Bayesian Tree-Structured Image Modeling using Wavelet-domain Hidden Markov Models

Justin K. Romberg, Hyeokho Choi and Richard G. Baraniuk

Department of Electrical and Computer Engineering, Rice University
6100 S. Main St., Houston, TX 77005, USA

Submitted to IEEE Transactions on Image Processing, March 2000
Revised October 2000, February 2001

EDICS Numbers: 2-MODL, 2-WAVP, 2-REST

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 probability density of the wavelet coefficients of real-world data. One potential drawback to the HMT framework is the need for computationally expensive iterative training to fit an HMT model to a given data set (using the Expectation-Maximization algorithm, for example). In this paper, we greatly simplify the HMT model by exploiting the inherent self-similarity of real-world images. This simplified model specifies the HMT parameters with just nine meta-parameters (independent of the size of the image and the number of wavelet scales). We also introduce a Bayesian *universal HMT (uHMT)* that fixes these nine parameters. The uHMT requires no training of any kind. While extremely simple, we show using a series of image estimation/denoising experiments that these new models retain nearly all of the key image structure modeled by the full HMT. Finally, we propose a fast shift-invariant HMT estimation algorithm that outperforms other wavelet-based estimators in the current literature, both visually and in mean-square error.

Contact author: Justin Romberg. Email: jrom@rice.edu, choi@ece.rice.edu, richb@rice.edu. Web: www.dsp.rice.edu. Fax: 713.348.6196. This work was supported by the National Science Foundation grants MIP-9457438 and CCR-9973188, DARPA/AFOSR grant F49620-97-1-0513, ONR grant N00014–99–1–0813, and the Texas Instruments Leadership University Program.

I. Introduction

In statistical image processing, we view an image $\boldsymbol{x}$ as a realization of a random field with joint probability density function (pdf) $f(\boldsymbol{x})$. Viewing $\boldsymbol{x}$ as random allows us to take a Bayesian approach to image processing: we can incorporate knowledge of an image's characteristics into $f(\boldsymbol{x})$. Solutions to problems such as estimation, detection, and compression rely on $f(\boldsymbol{x})$; the more accurately it can be specified, the better the solutions. Of course, we rarely have enough information to specify the joint pdf exactly. Our goal is to construct a realistic *model* that approximates $f(\boldsymbol{x})$ and allows efficient processing algorithms.

There have been several approaches to model the local joint statistics of image pixels in the spatial domain, the Markov random field model [1] being the most prevalent. However, spatial-domain models are limited in their ability to describe large-scale image 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. For such images, the wavelet transform provides a powerful domain for modeling [2].

The wavelet transform records the differences in the image at different scales (resolutions). As such, the portions of the image which do not vary significantly from scale to scale (the "smooth" regions) will be captured by a few large values at coarse scales. The portions of the image that do vary from scale to scale are typically regions around edges, and are represented by large values at each scale in the wavelet transform.

The following *primary properties* of the wavelet transform make wavelet-domain statistical image processing attractive [2,3]:

P1. Locality: Each wavelet coefficient represents image content local in space 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 locations.

Properties **P1** and **P2** lead to a natural arrangement of the wavelet coefficients into three subbands representing the horizontal, vertical, and diagonal edges. Each of these subbands has a *quad-tree* structure; regions of analysis in the image at one scale are divided up into four smaller regions at the next (finer) scale (see Fig. 1).

Properties **P1 - P3** induce two properties for the wavelet *coefficients* of real-world images:

P4. Energy Compaction: The wavelet transforms of real-world images tend to be sparse. A wavelet coefficient is large only if edges are present within the support of the wavelet.

P5. Decorrelation: The wavelet coefficients of real-world images tend to be approximately decorrelated.

The Compaction property **P4** follows intuitively from two observations[1]: that edges constitute only a very small portion of a typical image, and that a wavelet coefficient is large only if edges are present within the support of the wavelet. Consequently, we can closely approximate an image using just a few (large) wavelet coefficients. Finally, the Decorrelation property (**P5**) indicates that the dependencies between wavelet coefficients are predominantly local.

The primary properties give the wavelet coefficients of natural images significant statistical structure, which we codify in the following *secondary properties* [6]:

S1. NonGaussianity: The wavelet coefficients have peaky, heavy-tailed marginal distributions [7, 8].

S2. Persistency: Large/small values of wavelet coefficients tend to propagate through the scales of the quad-trees [9, 10].

NonGaussianity is simply a statistical restatement of Energy Compaction (**P4**). Persistency is a consequence of the Edge Detection (**P3**) and Multiresolution (**P2**) properties.

These secondary properties give rise to joint wavelet statistics that are succinctly captured by the wavelet-domain *hidden Markov tree (HMT) model* [6]. The HMT models the nonGaussian marginal pdf (**S1**) as a two-component Gaussian mixture. The components are labeled by a hidden state signifying whether the coefficient is small or large. The Gaussian component corresponding to the small state has a relatively small variance, capturing the peakiness around zero, while the component corresponding to the large state has a relatively large variance, capturing the heavy tails[2]. The persistence of wavelet coefficient magnitudes across scale (**S2**) is modeled by linking these hidden states across scale in a Markov tree (see Fig. 4 below). A state transition matrix for each link quantifies statistically the degree of persistence of large/small coefficients. Given

[1] While **P4** and **P5** can be made precise mathematically (see [4], for example), we present them here using intuitive arguments based on the nature of images.

[2] Of course, no Gaussian density has heavy tails in the strict sense. Here a Gaussian with a large variance captures the shape of the heavy-tailed density in the region where large values are likely.

a set of training data (usually in the form of one or more observed images), maximum likelihood estimates of the mixture variances and transition matrices can be calculated using the Expectation-Maximization (EM) algorithm [6]. These parameter estimates yield a good approximation of the joint density function $f(\mathbf{w})$ of the wavelet coefficients and thus $f(\boldsymbol{x})$.

In its most general form, the HMT model for an n-pixel ($N \times N$) image has approximately $4n$ parameters, making it too cumbersome for almost all applications. In [6], the number of parameters was reduced to approximately $4J$, with J the number of wavelet scales ($\sim \log_2 N$, typically 4–10), by assuming that the model parameters are the same at each scale. This reduction makes it feasible, but still computationally costly, to fit a model to one n-pixel training image.

In this paper, we leverage additional wavelet domain image structure not yet exploited by the HMT to obtain a reduced-parameter HMT model. This new model is constructed using two empirical *tertiary properties* of image wavelet coefficients. These tertiary properties reflect the *self-similar* nature of images and their resulting generalized $1/f$ spectral behavior [11, 12]:

T1. Exponential decay across scale: The magnitudes of the wavelet coefficients of real-world images decay exponentially across scale [2].

T2. Stronger persistence at fine scales: The persistence of large/small wavelet coefficient magnitudes becomes exponentially stronger at finer scales.

Using **T1** and **T2**, we will develop a reduced-parameter HMT model that is described with just 9 meta-parameters (independent of the size of the image and the number of wavelet scales). As an added bonus, we will observe that these 9 parameters take similar values for many real-world images, allowing us to fix a "universal" set of parameters, resulting in a *universal HMT (uHMT).* Using the uHMT model, the parameter values are completely determined, giving us a prior $f(\mathbf{w})$ for the wavelet transforms of real-world images. With the prior specified, we avoid the costly image specific training required with an empirical Bayesian approach (as in [6] and Section III-E), making HMT-based processing practical in more settings.

While the uHMT is certainly less specific in its modeling of a particular image, it captures the statistics of a broad class of real-world images sufficiently for many applications. Figure 2, which compares denoising results using algorithms based on the uHMT to other methods in the literature, demonstrates the effectiveness of the uHMT. We observe in Fig. 2 that the image estimation (denoising) performance of the uHMT model is extremely close to the more complicated HMT model. Furthermore, the simplicity of the uHMT model allows us to apply it in situations where the cost of the HMT would be prohibitive. For instance, we will develop an $O(n \log n)$ shift-invariant uHMT based estimation scheme in Section V below that offers state-of-the-art denoising performance, as seen from Fig. 2 and column 1 of Tables I–III.

In contrast to other hidden Markov model techniques in the literature, the uHMT is simple and easy to use. The uHMT offers the performance of a complicated model with the computational efficiency of a simple model. In [7], shrinkage rules are introduced using a two state independent Gaussian mixture model for the prior on the wavelet coefficients. A Generalized Gaussian distribution (GGD) with auto regressive dependencies between neighboring coefficients (both within and across scales) is used to model wavelet coefficients in [8]. In [14], maximum a posteriori estimation for GGD models and its equivalence to hard thresholding and MDL estimation is discussed. An independent two state mixture model, where the "low" state is a point mass at zero, is analyzed in [15] with relations between realizations of this model to functions in Besov spaces. In [16], the wavelet coefficients are modeled as Gaussian, with the variance estimated from neighbors at the same scale. Finally, a HMT model with parameters estimated from a noisy observation of an image is used in [6].

After reviewing the wavelet transform in Section II and the HMT model in Section III, we introduce the HMT meta-parameters and the uHMT in Section IV. Bayesian estimation with the HMT is reviewed in III-E, and revisited in IV-C with the uHMT. Section V develops the new redundant wavelet estimation technique. We close in Section VI with a discussion and conclusions.

II. Discrete Wavelet Transform

The 2-D discrete wavelet transform (DWT) represents an image $x(s) \in L^2(\mathbb{R}^2)$ in terms of a set of shifted and dilated wavelet functions $\{\psi^{LH}, \psi^{HL}, \psi^{HH}\}$ and scaling function ϕ^{LL} [17]. When these shifted and dilated functions form an orthonormal basis for $L^2(\mathbb{R}^2)$, the image can be decomposed as

$$x(s) = \sum_{k \in \mathbb{Z}^2} u_{j_0,k} \phi^{LL}_{j_0,k}(s) + \sum_{b \in \mathcal{B}} \sum_{j \geq j_0} \sum_{k \in \mathbb{Z}^2} w^b_{j,k} \psi^b_{j,k}(s) \tag{1}$$

with $\phi^{LL}_{j_0,k} := 2^{j_0}\phi^{LL}(2^{j_0}s - k)$, $\psi^b_{j,k} := 2^j\psi^b(2^j s - k)$, and $b \in \mathcal{B} := \{LH, HL, HH\}$. The LH, HL, and HH denote the *subbands* of the wavelet decomposition. The expansion coefficients, called the *scaling coefficients* and *wavelet coefficients*, respectively, are given by

$$u_{j_0,k} := \int_{s \in \mathbb{R}^2} x(s)\phi_{j_0,k}(s)\; ds, \tag{2}$$

$$w^b_{j,k} := \int_{s \in \mathbb{R}^2} x(s)\psi^b_{j,k}(s)\; ds. \tag{3}$$

To keep the notation manageable, we will use an abstract index for the DWT coefficients and the basis functions, $w_{j,k}^b \to w_i$ and $\psi_{j,k}^b \to \psi_i$, unless the full notation is required.

