, with different discontinuity sets, without recourse to an oracle.

In particular, since one almost never has access to an oracle, the natural first impulse of one committed to the adaptive viewpoint would be to 'estimate' the break points – i.e. to perform some sort of edge detection. Unfortunately this is problematic when one is dealing with noisy blurred data. Edge detection is a whole topic in itself which has thousands of proposed solutions and (evidently, as one can see from the continuing rate of publication in this area) no convincing solution.

In using wavelets, one does not need edge detectors or any other problematic schemes, one simply extracts the big coefficients from the transform domain, and records their values and positions in an organized fashion.

We can lend a useful perspective to this phenomenon by noticing that the discontinuities in the underlying f are *point singularities*, and we are saying that wavelets need in some sense at most $\log(n)$ coefficients to represent a point singularity out to scale $1/n$.

It turns out that even in higher dimensions wavelets have a near-ideal ability to represent objects with point singularities.

The two-dimensional object $f_\beta(x_1, x_2) = 1/((x_1 - 1/2)^2 + (x_2 - 1/2)^2)^\beta$ has, for $\beta < 1/2$, a square-integrable singularity at the point $(1/2, 1/2)$ and is otherwise smooth. At each level of the $2D$ wavelet pyramid, there are effectively only a few wavelets which 'feel' the point singularity, other coefficients being effectively negligible. In approximation out to scale $1/n$, only about $O(\log(n))$ coefficients are required.

Another approach to understanding the representation of singularities, which is not limited by scale, is to consider rates of decay of the countable coefficient sequence. Analysis of wavelet coefficients of f_β shows that for any desired rate ρ, the N-th largest coefficient can be bounded by $C_\rho N^{-\rho}$ for all N. In short, the wavelet coefficients of such an object are very sparse.

Thus we have a slogan: *wavelets perform very well for objects with point singularities in dimensions* 1 *and* 2.

§3. Failure of Wavelets on Edges

We now briefly sketch why wavelets, which worked surprisingly well in representing point discontinuities in dimension 1, are less successful dealing with 'edge' discontinuities in dimension 2.

Suppose we have an object f on the square $[0, 1]^2$ and that f is smooth away from a discontinuity along a C^2 curve Γ. Let's look at the number of substantial wavelet coefficients.

A grid of squares of side 2^{-j} by 2^{-j} has order 2^j squares intersecting Γ. At level j of the two-dimensional wavelet pyramid, each wavelet is localized near a corresponding square of side 2^{-j} by 2^{-j}. There are therefore $O(2^j)$ wavelets which 'feel' the discontinuity along Γ. Such a wavelet coefficient is controlled by

$$|\langle f, \psi_{j,k_1,k_2}\rangle| \le \|f\|_\infty \cdot \|\psi_{j,k_1,k_2}\|_1 \le C \cdot 2^{-j};$$

and in effect no better control is available, since the object f is not smooth within the support of ψ_{j,k_1,k_2} [14]. Therefore there are about 2^j coefficients of size about 2^{-j}. In short, the N-th largest wavelet coefficient is of size about $1/N$. The result (2) follows.

We can summarize this by saying that in dimension 2, discontinuities across edges are spatially distributed; because of this they can interact rather extensively with many terms in the wavelet expansion, and so the wavelet representation is not sparse.

In short, wavelets do well for point singularities, and not for singularities along curves. The success of wavelets in dimension 1 derived from the fact that all singularities in dimension 1 are point singularities, so wavelets have a certain universality there. In higher dimensions there are more types of singularities, and wavelets lose their universality.

