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, Houston, TX 77005, USA

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 density 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 Expectation-Maximization algorithm, for example). In this paper, we propose two reduced-parameter HMT models that capture the general structure of a broad class of real-world images. In the *image HMT (iHMT)* model we use the fact that for a large class of images the structure of the HMT is self-similar across scale. This allows us to reduce the complexity of the iHMT to just nine easily trained parameters (independent of the size of the image and the number of wavelet scales). In the *universal HMT (uHMT)* we take a Bayesian approach and fix these nine parameters. The uHMT requires no training of any kind. While simple, we show using a series of image estimation/denoising experiments that these two new models retain nearly all of the key structure modeled by the full HMT. Finally, we propose a fast shift-invariant HMT estimation algorithm that outperforms all other wavelet-based estimators in the current literature, both in mean-square error and visual metrics.

1. INTRODUCTION

In statistical image processing, we view an image $\boldsymbol{x}$ as a realization of a random field with joint probability density function (pdf) $f(\boldsymbol{x})$. All statistical image processing problems, such as estimation, detection, and classification, rely on knowledge of $f(\boldsymbol{x})$; the more accurately it can be specified, the better the solutions. Of course, the joint pdf is rarely known exactly. In such cases, it is attractive to use a *model* that approximates $f(\boldsymbol{x})$.

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 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. Many 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 *primary properties* of wavelet transforms make wavelet-domain statistical image processing attractive:[2,3]

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. 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.

P4. Decorrelation: The wavelet coefficients of real-world images tend to be approximately decorrelated.

P5. 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 in a quad-tree structure with three subbands representing the horizontal, vertical, and diagonal edges in the image (see Fig. 1). Due to the Compaction property **P3**, which follows from the fact that the edges constitute only a very small portion of a typical image, 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.

Email: jrom@rice.edu, choi@ece.rice.edu, richb@rice.edu. Web: www.dsp.rice.edu. This work was supported by the National Science Foundation, grant MIP-9457438, DARPA/AFOSR, grant F49620-97-1-0513, and Texas Instruments.

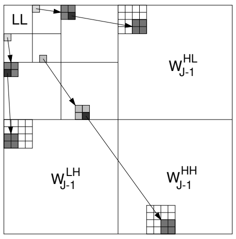
(a)  (c)  (c) 

Figure 1. (a) Cameraman image. (b) 2-D wavelet transform represents an image in terms of (lowpass) scaling coefficients and subbands of (bandpass) wavelet coefficients that detect edges in the horizontal, vertical, and diagonal directions. (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.

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 quad-trees.

NonGaussianity follows immediately from Energy Compaction (**P3**). Persistency follows from the Edge Detection (**P5**) 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*.[4]

The HMT models the nonGaussian marginal pdf (**S1**) as a Gaussian mixture whose components are labeled by a hidden state that signifies the large/small magnitude of the coefficient. The persistence of wavelet coefficient magnitude across scale (**S2**) is modeled by linking the hidden states across scale in a Markov tree. A state transition matrix for each link quantifies statistically the degree of persistence of large/small coefficients. The parameters of the HMT models can be fit to a set of training data using the Expectation-Maximization (EM) algorithm. The training yields an approximate maximum likelihood estimate of the model parameters given the training data; these parameters yield 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 not available, we run the risk of *over-fitting*. In Crouse et al.,[4] the total number of HMT parameters is reduced to approximately $4L$, with L the number of wavelet scales (typically 4–10). While a significant reduction, 40 parameters can still be a large number for some applications.

Often, the a priori image information takes the form of training data. Training algorithms such as the EM algorithm, especially for large data sets or data that have been severely corrupted by noise, can be computationally prohibitive. This makes the wavelet HMT models impractical for applications requiring computationally efficient processing. Furthermore, in many applications, training data is unavailable. In such cases, an empirical Bayesian approach can be taken and a model fit to the data at hand. While simple, if the observed data is corrupted (by noise, for example), then training may not be robust and the model parameters will not characterize the joint image pdf accurately.

