Shift-Invariant Denoising using Wavelet-Domain Hidden Markov Trees

Justin K. Romberg, Hyeokho Choi and Richard G. Baraniuk *
Rice University
Dept. of Electrical and Computer Engineering
Houston, TX 77251–1892, USA
jrom@rice.edu, choi@ece.rice.edu, richb@rice.edu
Web: www.dsp.rice.edu

Abstract

Wavelet-domain hidden Markov models have proven to be useful tools for statistical signal and image processing. The hidden Markov tree (HMT) model captures the key features of the joint statistics of the wavelet coefficients of real-world data. One potential drawback to the HMT framework is the need for computationally expensive iterative training (using the EM algorithm, for example). In this paper, we use image structure not yet recognized by the HMT to show that the HMT parameters of real-world, grayscale images have a certain form. This leads to a description of the HMT model with just nine meta-parameters (independent of the size of the image and the number of wavelet scales). We also observe that these nine meta-parameters are similar for many images. This leads to a universal HMT (uHMT) *model for grayscale images. Algorithms using the uHMT require no training of any kind. While simple, a series of image estimation/denoising experiments show that the uHMT retains nearly all of the key structures modeled by the full HMT. Based on the uHMT model, we develop a shift-invariant wavelet denoising scheme that outperforms all algorithms in the current literature.*

1. Introduction

Statistical image processing problems, such as estimation, detection, and classification, rely on knowledge of the joint probability density function (pdf), $f(\boldsymbol{x})$, of the image $\boldsymbol{x}$. Since $f(\boldsymbol{x})$ is usually not known or is too complex to specify exactly, *models* that accurately approximate $f(\boldsymbol{x})$ are critical to image processing algorithms.

There have been several approaches to modeling the local joint statistics of image pixels in the spatial domain, the Markov random field model [4] being the most prevalent. However, spatial-domain models are limited in their ability to describe large-scale behavior. Markov random field models can be improved by incorporating a larger neighborhood of pixels, but this rapidly increases their complexity.

Transform-domain models are based on the idea that often a linear, invertible transform will "restructure" the image, leaving transform coefficients whose structure is simpler to model. Real-world images are well characterized by their *singularity* (edge and ridge) structure. The wavelet transform captures this singularity structure, and provides a natural and powerful domain for image modeling [6]. We aim, therefore, to model the joint pdf of the wavelet coefficients $\mathbf{w}$ of an image, and design wavelet-domain processing algorithms based on this model of $f(\mathbf{w})$.

2. Images in the Wavelet Domain

The wavelet transform is an atomic decomposition of an image with basis functions that are shifted and dilated versions of an oscillating mother wavelet [6]. The *primary properties* of wavelet transforms make wavelet-domain statistical image processing attractive [6, 8]:

- **P1. Locality:** Each wavelet coefficient represents the image content localized in spatial location and frequency.
- **P2. Multiresolution:** The wavelet transform represents the image at a nested set of scales.
- **P3. Edge Detection:** Wavelets act as local edge detectors. The edges in the image are represented by large wavelet coefficients at the corresponding spatial locations.
- **P4. Decorrelation:** The wavelet coefficients of real-world images tend to be approximately decorrelated.
- **P5. Energy Compaction:** The wavelet transforms of real-world images tend to be sparse. A wavelet co-

* This work was supported by the National Science Foundation, grant no. MIP–9457438, DARPA, grant no. DARPA/AFOSR F49620–97–1–0513, ONR, grant no. N00014–99–1–0813, Texas Instruments, and the Rice Consortium for Computational Seismic/Signal Interpretation.

efficient is large only if edges are present within the support of the wavelet.

Properties **P1** and **P2** lead to a natural arrangement of the wavelet coefficients in a quadtree structure with three subbands representing the horizontal, vertical, and diagonal edges in the image (see Fig. 1). The Compaction property (**P5**) follows from the fact that the edges constitute only a very small portion of a typical image; consequently, we can closely approximate an image by just a few (large) wavelet coefficients. Furthermore, the Decorrelation property (**P4**) indicates that the dependencies between wavelet coefficients are predominantly local. The primary properties give wavelet transforms significant structure, which we codify in the following *secondary properties*:

S1. NonGaussianity: The wavelet coefficients have peaky, heavy-tailed marginal distributions.

S2. Persistency: Large/small values of wavelet coefficients tend to propagate through the scales of the quadtrees.