For balance, we need to say that wavelets *do* outperform classical methods. If we used sinusoids to represent an object of the above type, then we

have the result (1), which is far worse than that provided by wavelets. For completeness, we sketch the argument. Suppose we use for 'sinusoids' the complex exponentials on $[-\pi,\pi]^2$, and that the object f tends smoothly to zero at the boundary of the square $[0,1]^2$, so that we may naturally extend it to a function living on $[-\pi,\pi]^2$. Now typically the Fourier coefficients of an otherwise smooth object with a discontinuity along a curve decay with wavenumber as $|k|^{-3/2}$ (the very well-known example is f = indicator of a disk, which has a Fourier transform described by Bessel functions). Thus there are about R^2 coefficients of size $\geq c \cdot R^{-3/2}$, meaning that the N-th largest is of size $\geq c \cdot N^{-3/4}$, from which (1) follows.

In short: neither wavelets nor sinusoids really sparsify two-dimensional objects with edges (although wavelets are better than sinusoids).

§4. Ideal Representation of Objects with Edges

We now consider the optimality result (3), which is really two assertions. On the one hand, no reasonable scheme can do better than this rate. On the other hand, a certain adaptive scheme, with intimate connections to adaptive triangulation, which achieves it. For more extensive discussion see [10,11,13].

In talking about adaptive representations, we need to define terms carefully, for the following reason. For any f, there is always an adaptive representation of f that does very well: namely the orthobasis $\Psi = \{\psi_0, \psi_1, \ldots\}$ with first element $\psi_0 = f/\|f\|_2$! This is, in a certain conception, an 'ideal representation' where each object requires only one nonzero coefficient. In a certain sense it is a useless one, since all information about f has been hidden in the definition of representation, so actually we haven't learned anything. Most of our work in this section is in setting up a notion of adaptation that will free us from fear of being trapped at this level of triviality.

Dictionaries of Atoms

Suppose we are interested in approximating a function in $L^2(T)$, and we have a countable collection $\mathcal{D} = \{\phi\}$ of atoms in $L^2(T)$; this could be a basis, a frame, a finite concatenation of bases or frames, or something even less structured.

We consider the problem of m-term approximation from this dictionary, where we are allowed to select m terms $\phi_1, \ldots, \phi_m$ from $\mathcal{D}$ and we approximate f from the L^2-closest member of the subspace they span:

$$\tilde{f}_m = Proj\{f|\mathrm{span}(\phi_1, \ldots, \phi_m)\}.$$

We are interested in the behavior of the m-term approximation error

$$e_m(f;\mathcal{D}) = \|f - \tilde{f}_m\|_2^2,$$

where in this *provisional* definition, we assume $\tilde{f}_m$ is a best approximation of this form after optimizing over the selection of m terms from the dictionary.

However, to avoid a trivial result, we impose regularity on the selection process. Indeed, we allow rather arbitrary dictionaries, including ones which

enumerate a dense subset of $L^2(T)$, so that in some sense the trivial result $\phi_1 = f/\|f\|_2$ $e_m = 0$, $\forall m$ is always a lurking possibility. To avoid this possibility we forbid arbitrary selection rules. Following [10] we propose

Definition. *A sequence of selection rules* $(\sigma_m(\cdot))$ *choosing* m *terms from a dictionary* $\mathcal{D}$,

$$\sigma_m(f) = (\phi_1, \ldots, \phi_m),$$

is said to implement **polynomial depth search** *if there is a single fixed enumeration of the dictionary elements and a fixed polynomial* $\pi(t)$ *such that terms in* $\sigma_m(f)$ *come from the first* $\pi(m)$ *elements in the dictionary.*

Under this definition, the trivial representation based on a countable dense dictionary is not generally available, since in any fixed enumeration, a decent 1-term approximation to typical f will typically be so deep in the enumeration as to be unavailable for polynomial-depth selection. (Of course, one can make this statement quantitative, using information-theoretic ideas).

More fundamentally, our definition not only forbids trivialities, but it allows us to speak of optimal dictionaries and get meaningful results. Starting now, we think of dictionaries as *ordered*, having a first element, second element, etc., so that different enumerations of the same collection of functions are *different* dictionaries. We define the m-optimal approximation number for dictionary $\mathcal{D}$ and limit polynomial π as

$$e_m(f; \mathcal{D}; \pi) = \|f - \tilde{f}_m\|_2^2,$$

where $\tilde{f}_m$ is constructed by optimizing the choice of m atoms among the first $\pi(m)$ in the fixed enumeration. Note that we use *squared error* for comparison with (1)-(3) in the Introduction.

