Information-theoretic Interpretation of Besov Spaces

Hyeokho Choi and Richard Baraniuk

Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005, USA

ABSTRACT

Besov spaces classify signals and images through the Besov norm, which is based on a deterministic smoothness measurement. Recently, we revealed the relationship between the Besov norm and the likelihood of an independent generalized Gaussian wavelet probabilistic model. In this paper, we extend this result by providing an information-theoretic interpretation of the Besov norm as the Shannon codelength for signal compression under this probabilistic mode. This perspective unites several seemingly disparate signal/image processing methods, including denoising by Besov norm regularization, complexity regularized denoising, minimum description length (MDL) processing, and maximum smoothness interpolation. By extending the wavelet probabilistic model (to a locally adapted Gaussian model), we broaden the notion of smoothness space to more closely characterize real-world data. The locally Gaussian model leads directly to a powerful wavelet-domain Wiener filtering algorithm for denoising.

Keywords: Besov spaces, wavelets, information theory, compression, Shannon codelength, signal estimation, MDL

1. INTRODUCTION

Many powerful wavelet-domain signal and image processing algorithms owe their success to improved modeling of the wavelet coefficients of real-world data. Signal and image models naturally fall into two categories: deterministic and probabilistic. Deterministic Besov space models, which correspond to function spaces of certain smoothness, have been used for motivating, designing, and analyzing many algorithms.[1–3] Probabilistic models have been used extensively in statistical estimation, in particular Bayesian inference algorithms.[4,5] Recently, we established a strong connection between deterministic Besov space models and a class of very simple wavelet-domain probabilistic models.[3] In this paper, we extend these ideas to a new information-theoretic interpretation of Besov spaces that connects to image-compression theory.

Many real-world signals and images consist of piecewise smooth sections separated by discontinuities. The Besov norm essentially measures the smoothness between these discontinuities. The Besov spaces a signal or image belongs to is also characterized by the asymptotic decay of its nonlinear approximation error.[6] While the space-domain definition of the Besov norm is somewhat complicated, there exists a simple and clear equivalent formula in terms of the wavelet coefficients (because the smoothness of a signal is related to the wavelet coefficient decay).

The wavelet-domain properties of Besov space functions have been explored deeply in the world of nonlinear approximation theory,[6] and many practical processing algorithms have been developed.[1,3,7] Donoho[7] showed that the simple wavelet shrinkage algorithm for noise removal is optimal for the functions in Besov spaces. In the area of image compression, the seminal paper of DeVore et al[8] first demonstrated the relationship between the Besov space properties of real-world data and image compression algorithms. Recently, the Besov smoothness of images was shown to be directly related to the rate-distortion analysis of practical image compression algorithms.[9] Chambolle et al[1] recently formulated a Besov norm penalized least square denoising scheme that yields simple wavelet shrinkage algorithms for certain Besov parameters. Choi et al[2] proposed a maximum Besov smoothness interpolation algorithm that is well-suited to data containing singularities.

Recently, we illustrated a strong connection[3] between the Besov norm and a simple wavelet-domain generalized Gaussian statistical model that has been used extensively to model real-world signals in the wavelet domain.[10,11] In this paper, we extend this result and give a new interpretation of the Besov norm as a compressibility measure. In particular, we show that Besov norm is equivalent to the *Shannon codelength* of the signal/image under the independent generalized Gaussian model considered in Choi et al.[3] This information-theoretic interpretation of Besov norm unifies many seemingly different algorithms, including Besov regularization,[1] MDL denoising,[12] compression-based denoising,[13] maximum smoothness interpolation,[2] and so on.

Email: choi@ece.rice.edu, richb@rice.edu. Web: www.dsp.rice.edu. This work was supported by NSF grants MIP–9457438 and CCR–9973188, DARPA/AFOSR grant F49620–97–1–0513, ONR grant N00014–99-1–0813, and Texas Instruments.

The limitations of the underlying generalized Gaussian statistical model manifest the problems of Besov characterization of real-world images. To overcome this difficulty, we extend this model to the local Gaussian model implicit in the expectation-quantization (EQ) image coder[14,15] to more accurately characterize real-world data.

2. WAVELET STATISTICAL MODELS AND BESOV SPACE

2.1. Wavelet transform

The discrete wavelet transform (DWT) represents a one-dimensional (1-D) signal $z(t)$ in terms of shifted versions of a lowpass scaling function $\phi(t)$ and shifted and dilated versions of a prototype bandpass wavelet function $\psi(t)$.[16] For special choices of $\phi(t)$ and $\psi(t)$, the functions $\psi_{j,k}(t) \equiv 2^{j/2}\psi(2^j t - k)$ and $\phi_{j,k}(t) \equiv 2^{j/2}\phi(2^j t - k)$ with $j, k \in \mathbb{Z}$ form an orthonormal basis, and we have the representation[16]

$$z \;=\; \sum_k u_{j_0,k}\,\phi_{j_0,k} \;+\; \sum_{j=j_0}^{\infty}\sum_k w_{j,k}\,\psi_{j,k}, \tag{1}$$

with $u_{j,k} \;\equiv\; \int z(t)\,\phi^*_{j,k}(t)\,dt$ and $w_{j,k} \;\equiv\; \int z(t)\,\psi^*_{j,k}(t)\,dt$.

The *wavelet coefficient* $w_{j,k}$ measures the signal content around time $2^{-j}k$ and frequency $2^j f_0$. The scaling coefficient $u_{j,k}$ measures the local mean around time $2^{-j}k$. The DWT (1) employs scaling coefficients only at scale j_0; wavelet coefficients at scales $j > j_0$ add higher resolution details to the signal.