NonGaussianity follows immediately from Energy Compaction (**P5**). Persistency follows from the Edge Detection (**P3**) and Multiresolution (**P2**) properties.

In general, the wavelet coefficients $\mathbf{w}$ are indexed by two integers: one for the scale (dilation), and one for the shift. In this paper, we will adopt an abstract indexing system and use only one integer whose value ranges from 1 to N^2 for an $N \times N$ image.

3. Hidden Markov Tree Models

The secondary properties of wavelet transforms give rise to joint wavelet statistics that are succinctly captured by the *wavelet-domain hidden Markov tree (HMT) model*, introduced by Crouse et al. (see [3] for a more detailed discussion).

The HMT models the nonGaussian marginal pdf $f(w_i)$ (**S1**) as a Gaussian mixture whose components are labeled by a *hidden state* $S_i \in \mathsf{S}, \mathsf{L}$. The S_i dictate from which of the two components in the mixture model w_i is drawn, and thus characterize (in the statistical sense) the magnitude of w_i. State S corresponds to a zero-mean, low-variance Gaussian, while state L corresponds to a zero-mean, high-variance Gaussian. If we let

$$g(x;\mu,\sigma^2) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{1}$$

denote the Gaussian pdf, then we can write

$$\begin{aligned} f(w_i|S_i=\mathsf{S}) &:= g(w_i;0,\sigma^2_{\mathsf{S};i}), & (2)\\ f(w_i|S_i=\mathsf{L}) &:= g(w_i;0,\sigma^2_{\mathsf{L};i}) & (3)\end{aligned}$$

with $\sigma^2_{\mathsf{L}} > \sigma^2_{\mathsf{S}}$. The marginal pdf $f(w_i)$ is the convex combination of the conditional densities

$$f(w_i) = p_i^{\mathsf{S}} g(w_i;0,\sigma^2_{\mathsf{S};i}) + p_i^{\mathsf{L}} g(w_i;0,\sigma^2_{\mathsf{L};i}), \tag{4}$$

with $p_i^{\mathsf{S}} = 1 - p_i^{\mathsf{L}}$. The p_i^{S} and p_i^{L} can be interpreted as the probability that w_i is small or large (in the statistical sense), respectively.

The persistence of wavelet coefficient magnitudes across scale (**S2**) is modeled by linking the hidden states in a Markov tree. The resulting dependency graph has a quadtree topology that mirrors the quadtree topology of the wavelet coefficients, see Fig. 1(b). Each subband is represented with its own quadtree; this assumes that the subbands are statistically independent.

Each parent→child state-to-state link has a corresponding transition matrix that quantifies statistically the degree of persistence of large/small coefficients:

$$A_i := \begin{bmatrix} p_i^{\mathsf{S}\to\mathsf{S}} & p_i^{\mathsf{S}\to\mathsf{L}} \\ p_i^{\mathsf{L}\to\mathsf{S}} & p_i^{\mathsf{L}\to\mathsf{L}} \end{bmatrix} \tag{5}$$

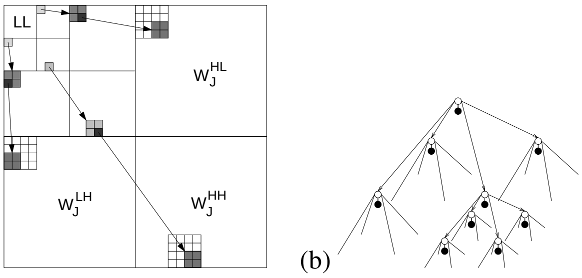
with $p_i^{\mathsf{S}\to\mathsf{L}} = 1 - p_i^{\mathsf{S}\to\mathsf{S}}$ and $p_i^{\mathsf{L}\to\mathsf{S}} = 1 - p_i^{\mathsf{L}\to\mathsf{L}}$.

Denote the parameters needed to specify a HMT model by the vector $\mathbf{\Theta}$. Members of $\mathbf{\Theta}$ are the mixture variances for each state, $\sigma_{\mathsf{S},i}$ and $\sigma_{\mathsf{L},i}$, the transition probabilities $p_i^{\mathsf{S}\to\mathsf{S}}$ and $p_i^{\mathsf{L}\to\mathsf{L}}$, and a mass function for the hidden state of the root node, p_0^{L}. These parameters can be fit to a given set of training data using the Expectation-Maximization (EM) algorithm [3]. The training yields an approximate maximum likelihood estimate of the model parameters given the training data, yielding a good approximation of the joint density function $f(\mathbf{w})$ of the wavelet coefficients and thus $f(\boldsymbol{x})$.