Approximating Classes of Functions

Suppose we now have a class $\mathcal{F}$ of functions whose members we wish to approximate. Suppose we are given a countable dictionary $\mathcal{D}$ and polynomial depth search delimited by polynomial $\pi(\cdot)$.

Define the error of approximation by this dictionary over this class by

$$e_m(\mathcal{F}; \mathcal{D}, \pi) = \max_{f \in \mathcal{F}} e_m(f; \mathcal{D}, \pi).$$

We may find, in certain examples, that we can establish bounds

$$e_m(\mathcal{F}; \mathcal{D}, \pi) = O(m^{-\rho}), \qquad m \to \infty,$$

for all $\rho < \rho^*$. At the same time, we may have available an argument showing that for every dictionary and every polynomial depth search rule delimited by $\pi(\cdot)$,

$$e_m(\mathcal{F}; \mathcal{D}, \pi) \geq c m^{-\rho^*}, \qquad m \geq m_0(\pi).$$

Then it seems natural to say that ρ^* is the optimal rate of m-term approximation from any dictionary when polynomial depth search delimited by $\pi(\cdot)$.

Starshaped Objects with C^2 Boundaries

We define Star-Set$^2(C)$, a class of star-shaped sets with C^2-smooth boundaries, by imposing regularity on the boundaries using a kind of polar coordinate system. Let $\rho(\theta) : [0, 2\pi) \to [0, 1]$ be a radius function and $b_0 = (x_{1,0}, x_{2,0})$ be an origin with respect to which the set of interest is star-shaped. With $\delta_i(x) = x_i - x_{i,0}$, $i = 1, 2$, define functions $\theta(x_1, x_2)$ and $r(x_1, x_2)$ by

$$\theta = \arctan(-\delta_2/\delta_1); \qquad r = ((\delta_1)^2 + (\delta_2)^2)^{1/2}.$$

For a starshaped set, we have $(x_1, x_2) \in B$ iff $0 \le r \le \rho(\theta)$. Define the class Star-Set$^2(C)$ of *sets* by

$$\{B : B \subset [\frac{1}{10}, \frac{9}{10}]^2, \frac{1}{10} \le \rho(\theta) \le \frac{1}{2} \quad \theta \in [0, 2\pi), \quad \rho \in C^2, |\ddot{\rho}(\theta)| \le C\},$$

and consider the corresponding *functional class*

$$\text{Star}^2(C) \;=\; \left\{ f \;=\; 1_B : \quad B \in \text{Star-Set}^2(C) \right\}.$$

The following lower rate bound should be compared with (3).

Lemma. *Let the polynomial $\pi(\cdot)$ be given. There is a constant c so that, for every dictionary $\mathcal{D}$,*

$$e_m(\text{Star}^2(C); \mathcal{D}, \pi) \ge c \frac{1}{m^2 \log(m)}, \qquad m \to \infty.$$

This is proved in [10] by the technique of hypercube embedding. Inside the class Star$^2(C)$ one can embed very high-dimensional hypercubes, and the ability of a dictionary to represent *all* members of a hypercube of dimension n by selecting $m \ll n$ terms from a subdictionary of size $\pi(m)$ is highly limited if $\pi(m)$ grows only polynomially.

To show that the rate (3) can be achieved, [13] adaptively constructs, for each f, a corresponding orthobasis which achieves it. It tracks the boundary of B at increasing accuracy using a sequence of polygons; in fact these are n-gons connecting equispaced points along the boundary of B, for $n = 2^j$. The difference between n-gons for $n = 2^j$ and $n = 2^{j+1}$ is a collection of thin triangular regions obeying $width \approx length^2$; taking the indicators of each region as a term in a basis, one gets an orthonormal basis whose terms at fine scales are thin triangular pieces. Estimating the coefficient sizes by simple geometric analysis leads to the result (3). In fact, [13] shows how to do this under the constraint of polynomial-depth selection, with polynomial Cm^7.