In practice, the image will be discretized on an $N \times N$ grid. This imposes a maximal level of decomposition $J := \log_2 N > j \geq j_0$, with 4^{j-1} wavelet coefficients in each subband and 4^{j-1} scaling coefficients at each scale. The $n := N^2$ scaling and wavelet coefficients in (2)–(3) for an $N \times N$ discrete image can be calculated using a 2-D separable filter bank [18] in $O(n)$ computations.

A wavelet coefficient $w_{j,k}^b$ at a scale j represents information about the image in the spatial region around $2^{-j}k$ ($k \in \mathbb{Z}^{\not{2}}$) [2]. At the next finest scale $j+1$, information about this region is represented by four wavelet coefficients; we call these the *children* of $w_{j,k}^b$. This leads to a natural quad-tree structuring of each of the 3 subbands, as shown in Fig. 1 and Fig. 4(a) [19]. In light of this natural tree structure, we will often refer to the wavelet coefficients as a *DWT tree* with w_i as a *node* in the tree. We also denote $\rho(i)$ as the parent and $c(i)$ as the set of children of node i. As j increases, the child coefficients add finer and finer details into the spatial regions occupied by their ancestors [19].

The Haar wavelet basis functions at a given scale are disjoint square waves [17]. In this case, the spatial divisions made by the wavelet quadtrees are exact (see Fig. 4(a)). For longer wavelets, the supports of adjacent wavelets at a given scale overlap. However, the wavelet coefficients still represent information in the $2^{-j} \times 2^{-j}$ dyadic squares to a good approximation.

The orthogonal wavelet transform is not shift-invariant. In fact, the wavelet coefficients of two different shifts of an image can be very different [13], with no simple relationship between them. We will find it useful to analyze and process the wavelet coefficients for each shift of the image. The resulting representation is called the *redundant wavelet transform* (RDWT) [19]. The RDWT is overcomplete, with $n \log n$ wavelet and scaling coefficients for an n-pixel image.

III. Wavelet-domain Hidden Markov Tree Models

In the Introduction, we made the notion of real-world image wavelet-domain structure precise with the secondary properties **S1** and **S2**. The hidden Markov tree (HMT) model, introduced by Crouse et al. in [6] and reviewed in this section, captures these properties simply and accurately. To match the non-Gaussian nature of the wavelet coefficients (**S1**), the HMT models the marginal pdf of each coefficient as a Gaussian mixture density with a hidden state that dictates whether a coefficient is large or small. To capture the dependencies between the wavelet coefficients, the HMT uses a probabilistic tree to model Markovian dependencies between the hidden states. Using

S2 above, this graph connects each parent to its four children and has the same quad-tree topology as the DWT tree discussed in Section II.

A. Capturing NonGaussianity: Mixture Models

The form for the marginal distribution of a wavelet coefficient w_i comes directly from the efficiency of the wavelet transform in representing real-world images: a few wavelet coefficients are large, but most are small. Gaussian mixture modeling runs as follows. Associate with each wavelet coefficient w_i an unobserved *hidden state* variable $S_i \in \{\mathsf{S}, \mathsf{L}\}$. The value of S_i dictates which of the two components in the mixture model generates w_i. State S corresponds to a zero-mean, low-variance Gaussian. If we let

$$g(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{4}$$

denote the Gaussian pdf, then we can write

$$f(w_i | S_i = \mathsf{S}) = g(w_i; 0, \sigma^2_{\mathsf{S};i}). \tag{5}$$

State L, in turn, corresponds to a zero-mean, high-variance Gaussian

$$f(w_i | S_i = \mathsf{L}) = g(w_i; 0, \sigma^2_{\mathsf{L};i}) \tag{6}$$

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

$$f(w_i) = p_i^{\mathsf{S}}\, g(w_i; 0, \sigma^2_{\mathsf{S};i}) + p_i^{\mathsf{L}}\, g(w_i; 0, \sigma^2_{\mathsf{L};i}), \tag{7}$$

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

$$p_{S_i} = \begin{bmatrix} p_i^{\mathsf{S}} \\ p_i^{\mathsf{L}} \end{bmatrix}. \tag{8}$$

be the state value probability mass function for S_i. 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 independent Gaussian mixture model (IM) is parameterized by a $\{p_i^{\mathsf{L}},\, \sigma^2_{\mathsf{S};i},\, \sigma^2_{\mathsf{L};i}\}$ triad for each wavelet coefficient w_i.

Wavelet coefficients have often been modeled as realizations from a zero-mean Generalized Gaussian distribution (GGD) [8, 14]. In fact, the GGD models the marginal densities of the wavelet coefficients more accurately than the Gaussian mixture, shown in Fig. 3, especially in the tails of the distribution. However, the Gaussian mixture model discussed above can approximate the Generalized Gaussian density arbitrarily well by adding more and more hidden states. Of course, as the number of states in the model increases, the model becomes more computationally complex and less robust. As can be seen in Fig. 3, we are matching the marginal histogram very closely using only two states. We can think of this two-state mixture model as an approximation to the Generalized Gaussian model, and will see that it is realistic enough for our purposes. The primary advantage of the Gaussian mixture model, as we will see in the next section, is that it sets up a framework for conveniently modeling the dependencies between wavelet coefficients. Although independence is a reasonable first-order approximation to the structure of the wavelet coefficients, significant gains are realized by modeling the dependencies between coefficients.

B. Capturing Persistence: Markov Trees

Secondary property **S2** states that the relative magnitude of a wavelet coefficient is closely related to the magnitude of its parent. This implies a type of Markovian relationship between the wavelet states, with the probability of a wavelet coefficient being "large" or "small" affected only by the size of its parent. The HMT models the dependence as Markov-1: given the state of a wavelet coefficient S_i, the coefficient's ancestors and descendents are independent of each other.

The HMT captures these dependencies by using a probabilistic tree that connects the hidden state variable of each wavelet coefficient with the state variable of each of its children. This leads to the dependency graph having the same quad-tree topology as the wavelet coefficients (see Fig. 4). Each subband is represented with its own quad-tree; this assumes that the subbands are independent.

Each parent→child state-to-state link has a corresponding state transition matrix [3]

$$A_i = \begin{bmatrix} p_i^{\mathsf{S}\to\mathsf{S}} & p_i^{\mathsf{S}\to\mathsf{L}} \\ p_i^{\mathsf{L}\to\mathsf{S}} & p_i^{\mathsf{L}\to\mathsf{L}} \end{bmatrix} \tag{9}$$

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

The parameters $p_i^{\mathsf{S}\to\mathsf{S}}$ ($p_i^{\mathsf{L}\to\mathsf{L}}$) can be read as "the probability that wavelet coefficient w_i is small (large) given that its parent is small (large)". We call these the *persistency probabilities.* We call $p_i^{\mathsf{L}\to\mathsf{S}}$ and $p_i^{\mathsf{S}\to\mathsf{L}}$ the *novelty probabilities*, for they give the probabilities that the state values will change from one scale to the next. Having large and small wavelet coefficient values propagate down the quad-tree (recall **S2**) requires more persistence than novelty, that is, $p_i^{\mathsf{S}\to\mathsf{S}} > p_i^{\mathsf{S}\to\mathsf{L}}$ and $p_i^{\mathsf{L}\to\mathsf{L}} > p_i^{\mathsf{L}\to\mathsf{S}}$.

The idea of persistency follows from our interpretation of the wavelet basis functions as local edge detectors. If there is an edge inside the spatial support of the basis function, then the

[3] This state transition matrix is the transpose of that presented in [20].

corresponding wavelet coefficient tends to be large (in magnitude). Since the same edge is within the spatial support of at least one of the children basis functions, we have large values propagating down through scale. If, however, there are two edges inside the spatial support of a wavelet basis function, then their effects can cancel out, making the corresponding wavelet coefficient small. At some fine scale down the tree, however, the two edges are guaranteed to bifurcate, since the spatial resolution will be fine enough so that each edge is represented by its own (large) wavelet coefficient [9]. These wavelet coefficients will be large even though their parent is small. This is the idea behind novelty.

C. HMT Parameters

An HMT model is specified in terms of:

1. the mixture variances $\sigma^2_{\mathsf{S};i}$ and $\sigma^2_{\mathsf{L};i}$,
2. the state transition matrices A_i,
3. the probability of a large state at the root node for each i in the coarsest scale p_i^{L}.

Grouping these into a vector $\mathbf{\Theta}$, the HMT provides a parametric model for the joint probability density function $f(\mathbf{w}|\mathbf{\Theta})$ of the wavelet coefficients in each of the 3 subbands (we treat the subbands as independent [21]).

In general, the variance and transition parameters can be different for each wavelet coefficient. However, this makes the model too complicated for some applications. For example, if there is only one observation of an n-pixel image, then we are faced with the impossible task of fitting $4n$ parameters to n data points. To reduce the HMT complexity, we can make the simplifying assumption that each parameter is the same at each scale of the wavelet transform:

$$\left.\begin{aligned} \sigma^2_{\mathsf{S};b,j,k} &= \sigma^2_{\mathsf{S};j} \\ \sigma^2_{\mathsf{L};b,j,k} &= \sigma^2_{\mathsf{L};j} \\ A_{b,j,k} &= A_j \end{aligned}\right\} \quad \forall k \in \mathbb{Z}^{\nvDash},\ \forall \in \mathcal{B}. \tag{10}$$

This processes is referred to as *tying within scale* [6]. Parameter invariance within scale makes a tied HMT model less image-specific, since it prevents the model from expecting smooth regions or edges at certain spatial locations a priori.