We can easily construct two-dimensional (2-D) wavelets from the 1-D ψ and ϕ by setting for $\mathbf{x} \equiv (x, y) \in \mathbb{R}^2$, $\psi^{\mathrm{HL}}(x,y) \;\equiv\; \psi(x)\phi(y), \psi^{\mathrm{LH}}(x,y) \;\equiv\; \phi(x)\psi(y), \psi^{\mathrm{HH}}(x,y) \;\equiv\; \psi(x)\psi(y)$, and $\phi(x,y) \;\equiv\; \phi(x)\phi(y)$. If we let $\Psi \equiv \{\psi^{\mathrm{HL}}, \psi^{\mathrm{LH}}, \psi^{\mathrm{HH}}\}$, then the set of functions $\{\psi_{j,\mathbf{k}} \equiv 2^j\psi(2^j\mathbf{x}-\mathbf{k})\}_{\psi\in\Psi, j\in\mathbf{Z},\mathbf{k}\in\mathbf{Z}^2}$ and $\{\phi_{j,\mathbf{k}} \equiv 2^j\phi(2^j\mathbf{x}-\mathbf{k})\}_{j\in\mathbf{Z},\mathbf{k}\in\mathbf{Z}^2}$ forms an orthonormal basis for $L_2(\mathbb{R}^2)$. That is, for every $z \in L_2(\mathbb{R}^2)$, we have

$$z = \sum_{j>j_0,\mathbf{k}\in\mathbf{Z}^2,\psi\in\Psi} w_{j,\mathbf{k},\psi}\psi_{j,\mathbf{k}} + \sum_{\mathbf{k}\in\mathbf{Z}^2} u_{j_0,\mathbf{k}}\phi_{j_0,\mathbf{k}}, \tag{2}$$

with $w_{j,\mathbf{k},\psi} \;\equiv\; \int_{\mathbb{R}^2} z(\mathbf{x})\psi_{j,\mathbf{k}}(\mathbf{x})d\mathbf{x}$ and $u_{j_0,\mathbf{k}} \;\equiv\; \int_{\mathbb{R}^2} z(\mathbf{x})\phi_{j_0,\mathbf{k}}(\mathbf{x})d\mathbf{x}$. We will emphasize 2-D images in the sequel, although the results apply equally well in 1-D.

For the computer processing of images, we sample the continuous image $z(\mathbf{x})$ on a grid. With proper prefiltering, we can approximate the discrete samples as the scaling coefficients of $z(\mathbf{x})$ at a certain scale J; that is, the sampled image $z(\mathbf{k}) = u_{J,\mathbf{k}}$. Equivalently, we can build a continuous-time image $\widetilde{z}$ corresponding to $z(\mathbf{k})$ as

$$\widetilde{z} = \sum_{k\in\mathbf{Z}^2} u_{J,\mathbf{k}}\phi_{J,\mathbf{k}}, \tag{3}$$

or, using the wavelets

$$\widetilde{z} = \sum_{j_0<j<J,\mathbf{k}\in\mathbf{Z}^2,\psi\in\Psi} w_{j,\mathbf{k},\psi}\psi_{j,\mathbf{k}} + \sum_{k\in\mathbf{Z}^2} u_{j_0,\mathbf{k}}\phi_{j_0,\mathbf{k}}. \tag{4}$$

The coefficients $w_{j,\mathbf{k},\psi}$ and $u_{j_0,\mathbf{k}}$ are easily computed using the 2-D discrete-time wavelet filters and decimators operating on the samples $z(\mathbf{k})$.[16] As a notation, we define $\boldsymbol{w}$ as the set of all wavelet coefficients.

2.2. Besov function spaces

The theory of smoothness function spaces plays an ever more important role in signal and image processing. We will consider the family of *Besov spaces* $B^\alpha_q(L_p(I))$ over a finite domain I, for example, the square $[0,1]^2$, for $0 < \alpha < \infty$, $0 < p \leq \infty$, and $0 < q \leq \infty$. These spaces have, very roughly speaking, "α derivatives in $L_p(I)$"; the parameter q allows us to make finer distinctions in smoothness.[8]

For $r > 0$ and $h \in \mathbb{R}^2$, define the r-th difference of a function z as

$$\Delta^{(r)}_h z(t) \equiv \sum_{k=0}^{r} \begin{pmatrix} r \\ k \end{pmatrix} (-1)^k z(t + kh) \tag{5}$$

for $t \in I_h^r \equiv \{t \in I | t + rh \in I\}$. The $L_p(I)$-*modulus of smoothness* for $0 < p \leq \infty$ is defined as

$$\omega_r(z,t)_p \equiv \sup_{|h| \leq t} \|\Delta_h^{(r)} z\|_{L_p(I_h^r)}. \tag{6}$$

The Besov seminorm of index (α, p, q) is defined for $r > \alpha$, where $1 \leq p, q \leq \infty$, by

$$|z|_{B_q^\alpha(L_p(I))} \equiv \left\{ \int_0^\infty \left(\frac{\omega_r(z,t)_p}{t^\alpha} \right)^q \frac{dt}{t} \right\}^{1/q} \quad \text{if } 0 \leq q < \infty \tag{7}$$

and by

$$|z|_{B_\infty^\alpha(L_p(I))} \equiv \sup_{0<t<\infty} \left\{ \frac{\omega_r(z,t)_p}{t^\alpha} \right\} \text{ if } q = \infty. \tag{8}$$

The Besov space norm is defined as

$$\|z\|_{B_q^\alpha(L_p(I))} \equiv |z|_{B_q^\alpha(L_p(I))} + \|z\|_{L_p(I)}. \tag{9}$$

The Besov space $B_q^\alpha(L_p(I))$ is then the class of functions $z : I \longrightarrow \mathbb{R}$ satisfying $z \in L_p(I)$ and $|z|_{B_q^\alpha(L_p(I))} < \infty$. Various settings of the parameters yield familiar spaces. For example, when $p = q = 2$, $B_2^\alpha(L_2(I))$ is the Sobolev space $W^\alpha(L_2(I))$, and when $\alpha < 1$, $1 \leq p \leq \infty$, $q = \infty$, $B_\infty^\alpha(L_p(I))$ is the Lipschitz space.