Although space constraints prohibit a full explanation, our polynomial-depth search formalism also makes perfect sense in discussing the warped wavelet representations of the Introduction. Consider the noncountable 'dictionary' of all wavelets in a given basis, with all continuum warpings applied. Notice that for wavelets at a given fixed scale, warpings can be quantized with a certain finite accuracy. Carefully specifying the quantization of the warping, one obtains a countable collection of warped wavelets, for which polynomial depth search constraints make sense, and which is as effective as adaptive triangulation, but *not more so*. Hence (3) applies to (properly interpreted) deformation methods as well.

§5. Curvelet Construction

We now briefly describe the curvelet construction. It is based on combining several ideas, which we briefly review

- Ridgelets, a method of analysis suitable for objects with discontinuities across straight lines.
- Multiscale Ridgelets, a pyramid of windowed ridgelets, renormalized and transported to a wide range of scales and locations.
- Bandpass Filtering, a method of separating an object out into a series of disjoint scales.

We briefly describe each idea in turn, and then their combination.

Ridgelets

The theory of ridgelets was developed in the Ph.D. Thesis of Emmanuel Candès (1998). In that work, Candès showed that one could develop a system of analysis based on ridge functions

$$\psi_{a,b,\theta}(x_1, x_2) = a^{-1/2}\psi((x_1\cos(\theta) + x_2\sin(\theta) - b)/a). \tag{5}$$

He introduced a continuous ridgelet transform $R_f(a, b, \theta) = \langle \psi_{a,b,\theta}(x), f\rangle$ with a reproducing formula and a Parseval relation. He also constructed frames, giving stable series expansions in terms of a special discrete collection of ridge functions. The approach was general, and gave ridgelet frames for functions in $L^2[0,1]^d$ in all dimensions $d \geq 2$ – For further developments, see [3,5].

Donoho [12] showed that in two dimensions, by heeding the sampling pattern underlying the ridgelet frame, one could develop an orthonormal set for $L^2(\mathbb{R}^2)$ having the same applications as the original ridgelets. The orthonormal ridgelets are convenient to use for the curvelet construction, although it seems clear that the original ridgelet frames could also be used. The ortho-ridgelets are indexed using $\lambda = (j, k, i, \ell, \epsilon)$, where j indexes the ridge scale, k the ridge location, i the angular scale, and ℓ the angular location; ϵ is a gender token. Roughly speaking, the ortho-ridgelets look like pieces of ridgelets (5) which are windowed to lie in discs of radius about 2^i; $\theta_{i,\ell} = \ell/2^i$ is roughly the orientation parameter, and 2^{-j} is roughly the thickness.

A formula for ortho-ridgelets can be given in the frequency domain

$$\hat{\rho}_\lambda(\xi) = |\xi|^{-\frac{1}{2}}(\hat{\psi}_{j,k}(|\xi|)w^\epsilon_{i,\ell}(\theta) + \hat{\psi}_{j,k}(-|\xi|)w^\epsilon_{i,\ell}(\theta+\pi))/2\ .$$

Here the $\psi_{j,k}$ are Meyer wavelets for $\mathbb{R}$, $w^\epsilon_{i,\ell}$ are periodic wavelets for $[-\pi, \pi)$, and indices run as follows: $j, k \in \mathbb{Z}$, $\ell = 0, \ldots, 2^{i-1} - 1$; $i \geq 1$, and, if $\epsilon = 0$, $i = \max(1, j)$, while if $\epsilon = 1$, $i \geq \max(1, j)$. We let Λ be the set of such λ.

The formula is an operationalization of the *ridgelet sampling principle*:

- Divide the frequency domain in dyadic coronae $|\xi| \in [2^j, 2^{j+1}]$.
- In the angular direction, sample the j-th corona at least 2^j times.
- In the radial frequency direction, sample behavior using local cosines.