In this paper, we propose two reduced-parameter HMT models that capture the general structure of a large class of real-world images. In the *image HMT (iHMT)* model we use the fact that for a large class of images the structure of the HMT is *self-similar* across scale. This allows us to reduce the complexity of the iHMT to just 9 easily trained parameters (independent of the size of the image and the number of wavelet scales). In the *universal HMT (uHMT)* we take a Bayesian approach and fix these 9 parameters. The uHMT requires no training of any kind.

While the iHMT and certainly the uHMT are less specific in their modeling of a particular image, they capture the statistics of a broad class of real-world images sufficiently for many applications. For example, we observe in

Table 1. Image estimation results for 256×256 images corrupted with additive white Gaussian noise of $\sigma_n = 0.05$. Entries are the the (negative) mean-square error (MSE) in dB, $MSE := -20\log_{10}(||\widehat{x} - x||_2/N)$. Pixel intensity vales were normalized between 0 and 1. All results use the Daubechies-8 wavelet. "Cspin-HMT" is the shift-invariant estimator from Section 5; "uHMT" uses the "universal" parameters presented in Section 4.5; "Emp-HMT" uses the empirical Bayesian estimator of Section 3.5; "RDWT-Thresh" uses a hard thresholded redundant wavelet transform using the thresholds in Lang et al.[5]; "DWT-Thresh" uses a thresholded orthogonal wavelet transform using the thresholds in Lang et al.[5]; and "Wiener2" is the 2-D spatially adaptive Wiener filter command from Matlab.

Image	Cspin-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

columns 2 and 3 of Table 1 that the image estimation (denoising) performance of the uHMT model is extremely close the more complicated HMT model. Furthermore, the simplicity of the uHMT model allows us to apply it in situations where the cost of HMT would be prohibitive. We will develop a shift-invariant estimation scheme in Section 5 below that offers state-of-the-art denoising performance, as seen in column 1 of Table 1 and the example in Fig. 4.

After reviewing the wavelet transform in Section 2 and the HMT model in Section 3, we introduce the iHMT and uHMT in Section 4. Section 5 develops the new redundant wavelet estimation technique. We close in Section 6 with a discussion and conclusions. More details on these new models can be found in Romberg et al.[6]

2. DISCRETE WAVELET TRANSFORM

The 2-D discrete wavelet transform (DWT) represents an image $x(s,t) \in L^2(\mathbb{R}^2)$ in terms of a set of shifted and dilated wavelet functions $\{\psi^{LH}, \psi^{HL}, \psi^{HH}\}$ and scaling function ϕ^{LL}.[7] When these shifted and dilated functions form an orthonormal basis for $L^2(\mathbb{R}^2)$, the image is decomposed as

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

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

$$u_{j_0,k,m} = \int_{\mathbb{R}^2} x(s,t)\phi_{j_0,k,m}\, dsdt \tag{2}$$

$$w^B_{j,k,m} = \int_{\mathbb{R}^2} x(s,t)\psi^B_{j,k,m}\, dsdt. \tag{3}$$

To keep the notation manageable, we will use an abstract index for the DWT coefficients and the basis functions: $w^B_{j,k,m} \to w_i$, and $\psi^B_{j,k,m} \to \psi_i$ 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 $\log_2 N = J > j \geq j_0$, 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 calculated using a 2-D separable filter bank[8] in $O(n)$ computations.

Recall the quadtree structure of the wavelet coefficients from Fig. 1. Each parent wavelet coefficient encompasses the same spatial location but lower frequency band than its four children. 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 region occupied by their ancestors.[9]

For the Haar wavelet,[7] the basis functions are disjoint square waves. In this case, the spatial divisions made by the wavelet quadtrees are exact (see Fig. 2(a)). If we use longer wavelets, there will be some overlap between adjacent wavelets at a given scale; however, the energy will still be concentrated in the $2^{-j} \times 2^{-j}$ size regions shown.