Wavelets provide a simple characterization for the Besov spaces. For analyzing ϕ and ψ possessing $r > \alpha$ vanishing moments,[17] the Besov norm $\|z\|_{B_q^\alpha(L_p(I))}$ is equivalent to the sequence norm

$$\|z\|_{B_q^\alpha(L_p(I))} \asymp |u_{j_0,\mathbf{k}}| + \left(\sum_{j \geq j_0} \left(\sum_{\mathbf{k},\psi\in\Psi} 2^{\alpha j p} 2^{j(p-2)} |w_{j,\mathbf{k},\psi}|^p \right)^{q/p} \right)^{1/q}. \tag{10}$$

The three hyperparameters have natural interpretations: a p-norm of the wavelet coefficients is taken within each scale j, a q-norm is taken across scale, and the smoothness parameter α controls the rate of decay of the $w_{j,\mathbf{k},\psi}$ across scale (frequency). We will take (10) as the definition of the Besov norm in the following.

For signal and image processing applications, there is a particular case of interest: When $p = q$, the Besov norm reduces to

$$\|z\|_{B_p^\alpha(L_p(I))} \asymp |u_{j_0,\mathbf{k}}| + \left(\sum_{j \geq j_0, \mathbf{k}, \psi \in \Psi} 2^{\alpha j p} 2^{j(p-2)} |w_{j,\mathbf{k},\psi}|^p \right)^{1/p}, \tag{11}$$

which is a weighted l_p norm of the wavelet coefficients.

2.3. Independent generalized Gaussian wavelet model

Each wavelet basis function analyzes an image locally and produces a large wavelet coefficient when the corresponding region contains an edge. Conversely, if the region has no edges, the wavelet coefficient is small. Because the total area occupied by edge regions in a typical real-world image is very small, the majority of the wavelet coefficients are small, and the distribution of the coefficients has a heavy tail corresponding to the large edge coefficients. This *energy compaction* property of the wavelet transform results in a nonGaussian distribution of wavelet coefficients that is peaked at zero and heavy tailed.

Another desirable property of the wavelet transform is that it approximates the Karhunen-Loève transform for real-world images, and thus the resulting wavelet coefficients are approximately decorrelated. In simple cases, this permits us to consider the wavelet coefficients as independent random variables, simplifying the modeling and processing of wavelet coefficients enormously. The primary independent models employed to date are the generalized Gaussian distribution (GGD)[10,11] and Gaussian mixture distribution (GMD).[4,5] We will emphasize the GGD in the sequel.

The zero-mean generalized Gaussian density $\text{GGD}_\nu(0, \sigma^2)$ with shape parameter ν and variance σ^2 is defined as

$$f(x) \equiv \frac{\nu \eta(\nu)}{2\Gamma(1/\nu)} \frac{1}{\sigma} \exp\left\{ -[\eta(\nu)|x|/\sigma]^\nu \right\}, \tag{12}$$

with $\eta(\nu) \equiv \sqrt{\frac{\Gamma(3/\nu)}{\Gamma(1/\nu)}}$. The GGD model contains the Gaussian and Laplacian distribution as special cases, using $\nu = 2$ and $\nu = 1$, respectively.

In an independent GGD wavelet model, each wavelet coefficient is generated independently according to a zero-mean GGD. For the tractability of the model, all wavelet coefficients at each scale are assumed to be independent and identically distributed (iid). We will refer to the model under this assumption as *iid in scale.* That is, under the independent GGD model, we have $w_{j,\mathbf{k},\psi} \overset{\text{iid}}{\sim} \text{GGD}_{\nu_j}(0, \sigma_j^2)$. Due to the iid-in-scale assumption, the shape parameter ν_j and the variance σ_j^2 do not depend on the spatial location $\mathbf{k}$.

The shape parameter ν_j of the GGD function at scale j characterizes the peakiness and heavy tailedness of the wavelet coefficient distribution. In many cases, we can specify the shape parameters without reference to the given data, because most real-world images tend to have similar wavelet-domain energy compaction properties.[11] In practical applications, further simplification follows from assuming that the shape parameter ν_j is the same for all scales.

The variance σ_j^2 represents the signal energy at scale j. It can be empirically estimated based on the given data by estimating the variance of wavelet coefficients at scale j,[11] or it can be specified to decay exponentially. This is natural for signals and images with a $1/f$ type power spectrum.[4]

The simplest GGD model (with exponentially decaying variances) thus takes the form:

$$w_{j,\mathbf{k},\psi} \overset{\text{iid}}{\sim} f_{j,\mathbf{k},\psi}(w_{j,\mathbf{k},\psi}) = \text{GGD}_{\nu}(0, \sigma_j^2) \text{ with } \sigma_j = 2^{-j\beta}\sigma_0. \tag{13}$$

2.4. Likelihood interpretation of Besov norm

Given the probability density function (pdf) $f(x|\boldsymbol{\Theta})$ of the random variable X under a model $\boldsymbol{\Theta}$ and two realizations x_1 and x_2, we can compute the likelihoods of the data $f(x_1|\boldsymbol{\Theta})$ and $f(x_2|\boldsymbol{\Theta})$. Because the likelihoods indicate how "likely" the given data are under the given pdf, $f(x_1|\boldsymbol{\Theta}) > f(x_2|\boldsymbol{\Theta})$ implies that x_1 is more likely than x_2 under the model.