The sampling principle can be motivated by the behavior of Fourier transforms of functions with singularities along lines. Such functions have Fourier transforms which decay slowly along associated lines through the origin in the frequency domain. As one traverses a constant radius arc in Fourier space, one encounters a 'Fourier ridge' when crossing the line of slow decay. The ridgelet sampling scheme tries to represent such Fourier transforms by using wavelets in the angular direction, so that the 'Fourier ridge' is captured neatly by one or a few wavelets. In the radial direction, the Fourier ridge is actually oscillatory, and this is captured by local cosines. A precise quantitative treatment is given in [4].

Multiscale Ridgelets

Think of ortho-ridgelets as objects which have a "length" of about 1 and a "width" which can be arbitrarily fine. The multiscale ridgelet system renormalizes and transports such objects, so that one has a system of elements at all lengths and all finer widths.

In a light mood, we may describe the system impressionistically as "brush strokes" with a variety of lengths, thicknesses, orientations and locations.

The construction employs a nonnegative, smooth partition of energy function w, obeying $\sum_{k_1,k_2} w^2(x_1 - k_1, x_2 - k_2) \equiv 1$. Define a transport operator, so that with index Q indicating a dyadic square $Q = (s, k_1, k_2)$ of the form $[k_1/2^s, (k_1 + 1)/2^s) \times [k_2/2^s, (k_2 + 1)/2^s)$, by $(T_Q f)(x_1, x_2) = f(2^s x_1 - k_1, 2^s x_2 - k_2)$. The Multiscale Ridgelet with index $\mu = (Q, \lambda)$ is then

$$\psi_\mu = 2^s \cdot T_Q(w \cdot \rho_\lambda).$$

In short, one transports the normalized, windowed ortho-ridgelet.

Letting $\mathcal{Q}_s$ denote the dyadic squares of side 2^{-s}, we can define the subcollection of Monoscale Ridgelets at scale s:

$$\mathcal{M}_s = \{(Q, \lambda) : Q \in \mathcal{Q}_s, \lambda \in \Lambda\}.$$

Orthonormality of the ridgelets implies that each system of monoscale ridgelets makes a tight frame, in particular obeying the Parseval relation

$$\sum_{\mu \in \mathcal{M}_s} \langle \psi_\mu, f \rangle^2 = \|f\|_{L^2}^2.$$

It follows that the dictionary of multiscale ridgelets at all scales, indexed by

$$\mathcal{M} = \cup_{s \geq 1} \mathcal{M}_s,$$

is not frameable, as we have energy blow-up:

$$\sum_{\mu \in \mathcal{M}} \langle \psi_\mu, f \rangle^2 = \infty. \tag{6}$$

The Multiscale Ridgelets dictionary is simply too massive to form a good analyzing set. It lacks inter-scale orthogonality – $\psi_{(Q,\lambda)}$ is not typically orthogonal to $\psi_{(Q',\lambda')}$ if Q and Q' are squares at different scales and overlapping locations. In analyzing a function using this dictionary, the repeated interactions with all different scales causes energy blow-up (6).

Our construction of curvelets solves (6) by disallowing the full richness of the Multiscale Ridgelets dictionary. Instead of allowing 'brushstrokes' of all different 'lengths' and 'widths', we allow only those where $width \approx length^2$.

Subband Filtering

Our solution to the 'energy blow-up' (6) is to decompose f into subbands using standard filterbank ideas. Then we assign one specific monoscale dictionary $\mathcal{M}_s$ to analyze one specific (and specially chosen) subband.