The orthogonal wavelet transform is not shift-invariant. In fact, the wavelet coefficients of two different shifts of an image can be very different,[5] with no easy 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) .[9] The RDWT is overcomplete, with $n \log n$ wavelet and scaling coefficients for a size n image.

3. 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 in Crouse et al.,[4] captures these properties 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 hidden states that dictate whether a coefficient is large or small. To capture the key dependencies between the wavelet coefficients, the HMT uses a probabilistic tree to model Markovian dependencies between the hidden states. By **S2** above, this graph has the same quad-tree topology as the wavelet transform.

3.1. 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: some 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 S_i dictate from which of the two components in the mixture model w_i is drawn. 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 taking 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) would be parameterized by a p_i^{L}, $\sigma^2_{\mathsf{S};i}$, and $\sigma^2_{\mathsf{L};i}$ for each wavelet coefficient w_i. Although independence is a reasonable zeroth order approximation to the structure of the wavelet coefficients, significant gains are realized by modeling dependencies.

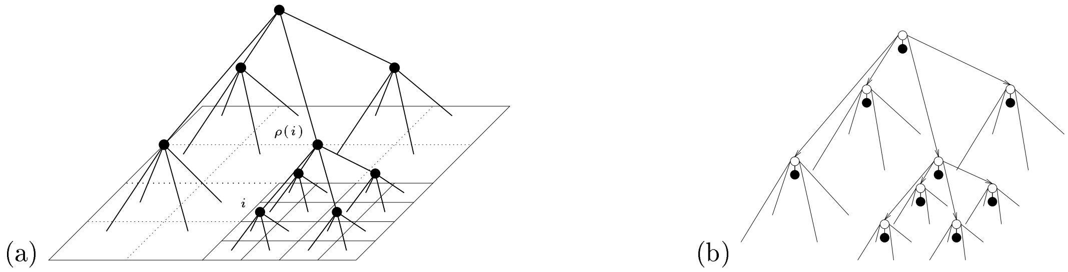


Figure 2. (a) Quad-tree organization of the wavelet coefficients in one subband. The four children wavelet coefficients divide the spatial localization of the parent coefficient. (b) 2-D HMT model. Each black node is a wavelet coefficient; each white node is the corresponding hidden state. Links represent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

3.2. Capturing Persistence: Markov Trees

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

In the HMT, these dependencies are captured 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. 2(b)). 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

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

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.* The parameters $p_i^{\mathsf{L}\to\mathsf{S}}$ and $p_i^{\mathsf{S}\to\mathsf{L}}$ are called the *novelty probabilities*, for they give the probabilities that the state values will change from one scale to the next. To have large and small coefficient values propagate down the quad-tree requires more persistence than novelty, that is, $p_i^{\mathsf{S}\to\mathsf{S}} > p_i^{\mathsf{S}\to\mathsf{L}}$ and $p_i^{\mathsf{L}\to\mathsf{L}} > p_i^{\mathsf{L}\to\mathsf{S}}$.

In the Introduction, we interpreted the wavelet basis functions as local edge detectors: if there is an edge inside the spatial support of the basis function, then the corresponding wavelet coefficient tends to be large. Since the same edge is within the spatial support of at least one of the child wavelet coefficient, the idea of persistency follows.

If, however, there are two edges inside the spatial support of a wavelet basis function, then their effect can cancel out, making the corresponding wavelet coefficient small. At some scale down the tree, however, the two edges will bifurcate. This is because the spatial resolution will be fine enough so that each edge is represented by its own wavelet coefficient. These wavelet coefficients will be large even though their parent is small. This is the idea behind novelty.

3.3. 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. a 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 parametric model for the joint probability density function $f(\mathbf{w}|\mathbf{\Theta})$.

In general, these parameters can be different for each wavelet coefficient. However, this could make the model too complicated for some applications. For example, if there is only one observation of an image, then there are more parameters to estimate than data points and over-fitting is certain to occur.