However, the likelihood $f(x_1|\boldsymbol{\Theta})$ is meaningful only when compared against another likelihood, such as $f(x_2|\boldsymbol{\Theta})$. That is, we cannot tell how likely x_1 is based merely on the value $f(x_1|\boldsymbol{\Theta})$. To be able to tell how likely an observation is under a given pdf model, we must normalize the likelihood appropriately. A natural way of normalizing is to compare the likelihood with the maximum likelihood achievable. We define the *normalized likelihood function* $f^N(x|\boldsymbol{\Theta})$ by

$$f^N(x|\boldsymbol{\Theta}) \equiv \frac{f(x|\boldsymbol{\Theta})}{\sup_x f(x|\boldsymbol{\Theta})}, \tag{14}$$

with the assumption that $0 < \sup_x f(x|\boldsymbol{\Theta}) < \infty$. Then, $f^N(x|\boldsymbol{\Theta}) \in [0, 1]$, and we can state that an observation x is "likely" if $f^N(x|\boldsymbol{\Theta})$ is close to 1 and "not likely" if it is close to 0. We can easily generalize the concept of normalized likelihood to finite random vectors using the joint pdf.

For discrete random processes, the joint pdf of an infinite random vector is not defined, and we cannot define the normalized likelihood as in (14). However, we can generalize the normalized likelihood using the limit of the normalized likelihoods when the limit exists. Let $X_1, X_2, \ldots$ be an infinite sequence of random variables. Then for the vector of the first n random variables $\boldsymbol{x}_n = \{X_k\}_{k=1}^n$, define the normalized likelihood $f_n^N(\boldsymbol{x}_n)$ using (14). Then, if $\lim_{n\to\infty} f_n^N$ exists, we define the normalized likelihood of the infinite sequence to be the limit.

For the independent GGD model (13), we can take the limit as we move to the fine scales, defining the normalized likelihood

$$f^N(\boldsymbol{w}) \equiv \lim_{J\to\infty} \frac{\prod_{j=0}^J \prod_{\mathbf{k},\psi} f_{j,\mathbf{k},\psi}(w_{j,\mathbf{k},\psi})}{\sup \prod_{j=0}^J \prod_{\mathbf{k},\psi} f_{j,\mathbf{k},\psi}(w_{j,\mathbf{k},\psi})}, \tag{15}$$

with $w_{j,\mathbf{k},\psi} \sim f_{j,\mathbf{k},\psi}(w_{j,\mathbf{k},\psi})$.

Now, consider an independent GGD wavelet model with shape parameter $\nu = p$ and exponentially decaying variance across scale:

$$\boldsymbol{\Theta}_p^\alpha \ : \ w_{j,\mathbf{k},\psi} \overset{\text{iid}}{\sim} \text{GGD}_p(0, \sigma_j^2) \text{ with } \sigma_j = 2^{-j(\alpha+2-2/p)}\sigma_0. \tag{16}$$

The normalized likelihood computed using the coefficients between scales 0 and J is then

$$\begin{aligned} f_J^N(\mathrm{w}) &= \prod_{j=0}^{J}\prod_{\mathbf{k},\psi} \exp\left\{-[\eta(p)|w_{j,\mathbf{k},\psi}|/\sigma_j]^p\right\} \\ &= \exp\left\{\sum_{0\leq j\leq J,\mathbf{k},\psi} -[\eta(p)|w_{j,\mathbf{k},\psi}|/\sigma_j]^p\right\}. \end{aligned} \tag{17}$$

Computing the negative log of the normalized likelihood function, we obtain the negative log normalized likelihood function:

$$\begin{aligned} -\log f^N(\boldsymbol{w}) &= \sum_{0\leq j,\mathbf{k},\psi} [\eta(p)|w_{j,\mathbf{k},\psi}|/\sigma_j]^p \\ &= [\eta(p)]^p \sum_{0\leq j,\mathbf{k},\psi} (2^{j(\alpha+2-2/p)}|w_{j,\mathbf{k},\psi}|)^p, \end{aligned} \tag{18}$$

which is equivalent to $\|z\|^p_{B_p^\alpha(L_p)}$, the homogeneous ($p=q$) Besov norm of the function z (within $|u_{j_0,\mathbf{k}}|$).[3] When $p=2$, we obtain an iid Gaussian model for the wavelet coefficients, and the corresponding normalized negative log likelihood is equivalent to the Sobolev norm of the function.

In terms of the normalized likelihood function for the iid GGD model $\boldsymbol{\Theta}_p^\alpha$, the Besov space $B_p^\alpha(L_p)$ can be equivalently defined as the set

$$\{\boldsymbol{w} : f^N(\boldsymbol{w}|\boldsymbol{\Theta}_p^\alpha) \neq 0\}. \tag{19}$$

Thus, functions in the Besov space $B_p^\alpha(L_p)$ are the "likely" images under the statistical model $\boldsymbol{\Theta}_p^\alpha$.

2.5. Besov balls for sampled data

In practice, the available data are given as finite samples of the signal or image, and we have no access to the wavelet coefficients beyond a certain finest scale. Suppose we have wavelet coefficients up to scale J and the coefficients with scale $j > J$ are unknown. Let $\boldsymbol{w}_J$ be the vector of available wavelet coefficients. Because the definition of Besov spaces is concerned with the asymptotic decay of the signal energy across scale, it is not clear how the Besov space theory can be applied in this case. Using the statistical connection in (19), we can modify the Besov space definition for the finite data $\boldsymbol{w}_J$.

The set of "likely" images can be modified as

$$\{\boldsymbol{w}_J : f_J^N(\boldsymbol{w}_J|\boldsymbol{\Theta}_p^\alpha) \geq \epsilon\}, \tag{20}$$

where $f_J^N(\boldsymbol{w}_J|\boldsymbol{\Theta}_p^\alpha)$ is the normalized pdf of the (finite) vector $\boldsymbol{w}_J$, and ϵ is a positive threshold that distinguishes between "likely" and "unlikely" images. We can choose ϵ so that the set (20) contains the images of interest.