In general, the HMT model for an $N \times N$ image has approximately $4n$ parameters, with $n := N^2$. In some applications, this large number of parameters could make the HMT model cumbersome. To accurately specify $4n$ parameters for an n-pixel image requires significant a priori information about the image. If this information is unavailable, we run the risk of over-fitting the model. Crouse et al. [3] reduces the total number of HMT parameters to approximately $4L$, with L the number of wavelet scales (typically 4–10), by tying within scale. While a significant reduction, a large quantity of a priori image information is still required to specify the parameters without over-fitting. In the next section, we will take advantage of additional image structure to reduce the number of parameters it takes to specify the HMT even further.

4. Reduced-Parameter HMT Image Models

By design, the HMT model captures the main statistical features of images in the wavelet domain. In this section, we show that the HMT parameters themselves are of a certain



(a) (b)

Figure 1. (a) Quadtree organization of the wavelet coefficients. The four children wavelet coefficients divide the spatial localization of the parent coefficient. (b) 2-D HMT model. Each black node is a wavelet coefficient; each white node is the corresponding hidden state. Links represent dependencies between states.

form; a form motivated by image structure not yet incorporated into the HMT model. This form is specified by nine *meta-parameters*, leading to a simpler, reduced-parameter model.

Furthermore, we have found that many real-world images have similar HMT meta-parameters. By fixing one set of parameters, called the *universal HMT (uHMT)*, we can take a strictly Bayesian approach to the estimation problem, and eliminate the need for training altogether.

4.1. Tertiary Properties of the Wavelet Coefficients

The wavelet transforms of real-world images exhibit strong statistical properties in addition to the primary (**P1**–**P5**) and the secondary (**S1**,**S2**) properties. The following are related to the statistical self-similarity of images across scale:

- **T1. Exponential decay across scale:** The magnitudes of the wavelet coefficients of real-world images tend to decay exponentially across scale.
- **T2. Stronger persistence at finer scales:** The persistence of large/small wavelet coefficient magnitudes becomes stronger at finer scales.

The exponential decay property (**T1**) stems from the overall smoothness of images. Roughly speaking, a typical real-world image consists of smooth regions separated by a finite number of discontinuities. This results in a $1/f$-type spectral behavior, which leads to the exponential decay of the wavelet coefficients across scale [6].

We can obtain intuition behind property **T2** by considering the simple yet powerful image model of Cohen and D'Ales [1]. They model an image as piecewise smooth with a finite number of discontinuities. Consider a 1-D slice from such an image. Clearly it is also piecewise smooth with a finite number (say M) of discontinuities.

Since there are a finite number of discontinuities and the spatial resolution of the wavelet coefficients becomes finer as j increases (**P2**), there is some j_{crit} such that for all $j \geq j_{\mathrm{crit}}$, each wavelet basis function has at most one discontinuity inside its spatial support. We call this condition *isolation of the edges*. Given no a priori information about the locations of the discontinuities, the fact that the spatial resolutions of the wavelet coefficients become finer exponentially implies that the probability that every edge is isolated goes to 1 exponentially.

By **P4**, for fine scales such that $j \gg j_{\mathrm{crit}}$ there will be on the order of M wavelet coefficients that are "large" when compared to other coefficients at the same scale (exactly M if we are using the Haar wavelet). Each of these large coefficients will also have a large child, since the children wavelet basis functions simply divide up the spatial support of the parent. Each of the small coefficients' children will have small children, since there is no chance for any of them to encounter an edge.

In 2-D, the situation is similar except that instead of a discontinuities at points, we now have discontinuities along curves. At j_{crit}, all wavelet basis functions that have spatial support intersecting this curve will be "large." Again, each of these coefficients will also have at least one large child, while the small coefficients will spawn only small children.

4.2. The HMT Meta-Parameters

Based on the tertiary properties of the wavelet transforms of real-world images, we can specify the HMT model parameters in terms of nine meta-parameters. The coefficient decay and the change in coefficient persistence are easily modeled by imposing structure on how the mixture variances and state transition probabilities change across scale. Because the tertiary properties are common to many real-world images, the resulting model describes the common overall behavior of real-world images in the wavelet domain.