D. HMT Algorithms

The HMT is a tree-structured hidden Markov model (HMM). Thus, the three standard problems of HMMs [22] apply equally well to the HMT:

Likelihood Determination. While the HMT is a model for the joint pdf of the wavelet coefficients, the closed form expression for $f(\mathbf{w}|\boldsymbol{\Theta})$ is prohibitively complicated. Fortunately, there is a fast $O(n)$ algorithm to compute $f(\mathbf{w}|\boldsymbol{\Theta})$ for a given $\mathbf{w}$ and $\boldsymbol{\Theta}$ called the Upward-Downward algorithm [6,22–24], involving a simple sweep through the tree.

State Path Estimation. Given a set of observations $\mathbf{w}$ and a model $\boldsymbol{\Theta}$, we can determine the probability that node i is in a given state (large or small) and the most likely sequence of hidden states. Using by-products of the upward-downward algorithm, we can calculate the probability $p(S_i = q|\mathbf{w}, \boldsymbol{\Theta})$ that an observed wavelet coefficient w_i has corresponding hidden state $q \in \{\mathsf{S}, \mathsf{L}\}$. The *Viterbi algorithm* [22,23], also of $O(n)$ complexity, finds the most likely state sequence that produced the observed wavelet coefficients.

Model Training. In many situations, we would like to fit the HMT parameters $\boldsymbol{\Theta}$ to a given set of training data. For example, we could desire the *most likely* $\boldsymbol{\Theta}$ that could give rise to the training observations $\mathbf{w}$ (the ML estimate):

$$\widehat{\boldsymbol{\Theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\Theta}} f(\mathbf{w}|\boldsymbol{\Theta}). \tag{11}$$

Since the state values are unknown (hidden), finding the ML estimate directly is intractable. However, if the states are known, finding $\widehat{\boldsymbol{\Theta}}_{\mathrm{ML}}$ is easy, since the coefficients are merely independent Gaussian random variables.

The Expectation-Maximization (EM) algorithm attacks this sort of "hidden data" problem [6, 22–24]. We start with an initial guess $\boldsymbol{\Theta}^0$ of the model parameters, and then for each iteration l we calculate $E[\log f(\mathbf{w}, \mathbf{S}|\boldsymbol{\Theta})|\mathbf{w}, \boldsymbol{\Theta}^l]$. Finding this expectation, called the "E step," amounts to calculating the state probabilities $p(S_i = q|\mathbf{w}, \boldsymbol{\Theta}^l)$, for which we use the upward-downward algorithm. The maximization, or "M step" consists of relatively simple, closed form updates of the parameters in $\boldsymbol{\Theta}^l$ to obtain $\boldsymbol{\Theta}^{l+1}$. As $l \to \infty$, $\boldsymbol{\Theta}^l$ approaches a local maximum of the likelihood function $f(\mathbf{w}|\boldsymbol{\Theta})$ [25]. The EM algorithm is $O(n)$ per iteration.

While simple, EM training for the HMT has several drawbacks. Being a hill-climber, the EM algorithm is guaranteed to convergence only to a local maximum of $f(\mathbf{w}|\boldsymbol{\Theta})$. Nevertheless, we obtain reasonable estimates in general. More importantly, convergence can be relatively slow. For large images, this can make training very computationally expensive. Even though each iteration of the algorithm is $O(n)$, there is nothing to limit the number of iterations it takes to converge. For example, convergence on a 512×512 image can take anywhere from minutes to hours on a standard workstation.

E. Application: Empirical Bayesian Estimation

To demonstrate the effectiveness of the HMT as a model for an image's wavelet coefficients, we estimate an image $\mathbf{x}$ submerged in additive white Gaussian noise. This is a straightforward extension to 2-D of the work in [6]. Given a noisy observation

$$\mathbf{v} = \mathbf{x} + \mathbf{n}, \tag{12}$$

with $\mathbf{n}$ a Gaussian random field whose components are independent and identically distributed (iid) with zero mean and known variance σ_n^2, we wish to estimate the underlying image $\mathbf{x}$. Translated into the wavelet domain, the problem is as follows:

$$\text{given} \quad \mathbf{y} = \mathbf{w} + \mathbf{n}', \quad \text{estimate} \quad \mathbf{w}, \tag{13}$$

where $\mathbf{n}'$ is again Gaussian iid with variance σ_n^2.

Since we are viewing $\mathbf{w}$ as a realization of a random field whose joint pdf $f(\mathbf{w}|\mathbf{\Theta})$ is modeled by the HMT, we take a Bayesian approach to this 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 parameters $\mathbf{\Theta}$ for the prior $f(\mathbf{w}|\mathbf{\Theta})$, Crouse et al. [6] take an *empirical Bayesian* approach. 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). A strictly Bayesian approach would require that we take the parameters as known (see Section IV-C below) or assign them a hyper-prior [7, 26].

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

$$\widehat{w}_i := E[w_i|\mathbf{y}, \mathbf{\Theta}] = \sum_{q \in \{\mathsf{S},\mathsf{L}\}} p(S_i = q|\mathbf{y}, \mathbf{\Theta}) \, \frac{\sigma_{q;i}^2}{\sigma_n^2 + \sigma_{q;i}^2} \, y_i \tag{14}$$

to obtain the minimum mean-square estimate (MMSE) of $\mathbf{w}$.

The results of this procedure for a number of test images are summarized in the third column of Tables I–III, and an example is shown in Fig. 2(f). The HMT empirical Bayesian estimator outperforms other DWT wavelet shrinkage techniques in terms of MSE, and in visual terms it is far superior, boasting estimates with sharper and more accurate edges. In fact, its MSE and visual performance are quite competitive with RDWT wavelet shrinkage [13, 27] (the current state-of-the-art in performance).

The estimator (14) is just one of the possible approaches to denoising using the HMT. Although (14) gives the estimate with the MMSE under the HMT model, the choice of squared error loss is somewhat arbitrary. Another Bayes estimator, e.g., a MAP estimator for 0/1 loss, could be used in its place. Alternatively, the model could be used outside the strict Bayesian framework. For instance, a thresholding technique based on the Viterbi algorithm can be used to determine which large wavelet coefficients are likely a part of the edge structure and should be kept (coefficients with associated hidden state L) and which ones are due to noise and should be killed [28].

IV. A Reduced-Parameter HMT Image Model

By design, the HMT model captures the main statistical features of the wavelet transforms of real-world images. In its raw form, however, the $4J$ parameters needed to model an image can make it unwieldy, even when tying within scale. This poses a number of problems. Directly specifying $4J$ parameters requires a tremendous amount of a priori information about the image, but without this information we run the risk of over-fitting the model. Training the parameters can be time consuming and may not be robust under unfavorable conditions. The empirical Bayes estimator of Section III-E works well, but requires the use of the EM algorithm, which at $O(n)$ computational complexity per iteration, can be very time consuming. All of these make the HMT inappropriate for applications with minimal available a priori information or that require rapid processing.

To address these problems, we must reduce the number of parameters in the HMT model. Because of this reduction in complexity, less a priori information will be needed to specify the model parameters. However, the HMT model will become less accurate: two images that have different parameterizations in the general form of the HMT may have the same parameterization in a reduced-parameter model.

The amount of parameter reduction that is appropriate depends on the application and the amount of information known about the images to be modeled. For example, in estimation/denoising the assumptions are usually very broad; that is, the noise-corrupted image is assumed "photograph-like." The estimator needs only to differentiate between image and noise. These entities have very different structure and hence can be modeled by very different HMTs (and thus differentiated using only a small set of parameters). In detection and classification, on the other hand, the differences in structure between the two hypotheses may be more subtle, and the models may need to be more specific and thus described by more parameters.

In [6] and in Section III, the modeling paradigm was to assign a different set of HMT parameters to each image, with the $4J$ parameters being specified by training on an observation. In this section we take a different approach. By taking advantage of image properties not yet explicitly recognized by the HMT, we will specify a set of (only 9) meta-parameters that determine the $4J$ HMT parameters.

These additional properties, introduced as the *tertiary properties of wavelet coefficients* (**T1**,**T2**) in the Introduction, are motivated by examining 1-D cross-sections (slices) of images (similar to the approach of [29]). These 1-D slices consist of piecewise smooth regions separated by a finite number of discontinuities. The extension of these properties to 2-D is not exact — they hold for images with only vertical, horizontal, and 45° diagonal edges — but still remains a good approximation.

A. Incorporating the Tertiary Properties of the Wavelet Coefficients into the HMT

The wavelet transforms of real-world images exhibit additional strong statistical properties in addition to the primary (**P1**–**P5**) and the secondary (**S1**,**S2**) properties. In designing our reduced-parameter HMT model, we will use the observed tertiary properties of the wavelet coefficients stated in the Introduction: as the scale becomes finer, the magnitude of the wavelet coefficients decreases exponentially (**T1**) and persistence becomes stronger (**T2**). The tertiary properties reflect the statistical *self-similarity* across scale observed in real images [11]. Zooming in on an image adds detail at every step, and since the statistics of these new details have predictable properties, we can use this fact to reduce the model complexity.

Based on the tertiary properties of the wavelet coefficients, we can specify functional forms for the parameters of an HMT model. The coefficient decay and change in coefficient persistence are easily modeled by imposing patterns how the mixture variances and state transition probabilities change across scale. Because the characterized 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.

A.1 Modeling wavelet coefficient decay

The wavelet coefficient exponential decay property (**T1**) stems from the overall smoothness of images. Roughly speaking, a typical grayscale image consists of a number of smooth regions separated by discontinuities. This results in a generalized $1/f$-type spectral behavior [11], which leads to an exponential decay of the wavelet coefficients across scale [2].