Under the independent GGD model, the negative log likelihood function is the truncated form of the Besov norm defined in (11), and the set of images is the "ball" in Besov space defined as $\{x : \|x\|_B \leq R\}$, with R the radius of the ball and $\|\cdot\|_B$ the truncated Besov norm for finite samples. Although we consider homogeneous ($p=q$) Besov spaces $B_p^\alpha(L_p)$ above, Besov balls can be defined for other scales of Besov spaces in the same way. In particular, the space $B_\infty^1(L_1)$ is desirable for modeling images.[1] For $p, q \geq 1$, the Besov balls are convex sets in l_p.

3. INFORMATION-THEORETIC INTERPRETATION OF BESOV SPACES

In the previous section, we showed that the Besov norm is a normalized negative log likelihood under a natural wavelet statistical model. Now, since the negative log likelihood is the classical Shannon codelength required to encode an image,[18] we can interpret the Besov norm as the Shannon codelength of the image. This is very reasonable: we expect that rougher images should take more bits to code and thus be more complex.

When the density function of each wavelet coefficient is symmetric and unimodal with zero mean, the supremum of the likelihoods in the denominator in (14) occurs when all the coefficients are zero. Because an all-zero wavelet expansion corresponds to a constant image, the normalization in (14) is with respect to a constant image. Therefore, the normalized likelihood (14) is the ratio of the likelihood of the image compared against the (simplest) constant image.

Computing the $-\log$ of the normalized likelihood, we can write

$$\begin{aligned} -\log f^N(\mathrm{w}) &= -\log(\text{Likelihood of image}) - [-\log(\text{Likelihood of constant image})] \\ &= \text{Shannon codelength of image} - \text{Shannon codelength of constant image}. \end{aligned} \tag{21}$$

Since $-\log f^N(\mathrm{w}) = [\eta(p)]^p \|z\|^p_{B^\alpha_p(L_p)}$, we obtain

$$[\eta(p)]^p \|z\|^p_{B^\alpha_p(L_p)} = \text{Shannon codelength of } z - \text{Shannon codelength of constant image}. \tag{22}$$

Because the constant image is the simplest image to code, the Besov norm (ignoring the constant $[\eta(p)]^p$) measures how much more codelength is required to code the image z beyond that required for the simplest image. In this sense, the Besov norm measures the coding complexity of the image (under the assumed iid GGD model).

In practical image compression algorithms, the GGD distributions are a popular choice for the wavelet coefficient pdf. For example, the EQ image coder models each wavelet coefficient as a GGD distribution with variance adapted to its spatial location.[14,15] If we ignore the spatial adaptation and simply incorporate the general (exponential) coefficient decay across scale, then the iid GGD model in (16) follows. Thus, the Besov norm is equivalent to the codelength resulting from a very crude form of EQ type image compression.

4. APPLICATIONS AND RELATION TO OTHER WORK

The problem of image estimation in additive white Gaussian noise (AWGN) can be formally described as estimating an unknown original image $\underline{x}$ from the corrupted observations

$$\underline{y} = \underline{x} + \underline{n}, \tag{23}$$

where $\underline{n}$ is white Gaussian noise with zero mean and known variance σ_n^2. Under this model, a regularized maximum likelihood (ML) estimator takes the form

$$\hat{\underline{x}} = \arg\min_{\underline{x}} \|\underline{y} - \underline{x}\|^2 + \mu\Phi(\underline{x}), \tag{24}$$

where $\Phi(\underline{x})$ is the regularization functional, which stabilizes the estimator. In Besov norm regularized estimation, the regularization functional $\Phi(\underline{x})$ is set to the Besov norm $\|\underline{x}\|_B$ of the image $\underline{x}$.[1] Since the Besov norm is equivalent to the negative log likelihood of the image $\underline{x}$, Besov regularization penalizes "unlikely" estimates according to the independent GGD image model. Thanks to the interpretation of the Besov norm as a complexity measure (Shannon codelength) under the iid GGD image model, this programme is equivalent to complexity-regularized denoising[13] and MDL type signal estimation.[12]

Considering that minimizing the Besov norm is equivalent to maximizing the signal likelihood under the iid GGD model, the minimum Besov norm signal interpolation of Choi et al[2] reduces to simply finding the interpolant signal that passes through the sample points having maximum likelihood.

In the Besov ball projection denoising algorithm,[3] we project the noisy observation onto the set of images having high likelihood (low complexity) under the iid GGD image model. This approach is similar to the confidence tube image denoising algorithm of Ishwar et al.[19]

5. LOCAL SMOOTHNESS CHARACTERIZATION VIA A LOCALLY GAUSSIAN MODEL

Unfortunately, Besov spaces and their corresponding statistical models have serious limitations. By developing more accurate models, we can more tightly characterize natural images. In this section, we investigate a locally Gaussian model inspired by the EQ image coder[14,15] that locally adapts the distribution at each wavelet coefficient.

5.1. Limitations of the iid GGD model

The iid GGD model considered in Section 2.3 captures the general properties of natural images. However, it is far from perfect in that it lacks any description of local characteristics. This is equivalent to saying that the Besov norm lacks spatial localization of image smoothness. Furthermore, any iid model ignores the dependencies between wavelet coefficients at different scales, as pointed out in Choi et al.[3] This is equivalent to the "shuffle invariance" of the Besov norm — any shuffling of wavelet coefficients within each scale does not affect the Besov norm.

Most state-of-the-art image compression algorithms adapt to the local statistics of the wavelet coefficients. Fortunately, the normalized likelihood function (14) and the Shannon codelength interpretation of Besov norm (21) are very general, and they can be applied to other statistical models without modification. Thus, using more accurate density models for wavelet coefficients, we can measure image smoothness and complexity more precisely.