To reduce HMT complexity, each parameter can be assumed to be the same at each scale of the wavelet transform:

$$\sigma^2_{\mathsf{S};B,j,k,m} = \sigma^2_{\mathsf{S};j}, \tag{10}$$

$$\sigma^2_{\mathsf{L};B,j,k,m} = \sigma^2_{\mathsf{L};j}, \tag{11}$$

$$A_{B,j,k,m} = A_j \quad \forall k,m \in \mathbb{Z},\ \forall B \in \mathcal{B}. \tag{12}$$

This is called *tying within scale*. Parameter invariance within scale makes a tied HMT model less image-specific. A priori, the tying within scale keeps the model from expecting smooth regions or edges at certain spatial locations.

3.4. HMT Algorithms

The HMT is a tree-structured hidden Markov model (HMM). Thus, the three standard problems of HMMs[10] apply equally well to the HMT.

3.4.1. Likelihood Determination

While the HMT is an elegant way of modeling the joint pdf of the wavelet coefficients, there is no closed form expression for $f(\mathbf{w}|\mathbf{\Theta})$. Fortunately, there is a fast $O(n)$ algorithm to compute $f(\mathbf{w}|\mathbf{\Theta})$ for a given $\mathbf{w}$ and $\mathbf{\Theta}$ called the *upward-downward algorithm*.[4,10–12]

3.4.2. State Path Estimation

Given a set of observations $\mathbf{w}$ and a model $\mathbf{\Theta}$, we would like to determine the probability that node i is in a given state, and the most likely sequence of hidden states. Using by-products of the upward-downward algorithm, we can calculate the probability $p(S_i = q|\mathbf{w}, \mathbf{\Theta})$ that an observed wavelet coefficient w_i has corresponding hidden state $q \in \{\mathsf{S}, \mathsf{L}\}$. The hidden state sequence problem is solved by the *Viterbi algorithm*.[10,11]

3.4.3. Model Training

In many situations, the parameters $\mathbf{\Theta}$ for the HMT model are a priori unknown but we have a set of training data. To train, we find the most likely $\mathbf{\Theta}$ that could give rise to the training observations $\mathbf{w}$:

$$\widehat{\mathbf{\Theta}}_{\mathrm{ML}} = \arg\max_{\mathbf{\Theta}} f(\mathbf{w}|\mathbf{\Theta}). \tag{13}$$

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

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

While simple, EM training for the HMT has several drawbacks. First, the convergence is relatively slow. For large images, this can make training very computationally expensive. Each iteration of the EM algorithm is $O(n)$, but there is nothing to limit the number of iterations it takes to converge. Second, being a hill-climber, the EM algorithm is guaranteed to convergence only to a local maximum of $f(\mathbf{w}|\mathbf{\Theta})$.

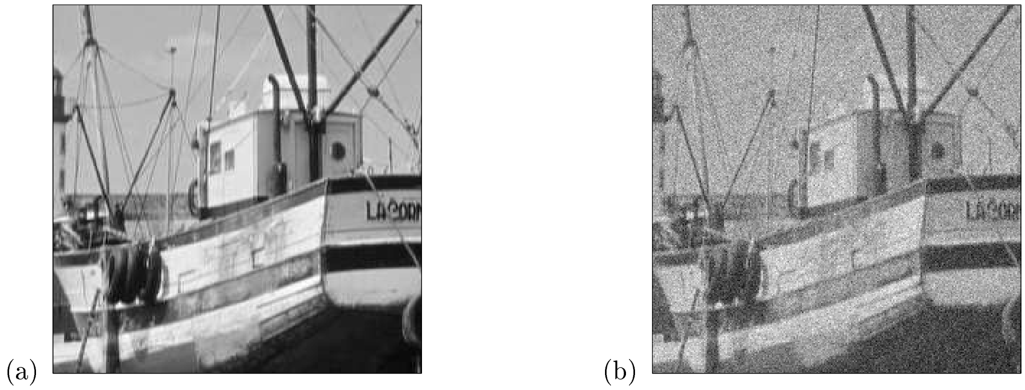


Figure 3. (a) Original 256×256 "Boats" image. (b) Noisy boats image, with $\sigma_n = 0.1$, MSE=20dB.