We can easily model the exponential decay of wavelet coefficients (**T1**) through the mixture variances of the HMT model. Since the HMT mixture variances characterize the magnitudes of the wavelet coefficients, we will require that they decay exponentially across scale:

$$\sigma^2_{\mathsf{S};j} = C_{\sigma_{\mathsf{S}}} 2^{-j\alpha_{\mathsf{S}}}, \quad (6)$$

$$\sigma^2_{\mathsf{L};j} = C_{\sigma_{\mathsf{L}}} 2^{-j\alpha_{\mathsf{L}}}. \quad (7)$$

To have $\sigma^2_{\mathsf{S};j} < \sigma^2_{\mathsf{L};j}$ for all scales, we require $\alpha_{\mathsf{S}} \geq \alpha_{\mathsf{L}}$. The result is an HMT for images with $1/f$ power spectra.

We will model the change in the degree of coefficient magnitude persistency by considering the way that the state transition probabilities change across scale.

Again, consider a 1-D signal consisting of smooth regions having M jump discontinuities. The isolation of edges at fine scales controls the persistency and novelty probabilities (and hence the form of the transition matrix) in the HMT. If each of the M edges in the 1-D slice is isolated then there is no opportunity for a novel large coefficient to come from a small parent; the only way a coefficient can be large is if its parent is large. Thus, $p_j^{\mathsf{S}\to\mathsf{L}} \to 0$ exponentially as $j \to \infty$. In other words, $p_j^{\mathsf{S}\to\mathsf{S}} \to 1$, since once a wavelet basis function lies over a smooth region, all of its children also lie over that smooth region. If a basis function lies over an edge, one and only one of its children will lie over the edge. This is an exact statement for the Haar basis functions, and a close approximation for longer wavelets. Therefore, the large wavelet coefficient gives rise to one large and one small wavelet coefficient and $p_j^{\mathsf{L}\to\mathsf{L}} \to \frac{1}{2}$. For a more in-depth analysis, see [7]

The edge isolation probability going to 1 exponentially means that the asymptotic values for persistency and novelty parameters are also approached exponentially. This gives a state transition matrix (see (5)) specified by four parameters:

$$A_j = \begin{bmatrix} 1 - C_{\mathsf{SS}} 2^{-\gamma_{\mathsf{S}} j} & C_{\mathsf{SS}} 2^{-\gamma_{\mathsf{S}} j} \\ \frac{1}{2} - C_{\mathsf{LL}} 2^{-\gamma_{\mathsf{L}} j} & \frac{1}{2} + C_{\mathsf{LL}} 2^{-\gamma_{\mathsf{L}} j} \end{bmatrix}. \tag{8}$$

The only parameter in the HMT not yet accounted for is the probability mass function on the hidden state value of the root coefficients (just one number in our case, $p_{j_0}^{\mathsf{L}}$, since the hidden state can only take two different vales). Taking this parameter as is, we can describe the HMT model with just nine meta-parameters:
$\mathbf{\Theta}_{\mathsf{i}} = \left\{\alpha_{\mathsf{S}}, \alpha_{\mathsf{L}}, C\sigma_{\mathsf{S}}, C_{\sigma_{\mathsf{L}}}, \gamma_{\mathsf{S}}, \gamma_{\mathsf{L}}, C_{\mathsf{SS}}, C_{\mathsf{LL}}, p_{j_0}^{L}\right\}$.

4.3. A "universal" HMT: The uHMT

The HMT meta-parameters not only give us a simpler description of the HMT model, but we have also observed that they take very similar values for many grayscale, photograph-like images. We trained HMT models for a set of normalized images and examined their parameters. The variance and persistence decays were measured by fitting a line to the log of the variance vs. scale for each state. The decays were very similar for all of the images. Since the images were normalized, the range over which the variances decayed was similar as well. These observations lead us to believe that a specific, "universal" set of HMT meta-parameters can reasonably characterize photograph-like images. We call the HMT with this set of parameters the *uHMT* model.