In Choi et al,[3] we indicated that the wavelet-domain hidden Markov tree model[5] captures the statistics of wavelet coefficients more accurately, relieving the localization and shuffle invariance problems of iid models mentioned above. However, the set of nonzero normalized likelihood defined using the hidden Markov tree model is no longer a linear space as mentioned in Choi et al,[3] and the normalized likelihood function does not correspond to a norm.[*] Even worse, the likelihood function for a wavelet-domain hidden Markov model has many local extrema that make the model less useful for defining a mathematical body similar to the Besov ball.

5.2. Local Gaussian model

The task of accurately specifying the image set boils down to accurately estimating the distribution of the image wavelet coefficients. This problem is central to any image compression algorithm. A typical wavelet-domain image compression consists of two main steps. First, the pdf of each wavelet coefficient is estimated as accurately as possible. (This step corresponds to confining the given image in a Besov-like set.) Then, the coefficient is quantized according to that pdf. (This corresponds to specifying the location of the given image in the set.) For efficient compression, we must make the confining set as small as possible to reduce the number of code bits needed to specify location in the set. For example, using the iid GGD model as the underlying density model for image compression, we need many more bits to code an image compared to other state-of-the-art image compression algorithms, because the confining set is too large. This sloppy image set results from the model inaccuracy of the iid GGD model.

The leading wavelet-domain image compression algorithms employ some kind of spatially adaptive pdf estimation procedure (often hidden in the compression algorithm). To minimize the number of bits to describe the pdf prediction algorithm itself, a very simple parametric form for the pdf functions, such as a zero-mean GGD or Laplacian,[†] is typically assumed. In this case, pdf estimation reduces to a variance (energy) estimation for each coefficient.

A representative image compression algorithm following the pdf estimation strategy is the EQ image coding algorithm.[14,15] EQ assumes a GGD distribution for each wavelet coefficient and estimates the variance of each pdf using the average energy in a small neighborhood around the coefficient. For compression, the window takes the form of a causal neighborhood following the proper scanning order of the coefficients both within and across scale. An image denoising algorithm using the methods similar to EQ image coder was recently proposed.[20] In this algorithm, each coefficient was modeled as a zero-mean Gaussian random variable, and its variance was estimated based on a local square window. The size of the window used for variance estimation was determined by a bootstrap method.

A natural question arises: What is the most natural parametric pdf for modeling the statistics of the wavelet coefficients. Mallat[10] showed that the histogram of wavelet coefficients w_i in each subband approximates a GGD with shape parameter strictly less than 2 (see Figure 1(a), for example). Since then, many algorithms have been based on GGD or Laplacian distributions.[11,14,15,21] Other algorithms have used Gaussian mixture models.[4,5,22,23] Heavy tailed models (heavier than the Gaussian) also seem natural to model the large wavelet coefficients corresponding to image singularities.

While these arguments are valid when the wavelet coefficients are considered identically distributed within scale, it is not clear what to do when we allow the within-scale statistics to be space-varying. For example, Figure 1(b) shows the histogram of the same coefficients as in Figure 1(a) but after normalizing each by the estimated local standard deviation of each coefficient (as in the EQ algorithm). That is, we plot $w_i/\widehat{\sigma}_i$, with the variance estimate

[*] Not to say that the set of natural images should be a linear normed vector space. However, characterization of such a set is beyond the scope of this paper.

[†] The zero-mean property follows from the fact that the wavelet coefficients oscillate positive and negative around edges.

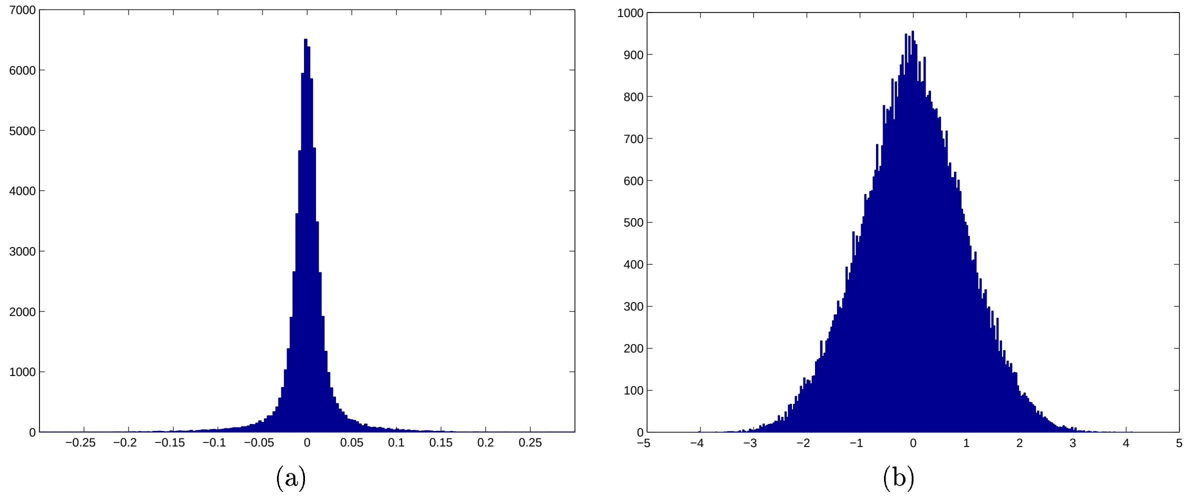

Figure 1. *(a) Histogram of wavelet coefficients from one subband of the Lena image. The distribution is close to a generalized Gaussian distribution with shape parameter less than 2. (b) Histogram after local variance normalization. The variance was estimated as the average energy of wavelet coefficients in* 3×3 *window around each coefficient. The distribution is close to a Gaussian.*

$\widehat{\sigma}_i^2$ the average energy of a 3×3 window of coefficients centered on coefficient i. This histogram is very close to a Gaussian distribution. The success of the denoising algorithm based on local Gaussian modeling[20] indicates that this is an accurate model. In fact, LoPresto et al[15] argued that a GGD is more appropriate than a Gaussian distribution in the EQ image compression algorithm only because in the compression case the variance must be estimated from *quantized* coefficients. If we are allowed to estimate the variance from the original coefficients, then a Gaussian seems a better choice.