We define coronae of frequencies $|\xi| \in [2^{2s}, 2^{2s+2}]$, and subband filters Δ_s extracting components of f in the indicated subbands; a filter P_0 deals with frequencies $|\xi| \le 1$. The filters decompose the energy exactly into subbands:

$$\|f\|_2^2 = \|P_0 f\|_2^2 + \sum_s \|\Delta_s f\|_2^2.$$

The construction of such operators is standard [15]; the coronization oriented around powers 2^{2s} is nonstandard – and essential for us. Explicitly, we build a sequence of filters Φ_0 and $\Psi_{2s} = 2^{4s}\Psi(2^{2s}\cdot)$, $s = 0, 1, 2, \ldots$ with the following properties: Φ_0 is a lowpass filter concentrated near frequencies $|\xi| \le 1$; Ψ_{2s} is bandpass, concentrated near $|\xi| \in [2^{2s}, 2^{2s+2}]$; and we have

$$|\hat{\Phi}_0(\xi)|^2 + \sum_{s \ge 0} |\hat{\Psi}(2^{-2s}\xi)|^2 = 1, \quad \forall \xi.$$

Hence, Δ_s is simply the convolution operator $\Delta_s f = \Psi_{2s} * f$.

Definition of Curvelet Transform

Assembling the above ingredients, we are able to sketch the definition of the Curvelet transform. We let M' consist of M merged with the collection of integral pairs (k_1, k_2) indexing unit-side squares in the plane.

The curvelet transform is a map $L^2(\mathbb{R}^2) \mapsto \ell^2(M')$, yielding Curvelet coefficients $(\alpha_\mu : \mu \in M')$. These come in two types.

At *coarse scale* we have wavelet scaling function coefficients

$$\alpha_\mu = \langle \phi_{k_1,k_2}, P_0 f \rangle, \qquad \mu = (k_1, k_2) \in M' \backslash M,$$

where ϕ_{k_1,k_2} is the Lemarié scaling function of the Meyer basis, while at *fine scale* we have Multiscale Ridgelets coefficients of the bandpass filtered object:

$$\alpha_\mu = \langle \Delta_s f, \psi_\mu \rangle, \qquad \mu \in M_s, s = 1, 2, \ldots.$$

Note well that each coefficient associated to scale 2^{-s} derives from the subband filtered version of $f - \Delta_s f$ – and not from f. Several properties are immediate:

- Tight Frame:
$$\|f\|_2^2 = \sum_{\mu \in M'} |\alpha_\mu|^2.$$

- Existence of Coefficient Representers (Frame Elements):
$$\alpha_\mu \equiv \langle f, \gamma_\mu \rangle.$$

- L^2 Reconstruction Formula:
$$f = \sum_{\mu \in M'} \langle f, \gamma_\mu \rangle \gamma_\mu.$$

- Formula for Frame Elements:
$$\gamma_\mu = \Delta_s \psi_\mu, \qquad \mu \in \mathcal{Q}_s.$$
In short, the curvelets are obtained by bandpass filtering of Multiscale Ridgelets with *passband* is rigidly linked to the *scale* of spatial localization

- Anisotropy Scaling Law: Linking the filter passband $|\xi| \approx 2^{2s}$ to the spatial scale 2^{-s} imposes that (1) most curvelets are negligible in norm (most multiscale ridgelets do not survive the bandpass filtering Δ_s); (2) the non-negligible curvelets obey $length \approx 2^{-s}$ while $width \approx 2^{-2s}$. So the system obeys approximately the scaling relationship
$$width \approx length^2.$$
It is here that the 2^{2s} coronization scheme comes into play.

§6. Why Should This Work?

The curvelet decomposition can be equivalently stated in the following form.

- *Subband Decomposition.* The object f is filtered into subbands:
$$f \mapsto (P_0 f, \Delta_1 f, \Delta_2 f, \ldots).$$

- *Smooth Partitioning.* Each subband is smoothly windowed into "squares" of an appropriate scale:
$$\Delta_s f \mapsto (w_Q \Delta_s f)_{Q \in \mathcal{Q}_s}.$$

- *Renormalization.* Each resulting square is renormalized to unit scale
$$g_Q = 2^{-s} (T_Q)^{-1} (w_Q \Delta_s f), \qquad Q \in \mathcal{Q}_s.$$

- *Ridgelet Analysis.* Each square is analyzed in the ortho-ridgelet system.

$$\alpha_\mu = \langle g_Q, \rho_\lambda \rangle, \qquad \mu = (Q, \lambda).$$

We can now give a crude explanation why the main result (4) holds. Effectively, the bandpass images $\Delta_s f$ are almost vanishing at x far from the edges in f. Along edges, the bandpass images exhibit *ridges* of width $\approx 2^{-2s}$ – the width of the underlying bandpass filter.