The universal meta-parameters are given by: $\alpha_{\mathsf{S}} = 3.1$, $\alpha_{\mathsf{L}} = 2.25$, $C_{\sigma_{\mathsf{S}}} = 2^{11}$, $C_{\sigma_{\mathsf{L}}} = 2^{11}$, $\gamma_{\mathsf{S}} = 1$, $\gamma_{\mathsf{L}} = 0.4$, $C_{\mathsf{SS}} = 2^{2.3}$, $C_{\mathsf{LL}} = 2^{0.5}$, and $p_0^{\mathsf{L}} = 1/2$.

The simplicity of the uHMT model also allows us to apply it in situations where the cost of a standard HMT would be prohibitive. For instance, we have developed a fast O($n \log n$) shift-invariant estimation scheme (discussed below in Section 5.2) based on the uHMT parameters that delivers state-of-the-art performance (see Fig. 3).

5. Applications to Image Denoising

5.1. Bayesian Estimation using the HMT

To demonstrate the effectiveness of the uHMT for modeling an image's wavelet coefficients, we estimate an image submerged in additive white Gaussian noise. Translated into the wavelet domain, the problem is as follows:

$$\text{given } \mathbf{y} = \mathbf{w} + \mathbf{n}, \text{ estimate } \mathbf{w}, \tag{9}$$

where $\mathbf{n}$ is a Gaussian random field whose components are independent and identically distributed with zero mean and known variance σ_n^2.

Since we are viewing $\mathbf{w}$ as a realization of a random field whose joint pdf is modeled by the HMT, we take a Bayesian approach to the estimation problem. The conditional density $f(\mathbf{y}|\mathbf{w})$ is given by the problem; it is an independent, Gaussian random field with mean $\mathbf{w}$. Using the HMT model for $f(\mathbf{w})$, we can solve the Bayes equation for the posterior $f(\mathbf{w}|\mathbf{y})$.

To obtain the model parameters, Crouse et al. takes an empirical Bayesian approach [3]. The HMT parameters used to model $f(\mathbf{w}|\mathbf{\Theta})$ are first estimated from the observed noisy data $\mathbf{y}$ and then "plugged-in" to the Bayes equation (after accounting for the noise).

For the Bayes estimator, we calculate the conditional mean of the posterior $f(\mathbf{w}|\mathbf{y}, \mathbf{\Theta})$ using the pointwise transformation

$$\widehat{w}_i = E[w_i|\mathbf{y}, \mathbf{\Theta}] = \sum_q p(S_i = q|\mathbf{y}, \mathbf{\Theta}) \frac{\sigma_{q;i}^2}{\sigma_n^2 + \sigma_{q;i}^2} y_i \tag{10}$$

to obtain the minimum mean-square estimate (MMSE) of $\mathbf{w}$. Results using the empirical Bayesian HMT estimator, summarized for images in [7], are competitive in both visual quality and PSNR to redundant wavelet shrinkage.

With the uHMT parameters, we have a prior on $\mathbf{w}$ and the estimation problem can be approached from a purely Bayesian standpoint. Since we have eliminated training, the estimation algorithm is truly $O(n)$ and takes only a few seconds to run on a workstation.

The results, given in [7], are almost identical to the more complicated empirical Bayes HMT approach, suggesting that we have lost almost nothing by totally eliminating training.

5.2. Shift-Invariant Estimation

Image estimates obtained using an orthogonal wavelet transform frequently exhibit visual artifacts, usually in the form of ringing around edges. These artifacts can be combatted by averaging together estimates obtained from all different shifts of the image [2]. The resulting shift-invariant

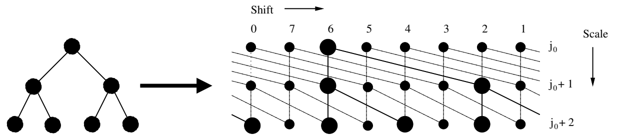


Figure 2. 1-D illustration of the wavelet tree at one shift (shift 6) in relation to the trees at other shifts. Note that the trees overlap — the same coefficient appears in more than one tree. There are $N \log N$ unique coefficients over all of the shifts of the length-N signal.

estimate is given by

$$\widehat{x} = \text{Average}(\mathbf{S}_{-k,-m}(\mathbf{D}(\mathbf{S}_{k,m}(y))))_{0 \le k,m \le N-1} \quad (11)$$

where $\mathbf{S}_{k,m}(y) = y(s-k, t-m)$ is the 2-D (circular) shift operator and $\mathbf{D}$ denotes the estimator (10). Implementing (11) directly would have computational complexity $O(n^2)$ and would thus be infeasible for large images. To streamline the algorithm, we must exploit the redundancies in the wavelet representations between different shifts of the image.