We can easily model the exponential decay of wavelet coefficients (**T1**) through the mixture variances of the wavelet HMT model. Since the HMT mixture variances characterize the magnitudes of the wavelet coefficients, we will require that they decay exponentially across scale as well (see also Fig. 5):

$$\sigma^2_{\mathsf{S};j} = C_{\sigma_{\mathsf{S}}} 2^{-j\alpha_{\mathsf{S}}}, \tag{15}$$

$$\sigma^2_{\mathsf{L};j} = C_{\sigma_{\mathsf{L}}} 2^{-j\alpha_{\mathsf{L}}}. \tag{16}$$

Since the wavelet coefficients representing edges in an image decay slower than those representing smooth regions, we need $\sigma^2_{\mathsf{S};j} < \sigma^2_{\mathsf{L};j}$ for all scales, and thus require $\alpha_{\mathsf{S}} \geq \alpha_{\mathsf{L}}$. The result is an HMT for images with a generalized $1/f$ power spectrum.

The four meta-parameters $C_{\sigma_{\mathsf{S}}}$, α_{S}, $C_{\sigma_{\mathsf{L}}}$, and α_{L} characterize the marginal densities of the wavelet coefficients. Having marginals of this form not only meshes with the statistical self-similarity of images, but is also related to smoothness characterization using Besov spaces [15, 30]. Roughly speaking, a Besov space $B^s_q(L^p)$ contains functions with s derivatives in L^p, with q making finer smoothness distinctions [31]. For $s < 1$, $B^s_q(L^p)$ contains functions that are uniformly regular but have isolated discontinuities [2]. These properties are similar to those of real-world images; Besov spaces have been successfully used as image models for several applications [31, 32].

The fact that wavelets form an unconditional basis for all Besov spaces $B^s_q(L^p)$ means the Besov norm can be computed equivalently (subject to the constraint that the analysis wavelet ψ is smoother than the image itself) as a simple sequence norm on the wavelet coefficients [33]:

$$||x||_{B^s_q(L^p)} \asymp c_0||u_{j_0}||_p + \left[\sum_{j\geq j_0} 2^{js'q}\left(\sum_{k,b}|w^b_{j,k}|^p\right)^{q/p}\right]^{1/q} \tag{17}$$

where "$\asymp$" denotes equivalent norm, $s' = s+1-2/p, q < \infty$, and $c_0 = 2^{1/2-1/p}$. We say $x \in B^s_q(L^p)$ if $||x||_{B^s_q(L^p)} < \infty$.

For (17) to be finite, the p-norm of the wavelet coefficients at each scale must fall off exponentially. The exponential decay of the variances in the HMT model captures this fact. In fact, it has been shown in [34] that a realization from an independent Gaussian mixture model having variance parameters of the form (15)–(16) lies in $B^s_q(L^p)$, $\alpha_{\mathsf{L}}/2 > s + 1/2$, with probability 1 (a proof for a very similar statement can be found in [15] and [35]). The equivalence between Besov spaces and wavelet domain statistical models is discussed in [30].

This connection between the form of the marginals of the wavelet coefficients and Besov spaces leads us to an important point. Modeling an image as lying in a certain Besov space places restrictions on the form of the wavelet coefficient marginals, but not on their dependency structure. By characterizing the dependencies between the wavelet coefficients, as done in the next section, we are essentially refining the Besov model to consider only images that have a similar edge structure to photograph-like images.

A.2 Modeling coefficient persistence

The edge structure of images manifests itself as dependencies between the wavelet coefficients. These dependencies are represented in the HMT model by the state transition matrix (9). In this section, we take advantage of the observation that these dependencies also exhibit self-similar structure from scale to scale, codified in (**T2**), to simplify the HMT model further by assigning an exponential form to the transition matrix.

We can obtain intuition behind the persistence property (**T2**) by considering a piecewise smooth 1-D image slice containing 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 the scale j increases (**P2**), there exists a scale j_{crit} such that for all $j \geq j_{\text{crit}}$ each wavelet basis function has at most one discontinuity inside its spatial support. We call this condition *isolation of edges*. Recalling **P3**, we expect that for fine scales such that $j \geq j_{\text{crit}}$ there will be approximately 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. We can combine these facts into a grammar for the magnitudes of wavelet coefficients at scales $j \geq j_{\text{crit}}$: a small parent has two children that are also small, a large parent has one small child and one large child. As the scale increases, more of the edges become isolated, and the coefficient magnitudes follow the grammar more closely as a whole.

In 2-D, the situation is similar except that instead of 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." Now each of these coefficients will have between one and 4 large children, while the small coefficients will spawn small children.

To incorporate (**T2**) into the HMT model, we examine how the isolation of edges at fine scales controls the persistency and novelty probabilities (and hence the form of the transition matrix).

The persistence of small values is intuitive. 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, $p_j^{\mathsf{S}\to\mathsf{L}} \to 0$ as $j \to \infty$. In other words, $p_j^{\mathsf{S}\to\mathsf{S}} \to 1$, since once a basis function lies over a smooth region, all of its children also lie over that smooth region.

The persistence of large values is somewhat more complicated. Consider a 1-D wavelet coefficient p (for "parent") lying over an isolated edge ($S_p = \mathsf{L}$) in the 1-D slice at scale $j \geq j_{\text{crit}}$. Call p's children c_1 and c_2. Since the edge is perfectly localized in space, one and only one of c_1 and c_2 will be large, since the Haar basis functions corresponding to c_1 and c_2 have disjoint supports. This means that

$$
\begin{aligned}
p(S_{c_1} = \mathsf{L}, S_{c_2} = \mathsf{S} | S_p = \mathsf{L}) &= 1/2, && (18)\\
p(S_{c_1} = \mathsf{S}, S_{c_2} = \mathsf{L} | S_p = \mathsf{L}) &= 1/2, && (19)\\
p(S_{c_1} = \mathsf{S}, S_{c_2} = \mathsf{S} | S_p = \mathsf{L}) &= 0, && (20)\\
p(S_{c_1} = \mathsf{L}, S_{c_2} = \mathsf{L} | S_p = \mathsf{L}) &= 0. && (21)
\end{aligned}
$$

Because the HMT does not jointly model the state values of the children given the state of the parent, it cannot capture the property that exactly one and only one of c_1 and c_2 is large. In fact, given S_p, under the HMT S_{c_1} and S_{c_2} are independent. Instead of modeling the joint distribution of S_{c_1} and S_{c_2} given S_p (18)–(21) exactly, the HMT approximates it as the product of the marginals $p(S_{c_1}|S_p)$ and $p(S_{c_2}|S_p)$ with

$$
\begin{aligned}
p(S_{c_1} = \mathsf{L} | S_p = \mathsf{L}) &= p(S_{c_1} = \mathsf{L}, S_{c_2} = \mathsf{S} | S_p = \mathsf{L}) + && (22)\\
&\quad\; p(S_{c_1} = \mathsf{L}, S_{c_2} = \mathsf{L} | S_p = \mathsf{L}) = 1/2, &&\\
p(S_{c_2} = \mathsf{L} | S_p = \mathsf{L}) &= p(S_{c_1} = \mathsf{S}, S_{c_2} = \mathsf{L} | S_p = \mathsf{L}) + && (23)\\
&\quad\; p(S_{c_1} = \mathsf{L}, S_{c_2} = \mathsf{L} | S_p = \mathsf{L}) = 1/2. &&
\end{aligned}
$$

As a result, the HMT persistency probability $p_j^{\mathsf{L}\to\mathsf{L}} \to 1/2$ as $j \to \infty$. Admittedly, this is an imperfect model, since for all values of j, there is a chance that the edge will disappear (since $p(S_{c_1} = \mathsf{S}, S_{c_2} = \mathsf{S} | S_p = \mathsf{L}) = 1/4$ under the HMT) or bifurcate [2] (since $p(S_{c_1} = \mathsf{L}, S_{c_2} = \mathsf{L} | S_p = \mathsf{L}) = 1/4$). However, capturing the exact behavior of (18)–(21) would necessitate the use of a more complicated joint model for the states of the children coefficients given the state of the parent coefficient. For wavelets other than the Haar, the support of the basis functions of the children are not necessarily disjoint. However, (18)–(21) hold within a reasonable approximation.

Extension of this analysis to 2-D is also not exact, except for horizontal, vertical, and diagonal edges. In 2-D, edges lie on curves in space, and the curve could conceivably intersect the spatial support of the basis functions of any of the children of a coefficient that has isolated the curve. However, the curves become straight lines asymptotically inside the support of the wavelet and so they encounter only two children in the limit.

We have observed for a number of grayscale images that the transition matrix entries approach their asymptotic values in a roughly exponential manner (see Fig. 5). This observation makes sense, since the transition probabilities rely on the edges being isolated, which becomes (approximately) exponentially more likely as scale increases. We therefore impose the following exponential form for the state transition matrix (9) specified by four parameters:

$$A_j = \begin{bmatrix} 1 - C_{\mathsf{SS}} 2^{-\gamma_{\mathsf{S}} j} & C_{\mathsf{SS}} 2^{-\gamma_{\mathsf{S}} j} \\ \frac{1}{2} - C_{\mathsf{LL}} 2^{-\gamma_{\mathsf{L}} j} & \frac{1}{2} + C_{\mathsf{LL}} 2^{-\gamma_{\mathsf{L}} j} \end{bmatrix}. \tag{24}$$

The transition matrix has the asymptotic form

$$A_\infty = \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}. \tag{25}$$

A.3 HMT meta-parameters

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 specify all of the HMT parameters with 9 meta-parameters:

$$\mathbf{\Theta}_{\mathrm{m}} = \{\alpha_{\mathsf{S}}, C_{\sigma_{\mathsf{S}}}, \alpha_{\mathsf{L}}, C_{\sigma_{\mathsf{L}}}, \gamma_{\mathsf{S}}, C_{\mathsf{SS}}, \gamma_{\mathsf{L}}, C_{\mathsf{LL}}, p_{j_0}^{L}\}. \tag{26}$$

The self-similarity of images is reflected in the self-similarity of the HMT model parameters. The fact that the model parameters can be captured by these functional forms means that statistical behavior of images at a fine scale is predictable from the statistical behavior at a coarser scale. Not only does the introduction of the HMT meta-parameters reduce the complexity of the model, but it integrates a key property of real-world images.

B. A "Universal" Grayscale Image Model: The uHMT

Now that we have an image model specified by a small set of meta-parameters $\mathbf{\Theta}_{\mathrm{m}}$, we must find a way of specifying them. The first possibility would be to derive a constrained EM algorithm to give pseudo-MLE estimates of $\mathbf{\Theta}_{\mathrm{m}}$ given an observation. Deriving the steps for this algorithm is difficult, and there is no guarantee that the training would be any faster than in the unconstrained case.