Based on the above considerations, we propose to locally fit a zero-mean Gaussian model to each wavelet coefficients, with the variance estimated as in Mihcak et al.[20] In a simplified scheme, we can fix the size of variance estimation window (use a 3×3 window, for example) with minimal impact on performance.

Although the locally fit Gaussian model assumes that each coefficient is *independent*, the local variance estimation procedure implicitly captures the dependencies between coefficients. Thus, the model more accurately describes the wavelet coefficient statistics than the iid GGD model. Since each coefficient is allowed to have a different variance, the overall statistics of an entire subband of coefficients will be a Gaussian mixture distribution that approximates a GGD.

More specifically, we will model each coefficient $w_{j,\mathbf{k}}$ as a zero-mean Gaussian distribution

$$w_{j,\mathbf{k}} \sim \mathcal{N}(0, \sigma_{j,\mathbf{k}}^2) \tag{25}$$

with variance $\sigma_{j,\mathbf{k}}^2$ computed as

$$\sigma_{j,\mathbf{k}}^2 = \frac{1}{L} \sum_{\mathbf{i}} w_{j,\mathbf{i}}^2 \tag{26}$$

and the summation taken over all L wavelet coefficients inside the local window centered at $\mathbf{k}$.

5.3. Local smoothness characterization

Using the locally Gaussian model (25)–(26), the likelihood computation and definition of the normalized likelihood and Besov ball are straightforward. The likelihood of the entire set of wavelet coefficients is given by the product

$$f(\boldsymbol{w}) = \prod_{j,\mathbf{k}} g(w_{j,\mathbf{k}}), \tag{27}$$

where $g(w) = \frac{1}{\sqrt{2\pi\sigma_{j,\mathbf{k}}^2}} \exp(-\frac{w^2}{2\sigma_{j,\mathbf{k}}^2})$ is the zero-mean Gaussian density function with variance $\sigma_{j,\mathbf{k}}^2$.

From (27) the negative normalized log likelihood becomes

$$-\log f^N(w) = \sum_{j,\mathbf{k}} \frac{w_{j,\mathbf{k}}^2}{2\sigma_{j,\mathbf{k}}^2}, \tag{28}$$

a weighted l_2 norm of the wavelet coefficients.[‡] This is similar to the Besov norm (11), which is a weighted l_p norm. However, in (28) the weighting is more flexible in that each coefficient is weighted differently, while in the Besov norm each coefficient receives a fixed exponential weighting across scale. As a notation for the norm defined using independent Gaussian model, we use $\|\cdot\|_{IG}$. That is,

$$\|z\|_{IG}^2 \equiv \sum_{j,\mathbf{k}} \frac{w_{j,\mathbf{k}}^2}{2\sigma_{j,\mathbf{k}}^2}, \tag{29}$$

where the $w_{j,\mathbf{k}}$'s are the wavelet coefficients of the image z, and the summation contains all wavelet coefficients. Note that while this measure is a valid norm, it is *not* shuffle invariant.

For the independent Gaussian model, we can define a set similar to a Besov ball (with radius R) for finite sampled images as

$$B^{IG}(R) \equiv \{z : \|z\|_{IG} \leq R\} = \left\{ z : \sum_{j,\mathbf{k}} \frac{w_{j,\mathbf{k}}^2}{2\sigma_{j,\mathbf{k}}^2} \leq R \right\}. \tag{30}$$

We call $B^{IG}(R)$ a *local Gaussian ball* with radius R. Since $B^{IG}(R)$ is a convex set, when we have a noisy image not lying in $B^{IG}(R)$ with wavelet coefficients $\{\widetilde{w}_{j,\mathbf{k}}\}$, the l_2 projection onto $B^{IG}(R)$ is given by solving the constrained minimization

$$\text{minimize} \sum_{j,\mathbf{k}} (w_{j,\mathbf{k}} - \widetilde{w}_{j,\mathbf{k}})^2 \qquad \text{subject to} \qquad \sum_{j,\mathbf{k}} \frac{w_{j,\mathbf{k}}^2}{2\sigma_{j,\mathbf{k}}^2} \leq R. \tag{31}$$

The minimizing solution can be easily found using Lagrange multipliers. Let

$$C = \sum_{j,\mathbf{k}} (w_{j,\mathbf{k}} - \widetilde{w}_{j,\mathbf{k}})^2 + 2\lambda \left(\sum_{j,\mathbf{k}} \frac{w_{j,\mathbf{k}}^2}{2\sigma_{j,\mathbf{k}}^2} - R \right). \tag{32}$$

Computing the derivative with respect to $w_{j,\mathbf{k}}$ and setting to zero, we obtain the projected coefficients $\widehat{w}_{j,\mathbf{k}}$

$$\widehat{w}_{j,\mathbf{k}} = \frac{\sigma_{j,\mathbf{k}}^2}{\sigma_{j,\mathbf{k}}^2 + \lambda} \widetilde{w}_{j,\mathbf{k}}, \tag{33}$$

where λ should be determined to satisfy the constraint $\sum_{j,\mathbf{k}} \frac{w_{j,\mathbf{k}}^2}{2\sigma_{j,\mathbf{k}}^2} = R$.