In the wavelet domain, each shift of the image corresponds to a different tree of wavelet coefficients. The wavelet coefficient trees for different shifts overlap, with common coefficients occupying entire subtrees (see 1-D example in Fig. 2). Averaging estimates for different shifts amounts to averaging the $p(S_i = q|\mathbf{y}, \mathbf{\Theta})$ for each tree in which w_i appears, and then using the result in (10) (we assume that $\mathbf{\Theta}$ is the same for each shift of the image). The way in which the wavelet coefficient trees of different shifts overlap allows an $O(n \log n)$ shift-invariant denoising algorithm [7]. The results (shown in Fig. 3(d) and summarized in [7]) indicate that this denoising algorithm defines the new state-of-the-art: in general, we gain a 1–1.5 dB gain over thresholding with the redundant wavelet transform [2, 5].

6. Conclusions

Hidden Markov Trees capture the primary aspects of image structure in the wavelet domain. In this paper, we have shown that additional image structure can be exploited by constraining the HMT parameters to have a certain form. As a result, an HMT model for real-world images can be described with 9 meta-parameters.

A set of "universal" parameters arises naturally from the form of the meta-parameters. These nine numbers completely specify a model for a large class of real-world images, eliminating any need for training in the estimation algorithm without compromising denoising performance. Having the model fully specified facilitates the implementation of a shift-invariant estimation algorithm which offers state-of-the-art performance.

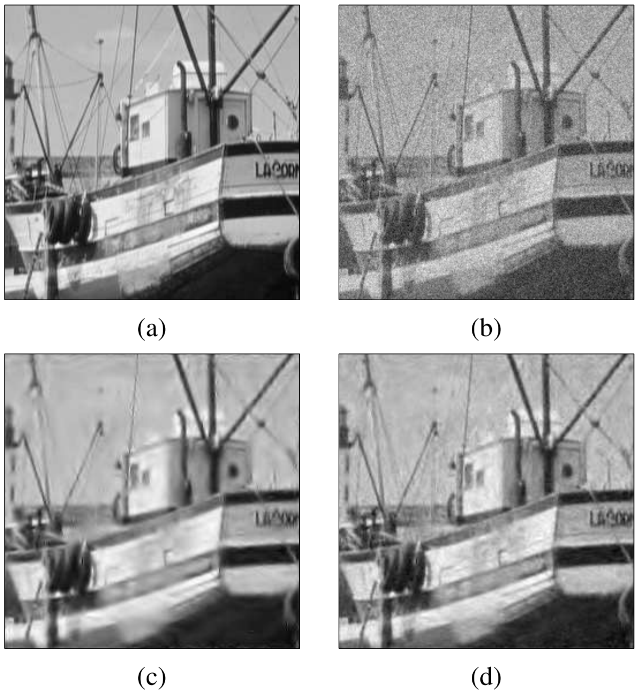


Figure 3. (a) Original 256×256 "Boats" image; (b) Noisy boats image, with $\sigma_n = 0.1$, PSNR=20dB. Boats image denoised using (c) redundant hard-thresholding using empirical best threshold [5], PSNR=26.3dB; (d) shift-invariant uHMT estimator, PSNR=27.4dB.

References

[1] A. Cohen and J. P. D'Ales. Nonlinear approximation of random functions. *SIAM J. Appl. Math.*, 57, April 1997.

[2] R. Coifman and D. Donoho. Translation-invariant denoising. In *Wavelets and Statistics*, Lecture Notes in Statistics. Springer-Verlag, 1995.

[3] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Proc.*, 46(4):886–902, April 1998.

[4] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. on Pattern Anal. Machine Intell.*, 6:721–741, 1984.

[5] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus, and R. O. Wells. Nonlinear processing of a shift invariant DWT for noise reduction. In *Proceedings of SPIE*, 1995.

[6] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 1998.

[7] J. K. Romberg, H. Choi, and R. G. Baraniuk. Bayesian tree-structured image modeling using wavelet-domain hidden Markov models. Preprint, available at www.dsp.rice.edu.

[8] M. Vetterli and J. Kovačević. *Wavlets and Subband Coding*. Prentice Hall, Englewood Cliffs:NJ, 1995.