Another possibility is to fix the meta-parameters directly. This yields an HMT model for a class of images, with each member in the class being treated as statistically equivalent. To see how much variation in the HMT meta-parameters there is across photograph-like images, we trained HMT models for a set of wavelet transforms (using the Daubechies-8 wavelet) of normalized photograph-like 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 many of the images (see Fig. 5). Since the images were normalized, the range over which the variances decayed was similar as well. These observations confirm that we can use a specific, "universal" set of HMT meta-parameters to reasonably characterize photograph-like images.

The universal parameters obtained by (jointly) fitting lines to the HMT parameters of four images (see Fig. 5) are given by

$$\mathbf{\Theta}_{\mathrm{m}} := \begin{cases} \alpha_{\mathsf{S}} &= 3.1 \\ C_{\sigma_{\mathsf{S}}} &= 2^{11} \\ \alpha_{\mathsf{L}} &= 2.25 \\ C_{\sigma_{\mathsf{L}}} &= 2^{11} \\ \gamma_{\mathsf{S}} &= 1 \\ C_{\mathsf{SS}} &= 2^{2.3} \\ \gamma_{\mathsf{L}} &= 0.4 \\ C_{\mathsf{LL}} &= 2^{0.5} \\ p_{j_0}^{\mathsf{L}} &= \frac{1}{2}. \end{cases} \tag{27}$$

The lines were fit to the HMT parameters starting at scale $j = 4$ ($j = 5$ for the $p^{\mathsf{L}\to\mathsf{L}}$ measurement). There are two reasons for this: 1) before this scale there is not enough data for an accurate estimate of the decays, and 2) these decay rates are really asymptotic properties. The parameter that is the most similar across all images is α_{L}. This is to be expected, since α_{L} corresponds to the decay rate of the wavelet coefficients lying over an edge, and hence is automatically independent of the image we are analyzing.

Of particular interest is the result $\alpha_{\mathsf{L}} = 2.25$. As mentioned above, in [34] it is shown that a realization from a independent mixture model is almost surely in a Besov space $B_q^s(L^p)$ if and only if $\alpha_{\mathsf{L}}/2 < s + 1/2$. Therefore, a realization from the uHMT model is almost surely in $B_q^s(L^p)$ if and only if $s < 0.625$. This closely agrees with [31], where Devore et. al found that real-world images lie in Besov spaces with $s < 0.6$.

Although we clearly lose accuracy by viewing all images we are interested in as statistically equivalent, this process totally eliminates the need for training. This can save a tremendous amount of computation, making real-time HMT processing possible.

C. Application: Bayesian Estimation with the uHMT

With the uHMT parameters, we have a fixed prior on the w_i and the estimation problem in Section III-E can be approached from a purely Bayesian standpoint. To find the conditional mean vector, the state probabilities $p(S_i = q|\mathbf{y}, \boldsymbol{\Theta}_{\mathrm{m}})$ are calculated using the upwards-downwards algorithm and used to evaluate (14). Since we have eliminated training, the estimation algorithm is truly $O(n)$ and takes only a few seconds to run on a workstation, slightly longer than simple wavelet thresholding algorithms, but much faster than the empirical Bayesian algorithm of Section III-E [6].

To test this new Bayesian estimator, we denoised the test images using the uHMT parameters presented in the last section. The estimation results are summarized in the second column of Tables I–III, and an example is given in Fig. 2(g). The results are almost identical to the much more complicated empirical Bayes HMT approach, suggesting that we have lost almost nothing by completely eliminating training.

V. Shift-Invariant HMT Image Estimation

Image estimates based on orthogonal wavelet transforms (DWTs) often exhibit visual artifacts, usually in the form of ringing around the edges. These artifacts result from the lack of shift-invariance in the DWT [27]. As we mentioned before, two different shifts of an image can have very different wavelet transforms. In particular, the wavelet domain characteristics of a singularity change as it shifts around.

For a shift-invariant estimation algorithm, we turn to the redundant wavelet transform (RDWT). Ideally, we would like to model the RDWT coefficients using an HMT in a similar fashion as in the orthogonal case. Unfortunately, the redundant transform does not have a tree-like structure, and capturing all of the important dependencies would require a complicated graph that would make Bayesian inference hard or impossible [6, 36].

Another way to make the image estimate shift-invariant is to follow the "cycle-spinning" programme proposed by Coifman and Donoho [27]. The estimation algorithm is applied to all shifts of the noisy image, and the results are averaged. The shift-invariant estimate of an $N \times N$ (n-pixel) image $\mathbf{x}$ that has been corrupted by noise, $\mathbf{v} = \mathbf{x} + \mathbf{n}$, is given by

$$\widehat{\mathbf{x}} = \frac{1}{N^2} \sum_{k,m} \widehat{\mathbf{x}}_{k,m}, \tag{28}$$

where $\widehat{\mathbf{x}}_{k,m}$ is the estimate of $\mathbf{x}$ using the (k, m) shift of $\mathbf{v}$ (there are N^2 possibilities, one for each pixel in the image). To calculate the estimate $\widehat{\mathbf{x}}_{k,m}$, shift the observation $\mathbf{v}$ by (k, m), apply the estimator, and unshift the result:

$$\widehat{\mathbf{x}}_{k,m} = \mathbf{S}_{-k,-m}(\mathbf{D}(\mathbf{S}_{k,m}(\mathbf{v}))), \tag{29}$$

where $\mathbf{S}_{k,m}(\mathbf{v}) = \mathbf{v}(s-k, t-m)$ is the 2-D shift operator and $\mathbf{D}$ denotes the estimator of Section III-E or IV-C.

This approach fits directly into the Bayesian framework. Since the estimate depends on the shift (k, m) of the data, (k, m) can be viewed as an unknown random variable. Since we have no a prior information about (k, m) except that $0 \leq k, m \leq N - 1$, we use a non-informative prior $p(k, m) = \frac{1}{N^2}$, meaning each possible shift is equally likely. Then the Bayes-optimal estimator becomes a weighted average over all shifts [26]

$$\widehat{x} = \sum_{k,m} p(k, m|y) \mathbf{S}_{-k,-m}(\mathbf{D}(\mathbf{S}_{k,m}(q))). \tag{30}$$

Since (28) makes the additional assumption that $p(k, m|y)$ is uniform, the estimator derives no information about the underlying shift given the observed data; equal weight to the estimates at each shift. If we calculated $p(k, m|y)$, we could use (30) and weight the estimate at each shift by its likelihood, but this is an expensive operation that lead to no significant gains in our experiments.

The algorithm (28), if implemented directly, would be computationally expensive, $O(n^2)$. However, if we assume that the same HMT model applies to all shifts (an assumption tacitly made in deriving the uHMT parameters), then the complexity can be reduced substantially. While the DWT tree for each shift of the image is unique, wavelet coefficients are shared between trees. There are $n \log n$ unique wavelet coefficients among the n DWT trees of an n-pixel image [27]. These $n \log n$ unique coefficients are the RDWT coefficients, as mentioned in Section II, and can be indexed by scale and shift. The DWT tree of a particular shift is embedded into the RDWT coefficients (see Fig. 6 for a 1-D example).

From the Figure, we see that the RDWT coefficients do not retain the same tree structure as the DWT; each node now has *two parents* and two children (in 1-D) with each parent coming from a different DWT tree (in the 2-D case, each node will have 4 parents and 4 children). Averaging the estimates of the image at different shifts in the spatial domain is equivalent to averaging together the estimates for each node from all trees in which it is included. Since each node still has two children, the downwards binary tree structure is preserved, and an $O(n \log n)$ algorithm can be obtained by a modification to the upward-downward algorithm (see [28] for details).

Our results using the uHMT parameters from Section IV-C in the shift-invariant estimator are summarized in the first column of Tables I–III, with an example shown in Fig. 2(h). As we see in the figure, the shift-invariant transform smoothes out the visual artifacts in the smooth regions of the image while keeping the edges sharp. We have also picked up an extra $\sim$ 1dB MSE performance over the uHMT and empirical Bayesian HMT models.

VI. Conclusions

Modeling lies at the core of any statistical image processing problem. An accurate model is of paramount importance for applications such as estimation, detection, compression and segmentation. Not only are models of great practical importance, but they also offer insight into the underlying natural structure of images.

Hidden Markov trees capture the primary aspects of image structure in the wavelet domain. In this paper, we have shown that the HMT parameters themselves have a certain form, described by the nine HMT meta-parameters, derived from the self-similar nature of real-world images. By constraining the HMT with these meta-parameters, we not only have a simpler, more concise image model, but we also incorporate more a priori information about the structure of images into the model.

The form of the HMT parameters not only agrees with the Besov space model of images, it expands on it. Besov space models capture the overall smoothness of images, a property which is reflected by the exponential decay of the mixture variances in the HMT. By including a characterization of dependencies between wavelet coefficients, the HMT also captures the edge structure of images, thus narrowing down the space of images represented by the model.

The "universal" HMT parameters (uHMT) arise naturally from the form of the HMT and accurately model a wide range of images. These nine numbers completely specify a HMT model for a large class of real-world images, eliminating any need for training and thus greatly simplifying processing algorithms and allowing real-time implementations.

With the uHMT parameters, we have specified a prior for photograph-like images. This allows us to take a Bayesian approach to statistical image processing problems; specifically, estimation in the presence of noise. Using a Bayesian approach, we are able to incorporate our knowledge of image structure into a "smart" wavelet shrinkage rule that takes into account coarse scale information while processing fine scale wavelet coefficients. The model helps predict which wavelet coefficients represent key features in the image (and thus should not be affected) and which ones represent noise (and thus should be shrunk).

Finally, the uHMT model also allows us to implement a shift-invariant estimator; a task that would be too computationally intensive if we had to train a model for every shift of the image. The shift-invariant uHMT estimator offers state-of-the-art performance in MSE and visual quality.

MATLAB code for the HMT-based denoising algorithms and the test images used for Tables I–III can be found at www.dsp.rice.edu/software/WHMT.

References

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

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

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

[4] D. L. Donoho, "Unconditional bases are optimal bases for data compression and for statistical estimation," *Applied and Computational Harmonic Analysis*, vol. 1, pp. 100–115, Dec 1993.