The partitioned bandpass images are broken into squares of side $2^{-s} \times 2^{-s}$. The squares which do not intersect edges have effectively no energy, and we ignore them. The squares which do intersect edges result from taking a nearly-straight ridge and windowing it. Thus the squares which 'matter' exhibit tiny *ridge fragments* of aspect ratio 2^{-s} by 2^{-2s}. After renormalization, the resulting g_Q exhibits a ridge fragment of about unit length and of width 2^{-s}. The ridge fragment is then analyzed in the ortho-ridgelet system, which should (we hope) yield only a few significant coefficients.

In fact, simple arguments of size and order give an idea how the curvelet coefficients roughly behave. We give an extremely loose description.

First, at scale 2^{-s}, there are only about $O(2^s)$ squares $Q \in \mathcal{Q}_s$ that interact with the edges. Calculating the energy in such a square using the size of the support and the height of the ridge leads to

$$(\text{length} \cdot \text{width})^{1/2} \cdot \text{height} \approx (2^{-s} \times 2^{-2s})^{1/2} \times 1.$$

Indeed, the height of the ridge is bounded by

$$\|\Delta_s f\|_\infty = \|\Psi_{2s} * f\|_\infty \le \|\Psi_{2s}\|_1 \|f\|_\infty = \|\Psi\|_1 \|f\|_\infty.$$

Since we are interested in uniformly bounded functions f, the height is thus bounded by a constant C. The calculation of the norm $\|g_Q\|_2 \approx 2^{-3/2}$ follows immediately (because of the renormalization, the height of the ridge g_Q is now 2^{-s}) .

Now *temporarily* suppose that for some fixed K not depending on Q

$$\textit{each ridge fragment } g_Q \textit{ is a sum of at most } K \textit{ ortho-ridgelets.} \tag{7}$$

This would imply that at level s we have a total number of coefficients

$$O(2^s) \text{ squares which 'matter' } \times K \text{coefficients/square},$$

while the norm estimate for g_Q and the orthonormality of ridgelets give

$$\text{coefficient amplitude} \le C \cdot 2^{-3s/2}.$$

The above assumptions imply that the N-th largest curvelet coefficient is of size $\le C \cdot N^{-3/2}$. Letting $|\alpha|_{(N)}$ denote the N-th coefficient amplitude, the tail sum of squares would obey

$$\sum_{N>m} |\alpha|_{(N)}^2 \le C \cdot m^{-2}. \tag{8}$$

This coefficient decay leads to (4) as follows. Let $\mu_1, \ldots, \mu_m$ enumerate indices of the m largest curvelet coefficients. Build the m-term approximation

$$\tilde{f}_m^C = \sum_{i=1}^{m} \alpha_{\mu_i} \gamma_{\mu_i}.$$

By the tight frame property,

$$\|f - \tilde{f}_m^C\|^2 \leq \sum_{N=m+1}^{\infty} |\alpha|_{(N)}^2, \leq C \cdot m^{-2},$$

where the last step used (8). This of course would establish (4) – in fact something even stronger, something fully as good as (3).

However, we have temporarily assumed (7) – which is not true. Each ridge fragment generates a countable number of nonzero ridgelet coefficients in general. The paper [6] gets (4) using much more subtle estimates.

§7. Discussion

Why Call These Things Curvelets?

The visual appearance of curvelets does not match the name we have given them. The curvelets waveforms look like brushstrokes; brushlets would have been an appropriate name, but it was taken already, by F. Meyer and R. Coifman, for an unrelated scheme (essentially, Gabor Analysis).