Note that (33) has exactly the same form as the Wiener filter for estimating a Gaussian random variable corrupted by additive Gaussian noise. When the variance of the corrupting Gaussian noise is σ_n^2, we choose R so that $\lambda = \sigma_n^2$ for optimal performance in the minimum mean square error sense. Thus, the denoising algorithm resulting from the l_2 projection onto a local Gaussian ball is the same as the wavelet-domain Wiener filtering proposed by Mihcak et al.[20] It has also the same form as the wavelet-domain Wiener filtering considered in Ghael et al[24] and analyzed in Choi et al,[25] although in this case the variance of each wavelet coefficient is estimated by using two different wavelet transforms.

6. CONCLUSIONS

In this paper, we have shown that the Besov norm has a natural interpretation as the Shannon codelength under an independent GGD model for the wavelet coefficients. This interpretation unifies many seemingly different wavelet processing algorithms, such as Besov regularization,[1] MDL,[12] and complexity regularization.[13] To overcome the

[‡] Comparing with the Sobolev norm of wavelet coefficients ($p = 2$ in (11)) , which is an exponentially weighted l_2 norm, we can interpret (28) as a locally adapted Sobolev norm.

shuffle invariance limitation of Besov norms, we extended our recent results[3] by considering a more accurate wavelet-domain image model, namely that underlying the EQ compression algorithm.[14,15] This locally Gaussian model leads to a new definition of a wavelet-domain norm extending the Besov norm. A local Gaussian ball and projection algorithm lead to wavelet-domain Wiener filtering. Further development of "local Besov spaces" and corresponding image processing algorithms seem worthwhile for natural image modeling.

REFERENCES

1. A. Chambolle, R. A. DeVore, N. Lee, and B. J. Lucier, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. on Image Proc.* **7**, pp. 319–355, July 1998.
2. H. Choi and R. Baraniuk, "Interpolation and denoising of nonuniformly sampled data using wavelet domain processing," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Proc. — ICASSP '99*, (Phoenix, AZ), March 1999.
3. H. Choi and R. Baraniuk, "Wavelet-domain statistical models and Besov spaces," in *Proc. of SPIE Conf. Wavelet Applications in Signal Proc. VII*, vol. 3813, pp. 489–501, (Denver), July 1999.
4. F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *J. Roy Stat. Soc. Ser. B* **60**, pp. 725–749, 1998.
5. M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Proc.* **46**, pp. 886–902, April 1998.
6. R. A. DeVore, "Nonlinear approximation," in *Acta Numerica*, pp. 51–150, 1998.
7. D. Donoho and I. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika* **81**, pp. 425–455, 1994.
8. R. A. DeVore, B. Jawerth, and B. J. Lucier, "Image compression through wavelet transform coding," *IEEE Trans. on Information Theory* **38**, pp. 719–746, March 1992.
9. A. Cohen, I. Daubechies, O. Guleryuz, and M. Orchard, "On the importance of combining wavelet-based non-linear approximation with coding strategies," *Preprint* , 2000.
10. S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, pp. 674–693, July 1989.
11. P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized-Gaussian priors," in *Proc. IEEE-SP Int. Symp. Time-Freq. and Time-Scale Anal.*, pp. 633–636, (Pittsburgh, PA), Oct. 6-9 1998.
12. N. Saito, "Simultaneous noise supression and signal compression using a library of orthonormal bases and the mdl criterion," in *Wavelets in Geophysics*, pp. 299–324, New York: Academic Press, 1994. Editors: E. Foufoula-Georgiou and P. Kumar.
13. J. Liu and P. Moulin, "Complexity-regularized image denoising," *Preprint* , 2000.
14. S. LoPresto, K. Ramchandran, and M. T. Orchard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework," in *Data Compression Conference '97*, pp. 221–230, (Snowbird, Utah), 1997.
15. S. LoPresto, K. Ramchandran, and M. T. Orchard, "Wavelet image coding via rate-distortion optimized adaptive classification," in *Proc. of NJIT Symposium on Wavelet, Subband and Block Transforms in Communications, New Jersey Institute of Technology*, 1997.
16. I. Daubechies, *Ten Lectures on Wavelets*, SIAM, New York, 1992.
17. S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, 1998.
18. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., New York, 1991.
19. P. Ishwar, K. Ratakonda, P. Moulin, and N. Ahuja, "Image denoising using multiple compaction domains," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc. — ICASSP '98*, pp. 1889–1892, (Seattle, WA), May 1998.
20. M. K. Mihcak, I. Kozintsev, and K. Ramchandran, "Spatially adaptice statistical modeling of wavelet image coefficients and its application to denoising," in *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Proc. — ICASSP '99*, (Phoenix, AZ), March 1999.
21. Y. Yoo, A. Ortega, and B. Yu, "Image subband coding using context-based classification and adaptive quantization," *IEEE Trans. Image Proc.* **8**(12), pp. 1702–1715, 1999.
22. J. Pesquet, H. Krim, and E. Hamman, "Bayesian approach to best basis selection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc. — ICASSP '96*, pp. 2634–2637, (Atlanta, GA), 1996.

23. H. Chipman, E. Kolaczyk, and R. McCulloch, "Adaptive Bayesian wavelet shrinkage," *Amer. Stat. Assoc.* **92**, 1997.
24. S. P. Ghael, A. M. Sayeed, and R. G. Baraniuk, "Improved wavelet denoising via empirical Wiener filtering," in *Proceedings of SPIE*, vol. 3169, pp. 389–399, (San Diego), July 1997.
25. H. Choi and R. Baraniuk, "Analysis of wavelet-domain Wiener filters," in *Proc. IEEE-SP Int. Symp. on Time-Freq. and Time-Scale Anal.*, pp. 613–616, (Pittsburgh, PA), Oct. 6-9 1998.