[5] Test images available at www.dsp.rice.edu/software/WHMT.

[6] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Proc.*, vol. 46, pp. 886–902, Apr. 1998.

[7] H. A. Chipman, E. D. Kolaczyk, and R. E. McCulloch, "Adaptive Bayesian wavelet shrinkage," *J. Amer. Stat. Assoc.*, vol. 92, 1997.

[8] E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," in *Proc. 31st Asilomar Conf.*, (Pacific Grove, CA), pp. 673–678, Nov. 1997.

[9] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. on Pattern Anal. Machine Intell.*, vol. 14, pp. 710–732, July 1992.

[10] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Proc.*, vol. 41, pp. 3445–3462, Dec. 1993.

[11] D. L. Ruderman and W. Bialek, "Statistics of natural images: Scaling in the woods," *Phys. Rev. Letters*, vol. 73(6), pp. 814–817, 1994.

[12] R. M. Dufour and E. L. Miller, "Statistical signal restoration with $1/f$ wavelet domain prior models," *Signal Processing*, vol. 78, pp. 209–307, 1998.

[13] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus, and R. O. Wells, "Noise reduction using an undecimated discrete wavelet transform," *IEEE Sig. Proc. Letters*, vol. 3, pp. 10–12, Jan. 1996.

[14] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors," *IEEE Trans. on Info. Theory*, vol. 45, pp. 909–919, Apr. 1999.

[15] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *J. Royal Stat. Soc., Series B*, vol. 60, pp. 725–749, 1998.

[16] K. Mihçak, I. Kozintev, K. Ramchandran, and P. Moulin, "Low complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Sig. Proc. Letters*, vol. 6, pp. 300–303, Dec. 1999.

[17] I. Daubechies, *Ten Lectures on Wavelets*. New York: SIAM, 1992.

[18] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. on Pattern Anal. Machine Intell.*, vol. 11, pp. 674–693, July 1989.

[19] C. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice Hall, 1998.

[20] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden Markov models," in *Proc. SPIE Conference on Mathematical Modeling, Bayesian Estimation, and Inverse Problems,*, vol. 3816, (Denver, CO), pp. 31–44, July 1999.

[21] H. Choi and R. G. Baraniuk, "Multiscale image segmentation using wavelet-domain hidden Markov models," *To appear in IEEE Trans. Image Proc.*, 2001.

[22] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–285, Feb. 1989.

[23] B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*. Cambridge, MA: MIT Press, 1998.

[24] O. Ronen, J. R. Rohlicek, and M. Ostendorf, "Parameter estimation of dependence tree models using the EM algorithm," *IEEE Sig. Proc. Letters*, Aug. 1995.

[25] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, April 1984.

[26] M. A. T. Figueiredo and R. D. Nowak, "Bayesian wavelet-based signal estimation using non-informative priors," in *Proc. 32nd Asilomar Conf.*, 1998.

[27] R. Coifman and D. Donoho, "Translation-invariant de-noising," in *Wavelets and Statistics* (Antoniadis/Oppenheim, ed.), vol. 103 of *Lecture Notes in Statistics*, Springer-Verlag, 1995.

[28] J. K. Romberg, "A universal hidden Markov tree image model," Master's thesis, Rice University, 1999. Available at www.dsp.rice.edu/publications.

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

[30] H. Choi and R. Baraniuk, "Wavelet-domain statistical models and Besov spaces," in *Proc. 44th SPIE Conf.*, (Denver, CO), July 1999.

[31] R. A. DeVore, B. Jawerth, and B. J. Lucier, "Image compression through wavelet transform coding," *IEEE Trans. Info. Theory*, vol. 38, pp. 719–746, Mar. 1992.

[32] A. Chambolle, R. A. DeVore, N. Lee, and B. J. Lucier, "Nonlinear wavelet image processing: Variational

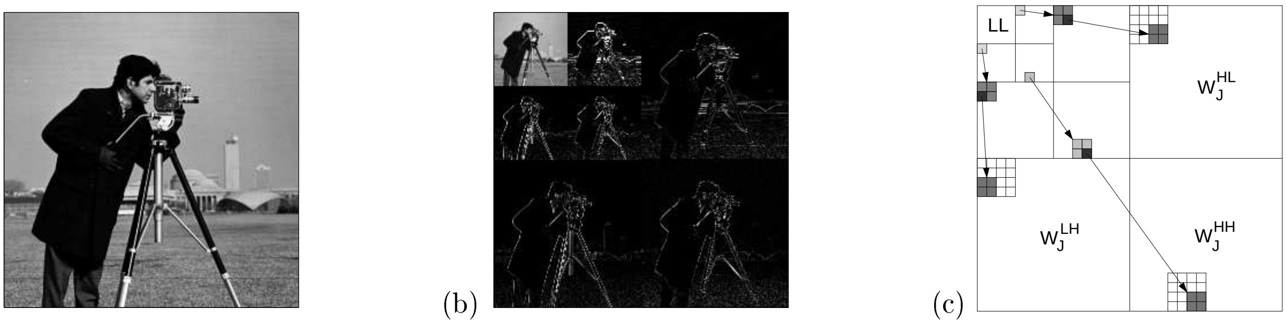
problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. on Image Proc.*, vol. 7, pp. 319–355, July 1998.

[33] Y. Meyer, *Wavelets and Operators.* Cambridge Univ. Press, 1992.

[34] H. Choi and R. Baraniuk, "Statistical wavelet models and function spaces," *preprint*, 2000.

[35] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Stochastic expansions in and overcomplete wavelet dictionary," *Probability Theory and Related Fields*, vol. 117, pp. 133–144, 2000.

[36] H. Lucke, "Which stochastic models allow Baum-Welch training?," *IEEE Trans. Signal Processing*, vol. 44, Nov. 1996.



(a) (b) (c)

Fig. 1. *(a) "camera" image [5]. (b) The 2-D wavelet transform represents an image in terms of (lowpass) scaling coefficients and 3 subbands of (bandpass) wavelet coefficients that detect edges in the horizontal (LH), vertical (HL), and diagonal directions (HH). (c) The wavelet subbands form three multiscale quad-trees, with each (parent) coefficient having four child coefficients in the next finer scale band. The child wavelets divide the support of the parent wavelet in four.*

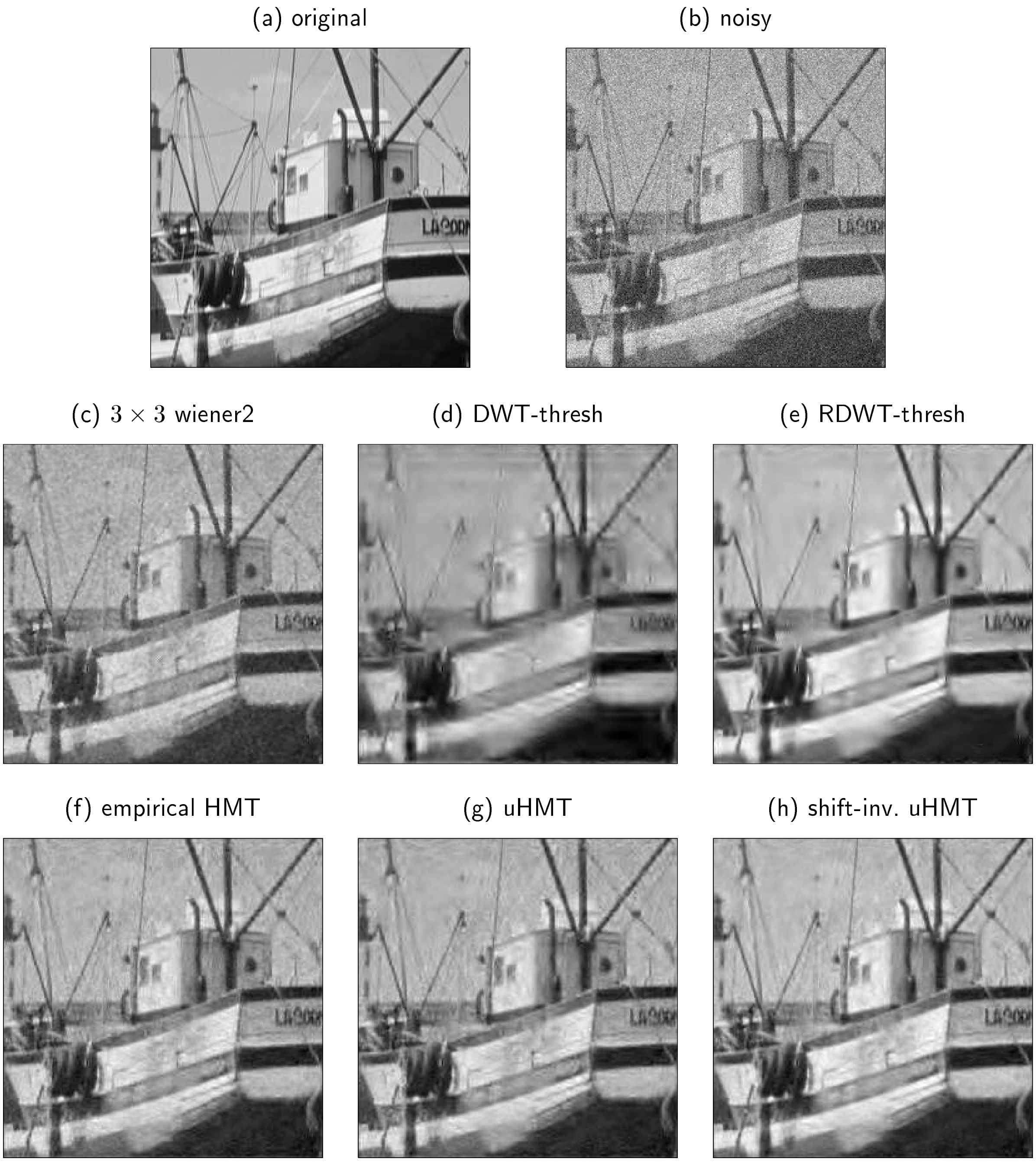


Fig. 2. *Images from the denoising experiment corresponding to the third row of Table II. (a) Original* 256×256 *"boats" image [5]. (b) Noisy boats image, with* $\sigma_n = 0.1$*, PSNR=20dB. Boats image denoised using (c) spatial domain* 3×3 *Wiener filter (*`wiener2` *command in MATLAB), PSNR=26.1dB; (d) soft thresholded discrete wavelet transform with threshold in [13], PSNR=22.5dB; (e) hard thresholded redundant wavelet transform (RDWT) with threshold chosen in [13], PSNR=26.3dB; (f) the empirical Bayesian HMT estimator of Section III-E, [6] PSNR=26.5dB; (g) the uHMT of Section IV-C, PSNR=26.4dB; (h) the shift-invariant uHMT estimation algorithm of Section V, PSNR=27.4dB.*