Our point of view is simply that curvelets exemplify a certain *curve scaling law* – $width = length^2$ – which is naturally associated to curves.

A deeper connection between curves and curvelets was alluded to in our talk at *Curves and Surfaces '99*. Think of a curve in the plane as a distribution supported on the curve, in the same way that a point in the plane can be thought of as a Dirac distribution supported on that point. The curvelets scheme can be used to represent that distribution as a superposition of functions of various lengths and widths obeying the scaling law $width = length^2$. In a certain sense this is a near-optimal representation of the distribution.

The Analogy and The Surprise

Sections 2 and 3 showed that wavelets do surprisingly well in representing *point singularities*. Without attempting an explicit representation of 'where the bad points are', wavelets do essentially as well as ideal adaptive schemes in representing the point singularities.

Sections 4–6 showed that the non-adaptive curvelets representation can do nearly as well in representing objects with discontinuities along curves as adaptive methods that explicitly track the shape of the discontinuity and use a special adaptive representation dependent on that tracking.

We find it surprising and stimulating that the curvelet representation can work so well despite the fact that it never constructs or depends on the existence of any 'map' of the discontinuity set.

We also find it interesting that there is a system of analysis which plays the role for curvilinear singularities that wavelets play for point singularities.

Acknowledgments. This research was supported by National Science Foundation grants DMS 98–72890 (KDI), DMS 95–05151, and by AFOSR MURI 95–P49620–96–1–0028.

References

1. Candès, E. J., Harmonic Analysis of Neural Networks, Appl. Comput. Harmon. Anal. **6** (1999), 197–218.
2. Candès, E. J., Ridgelets: theory and applications, Ph.D. Thesis, Statistics, Stanford ,1998.
3. Candès, E. J., Monoscale ridgelets for the representation of images with edges, Technical Report, Statistics, Stanford, 1999.
4. Candès, E. J., On the representation of mutilated Sobolev functions, Technical Report, Statistics, Stanford, 1999.
5. Candès, E. J., and D. L. Donoho, Ridgelets: The key to High-Dimensional Intermittency? Phil. Trans. R. Soc. Lond. A. **357** (1999), 2495-2509.
6. Candès, E. J., and D. L. Donoho, Curvelets, Manuscript, 1999.
7. Coifman, R. R., and M. V. Wickerhauser, Entropy-based algorithms for best basis selection, in IEEE Trans. Inform. Theory **38** (1992), 1713–1716.
8. Deng, B., and B. Jawerth, and G. Peters, and W. Sweldens, Wavelet Probing for Compression-based segmentation, in Proc. SPIE Symp. Math. Imaging: Wavelet Applications in Signal and Image Processing, 1993. Proceedings of SPIE conference July 1993, San Diego.
9. Donoho, D. L., Minimum Entropy Segmentation, in Wavelets: Theory, Algorithms and Applications, C. K. Chui, L. Montefusco and L. Puccio (eds.), Academic Press, San Diego, 1994, 233–270.
10. Donoho, D. L., and I. M. Johnstone, Empirical Atomic Decomposition, Manuscript, 1995.
11. Donoho, D. L., Wedgelets: nearly minimax estimation of edges, Ann. Statist. **27** (1999), 859-897.
12. Donoho, D. L., Orthonormal Ridgelets and Linear Singularities, Technical Report, Statistics, Stanford, 1998. To appear in SIAM J. Math. Anal.
13. Donoho, D. L., Sparse Components Analysis and Optimal Atomic Decomposition, Technical Report, Statistics, Stanford, 1998.
14. Meyer, Y., *Wavelets and Operators*, Cambridge University Press, 1992.
15. Vetterli, M., and J. Kovacevic, *Wavelets and Subband Coding*, Prentice Hall, 1995.

Department of Statistics
Stanford University
Stanford, CA 94305-4065
{emmanuel,donoho}@stat.stanford.edu
http://www-stat.stanford.edu/{∼emmanuel,∼donoho}/