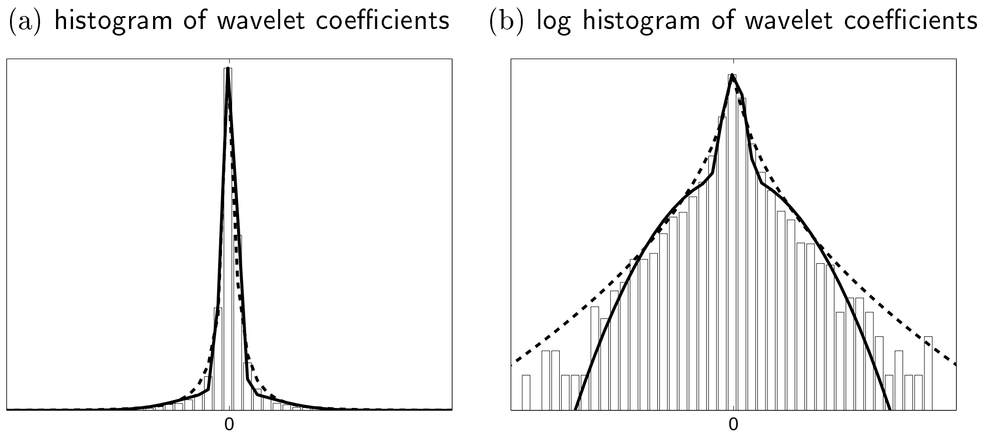


Fig. 3. *(a) Histogram and (b) log-histogram of the wavelet coefficients in one subband of the "fruit" image [5]. The dotted line is a generalized Gaussian approximation ($\nu = 0.5$). The solid line is a two-component Gaussian mixture model fitted to the data. Although the generalized Gaussian density is a better fit, by using only two states in the Gaussian mixture model, we achieve a close fit to the histogram. The Gaussian mixture model is not exact, but it allows simple and efficient algorithms, especially for capturing dependencies between wavelet coefficients.*

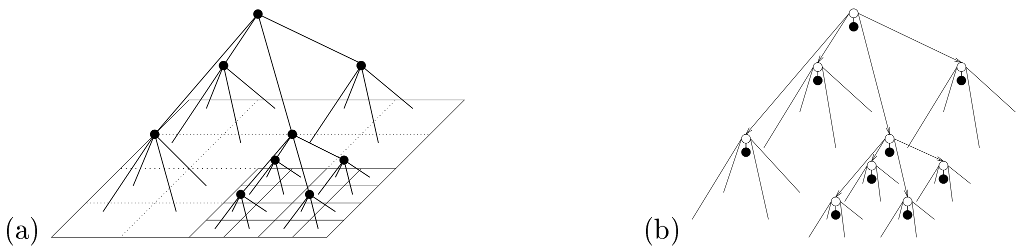


Fig. 4. *(a) Quad-tree organization of the wavelet coefficients (black nodes) in one subband of the wavelet transform. Four children wavelet coefficients divide the spatial localization of the parent coefficient. (b) 2-D HMT model. Each black node is a wavelet coefficient w_i; each white node is the corresponding hidden state (S_i in (6)–(7)). Links represent dependencies between states (quantified by (9)).*

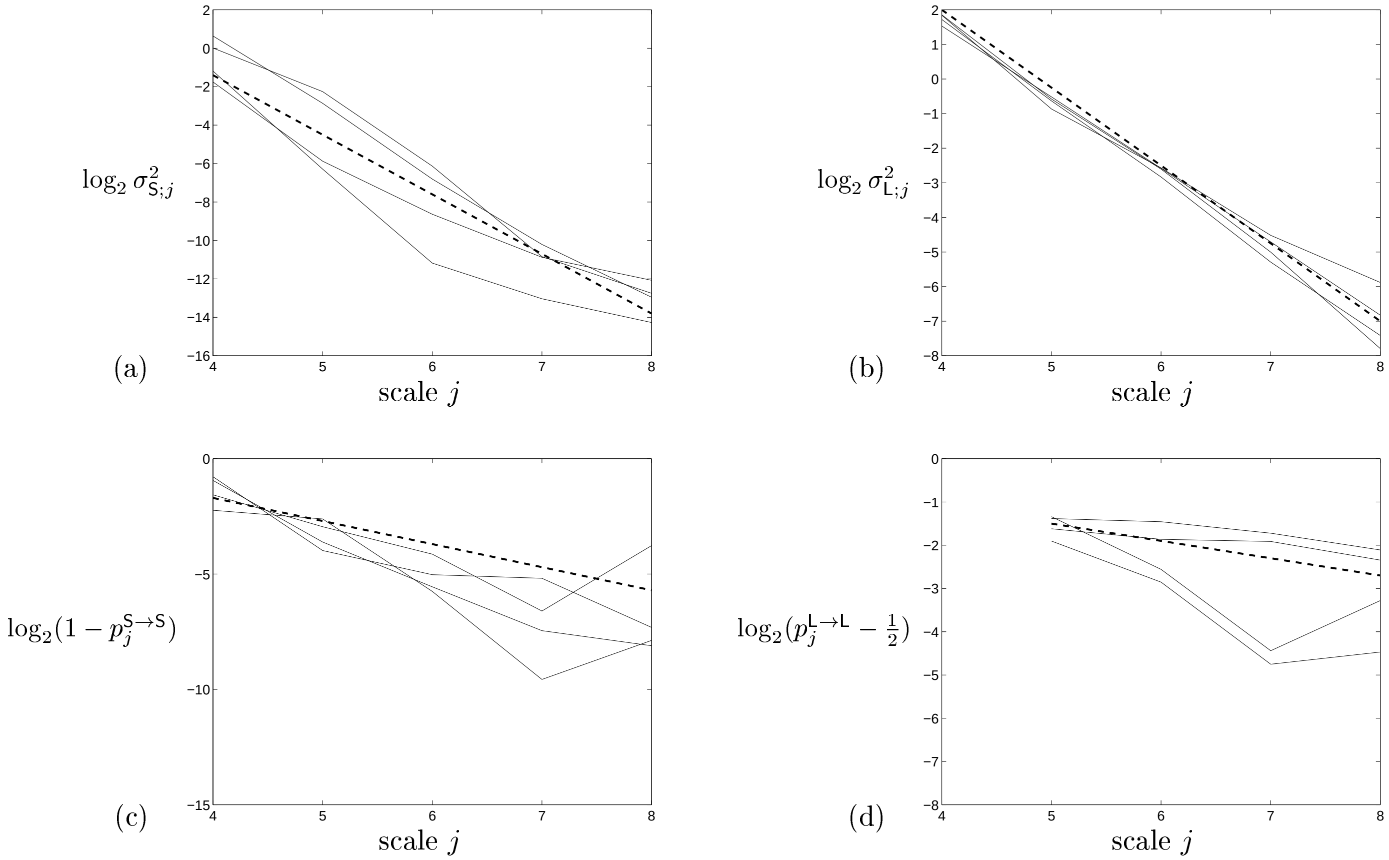


Fig. 5. *Universal HMT parameters against trained parameters for images. The solid lines are the parameters for the "baby", "camera", "fruit", and "lenna" images [5] plotted across scale. The dotted lines represent the uHMT parameters presented in Section IV-B. Reliable estimates for* $p_j^{\mathsf{L}\to\mathsf{L}}$ *require more data than the other 3 parameters, so the behavior is shown from scale* $j = 5$ *onwards.*

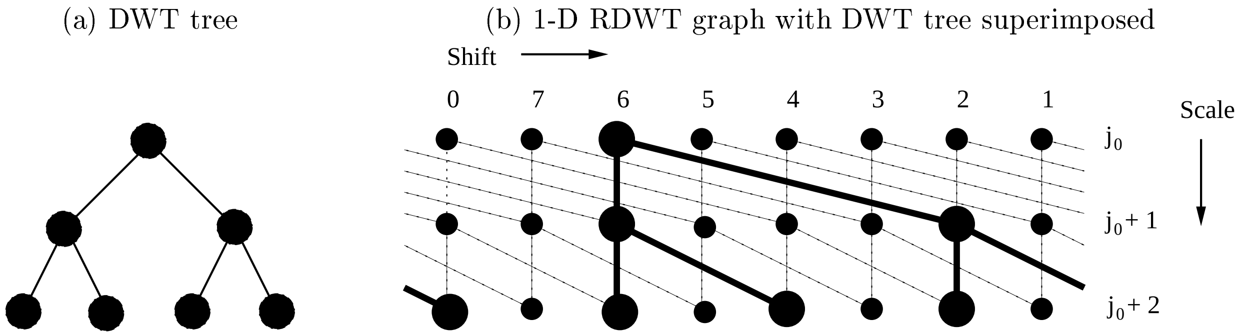


Fig. 6. *(a) DWT tree and (b) DWT tree at one shift (shift 6) embedded into the RDWT graph. Note that the DWT trees overlap — the same coefficient appears in more that one tree. There are* $n \log n$ *unique coefficients for a length-*n *signal. Also note that each node now has two parents as well as two children, and is included in* 2^j *different trees. In 2-D, the RDWT graph consists of overlapping quad-trees; each node has four parents and four children and is included in* 4^j *different trees.*

TABLE I

Image estimation results for 256×256 images corrupted with additive white Gaussian noise of $\sigma_n = 0.05$. Entries are the the peak signal-to-noise ratio (PSNR) in dB, PSNR $:= -20\log_{10}(||\widehat{x} - x||_2/N)$ *(larger numbers mean better performance). Pixel intensity vales were normalized between 0 and 1. All results use the Daubechies-8 wavelet. "si-HMT" is the shift-invariant estimator of Section V; "uHMT" uses the "universal" parameters presented in Section IV-C; "emp-HMT" uses the empirical Bayesian estimator of Section III-E; "RDWT-Thresh" uses a hard thresholded redundant wavelet transform using the thresholds in [13]; "DWT-Thresh" uses a thresholded orthogonal wavelet transform using the thresholds in [13]; and "wiener2" is the 2-D spatially adaptive Wiener filter command from* MATLAB.

Image	si-HMT	uHMT	emp-HMT	RDWT-Thresh	DWT-Thresh	wiener2
baby	**33.1**	32.4	32.6	32.7	28.6	32.1
birthday	**29.6**	28.9	29.1	27.5	24.4	28.1
boats	**31.4**	30.4	30.6	30.3	25.6	29.8
bridge	**28.9**	28.1	28.3	26.2	23.1	27.0
buck	33.7	32.5	32.8	**33.8**	27.8	33.0
building	**30.4**	29.7	30.0	29.0	24.8	28.9
camera	**31.1**	30.3	30.5	29.8	25.4	29.8
clown	**31.7**	30.6	30.9	30.6	25.8	30.7
fruit	**33.3**	32.2	32.6	32.8	27.8	32.6
kgirl	**32.6**	31.6	31.8	31.5	27.5	31.7
lenna	**31.3**	30.4	30.5	29.7	25.6	30.2

TABLE II

Estimation PSNR results for images corrupted with $\sigma_n = 0.1$.

Image	si-HMT	uHMT	emp-HMT	RDWT-Thresh	DWT-Thresh	wiener2
baby	**29.6**	28.9	29.2	29.5	25.6	27.2
birthday	**26.4**	25.8	25.8	25.3	22.4	25.5
boats	**27.4**	26.4	26.5	26.3	22.5	26.1
bridge	**25.3**	24.6	25.0	23.7	21.2	24.7
buck	29.6	28.4	28.6	**29.7**	24.2	27.6
building	**26.6**	25.9	26.3	25.8	22.2	25.6
camera	**27.0**	26.2	26.4	26.3	22.7	26.1
clown	**27.8**	26.8	26.8	26.5	22.8	26.5
fruit	**29.7**	28.5	28.6	29.0	24.6	27.2
kgirl	**29.3**	28.3	28.3	28.4	24.8	26.8
lenna	**27.6**	26.7	26.7	26.3	23.0	26.2

TABLE III

Estimation PSNR results for images corrupted with $\sigma_n = 0.2$.

Image	si-HMT	uHMT	emp-HMT	RDWT-Thresh	DWT-Thresh	wiener2
baby	**26.3**	25.8	25.4	26.1	23.0	21.3
birthday	**23.7**	23.1	23.0	23.0	20.3	20.8
boats	**24.1**	23.3	23.3	23.1	20.0	21.0
bridge	**22.7**	22.0	22.2	21.4	19.4	20.4
buck	**25.8**	24.7	24.5	25.6	21.1	21.3
building	**23.5**	22.8	23.0	23.0	19.9	20.8
camera	**23.7**	23.1	23.2	23.2	20.4	20.8
clown	**24.5**	23.7	23.6	23.2	20.2	21.1
fruit	**26.4**	25.3	25.0	25.3	21.8	21.3
kgirl	**26.4**	25.4	25.3	25.3	22.4	21.1
lenna	**24.5**	23.8	23.8	23.5	20.7	20.9

Figure and Table Captions

Figure 1: (a) “camera” image [5]. (b) The 2-D wavelet transform represents an image in terms of (lowpass) scaling coefficients and 3 subbands of (bandpass) wavelet coefficients that detect edges in the horizontal (LH), vertical (HL), and diagonal directions (HH). (c) The wavelet subbands form three multiscale quad-trees, with each (parent) coefficient having four child coefficients in the next finer scale band. The child wavelets divide the support of the parent wavelet in four.

Figure 2: Images from the denoising experiment corresponding to the third row of Table II. (a) Original 256×256 “boats” image [5]. (b) Noisy boats image, with $\sigma_n = 0.1$, PSNR=20dB. Boats image denoised using (c) spatial domain 3×3 Wiener filter (`wiener2` command in MATLAB), PSNR=26.1dB; (d) soft thresholded discrete wavelet transform with threshold in [13], PSNR=22.5dB; (e) hard thresholded redundant wavelet transform (RDWT) with threshold chosen in [13], PSNR=26.3dB; (f) the empirical Bayesian HMT estimator of Section III-E, [6] PSNR=26.5dB; (g) the uHMT of Section IV-C, PSNR=26.4dB; (h) the shift-invariant uHMT estimation algorithm of Section V, PSNR=27.4dB.

Figure 3: (a) Histogram and (b) log-histogram of the wavelet coefficients in one subband of the “fruit” image [5]. The dotted line is a generalized Gaussian approximation ($\nu = 0.5$). The solid line is a two-component Gaussian mixture model fitted to the data. Although the generalized Gaussian density is a better fit, by using only two states in the Gaussian mixture model, we achieve a close fit to the histogram. The Gaussian mixture model is not exact, but it allows simple and efficient algorithms, especially for capturing dependencies between wavelet coefficients.

Figure 4: (a) Quad-tree organization of the wavelet coefficients (black nodes) in one subband of the wavelet transform. Four children wavelet coefficients divide the spatial localization of the parent coefficient. (b) 2-D HMT model. Each black node is a wavelet coefficient w_i; each white node is the corresponding hidden state (S_i in (6)–(7)). Links represent dependencies between states (quantified by (9)).

Figure 5: Universal HMT parameters against trained parameters for images. The solid lines are the parameters for the “baby”, “camera”, “fruit”, and “lenna” images [5] plotted across scale. The dotted lines represent the uHMT parameters presented in Section IV-B. Reliable estimates for

$p_j^{\mathsf{L}\to\mathsf{L}}$ require more data than the other 3 parameters, so the behavior is shown from scale $j = 5$ onwards.

Figure 6: (a) DWT tree and (b) DWT tree at one shift (shift 6) embedded into the RDWT graph. Note that the DWT trees overlap — the same coefficient appears in more that one tree. There are $n \log n$ unique coefficients for a length-n signal. Also note that each node now has two parents as well as two children, and is included in 2^j different trees. In 2-D, the RDWT graph consists of overlapping quad-trees; each node has four parents and four children and is included in 4^j different trees.

Table 1: Image estimation results for 256×256 images corrupted with additive white Gaussian noise of $\sigma_n = 0.05$. Entries are the the peak signal-to-noise ratio (PSNR) in dB, $PSNR := -20\log_{10}(||\widehat{x} - x||_2/N)$ (larger numbers mean better performance). Pixel intensity vales were normalized between 0 and 1. All results use the Daubechies-8 wavelet. "si-HMT" is the shift-invariant estimator of Section V; "uHMT" uses the "universal" parameters presented in Section IV-C; "emp-HMT" uses the empirical Bayesian estimator of Section III-E; "RDWT-Thresh" uses a hard thresholded redundant wavelet transform using the thresholds in [13]; "DWT-Thresh" uses a thresholded orthogonal wavelet transform using the thresholds in [13]; and "wiener2" is the 2-D spatially adaptive Wiener filter command from MATLAB.

Table 2: Estimation PSNR results for images corrupted with $\sigma_n = 0.1$.

Table 3: Estimation PSNR results for images corrupted with $\sigma_n = 0.2$.

Author Biographies

Justin K. Romberg received the BSEE degree in 1997 and the MSEE degree in 1999 from Rice University, Houston, Texas, where he holds a Texas Instruments Fellowship as a PhD student in electrical and computer engineering. In 1995 he was a research engineer for the MITRE corporation, and he spent the summer of 2000 at the Xerox Palo Alto Research Center.

His research interests include multiscale statistical modeling, image processing, and applied harmonic analysis.

Hyeokho Choi was born in Korea in 1969. He received the B.S. degree in control and instrumentation engineering (summa cum laude) from Seoul National University, Seoul, Korea, in 1991 and the M.S. and Ph.D degrees in electrical engineering from the University of Illinois at Urbana-Champaign, in 1993 and 1998, respectively. His thesis research was in the area of computed imaging systems and signal processing.

Since January 1998, he has been at Rice University, Houston, TX, where he is currently a Research Professor in the Department of Electrical and Computer Engineering. His current research interests lie in the area of statistical signal processing, pattern recognition, wavelet theory, and imaging systems.

Richard G. Baraniuk was born on the lone prairie in Alberta, Canada. He received the BSc degree in 1987 from the University of Manitoba (Canada), the MSc degree in 1988 from the University of Wisconsin-Madison, and the PhD degree in 1992 from the University of Illinois at Urbana-Champaign, all in Electrical Engineering. In 1986, he was a research engineer with Omron Tateisi Electronics in Kyoto, Japan. After spending 1992–1993 with the Signal Processing Laboratory of Ecole Normale Supérieure, in Lyon, France, he joined Rice University, where he is currently a Professor of Electrical and Computer Engineering and DJ at KTRU. He spent Autumn 1998 at the Isaac Newton Institute of Cambridge University, England as the Rosenbaum Fellow. His research interests lie in the area of signal and image processing and include wavelets, probabilistic models, networks, time-frequency analysis, and rodeo theory.

Dr. Baraniuk received a NATO postdoctoral fellowship from NSERC in 1992, the National Young Investigator award from the National Science Foundation in 1994, a Young Investigator Award from the Office of Naval Research in 1995, the Rosenbaum Fellowship from the Newton Institute in 1998, the C. Holmes MacDonald National Outstanding Teaching Award from Eta Kappa Nu in 1999, the Charles Duncan Junior Faculty Achievement Award from Rice in 2000, and the ECE Young Alumni Achievement Award from the University of Illinois in 2000. He was co-author on a paper with Matthew Crouse and Robert Nowak that received the IEEE Signal Processing Society Junior Paper Award in 2001. He is a member of the "Signal Processing Theory and Methods" Technical Committee of the IEEE Signal Processing Society and serves on the editorial board of the journal *Applied and Computational Harmonic Analysis*.