3.5. Application: Empirical Bayesian Estimation

To demonstrate the effectiveness of the HMT for modeling an image's wavelet coefficients, we estimate an image submerged in additive white Gaussian noise. This is an extension to 2-D of the work in Crouse et al.[4] Translated into the wavelet domain, the problem is as follows:

$$\text{given} \quad \mathbf{y} = \mathbf{w} + \mathbf{n}, \quad \text{estimate} \quad \mathbf{w}, \tag{14}$$

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 $\mathbf{\Theta}$, Crouse et al.[4] takes 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 4.5 below) or assign a hyper-prior to them.[13,14]

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

$$\widehat{w}_i = E[w_i|\mathbf{y}, \mathbf{\Theta}] = \sum_q p(S_i = q|\mathbf{y}, \mathbf{\Theta}) \frac{\sigma_{q;i}^2}{\sigma_n^2 + \sigma_{q;i}^2} y_i \tag{15}$$

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 1-3, and an example is shown in Figs. 3–5. While the HMT empirical Bayesian estimator outperforms other DWT wavelet shrinkage techniques in terms of MSE, in visual terms it is far superior, with sharper and more accurate edges. In fact, its MSE and visual performance are quite competitive with RDWT wavelet shrinkage.

4. REDUCED-PARAMETER HMT IMAGE MODELS

By design, the HMT described in the previous section captures the main features of the wavelet transforms of real-world images. In its raw form, however, the number of parameters needed to model an image can make it unwieldy, even when tying within scale. To model an image, the HMT requires at least $4L$ different parameters where L is the number of scales in the wavelet decomposition. This poses problems. Directly specifying $4L$ parameters would require a tremendous amount of a priori information about the image, and without this information, we would 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 3.5 works well, but requires the use of the EM

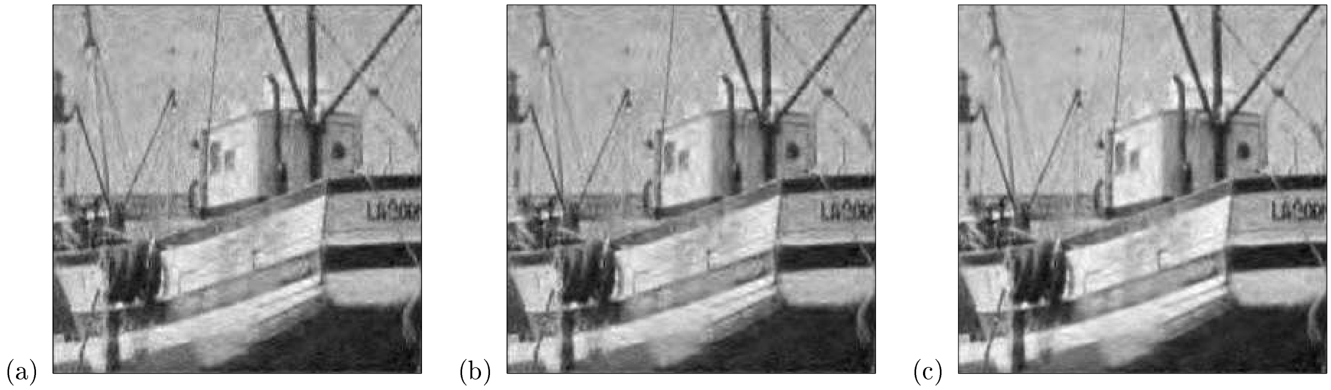


Figure 4. Boats image denoised using (a) the empirical Bayesian HMT estimator of Section 3.5,[4] MSE=26.5dB; (b) the uHMT of Section 4.5, MSE=26.4dB; (c) the shift-invariant uHMT estimation algorithm of Section 5, MSE=27.4.

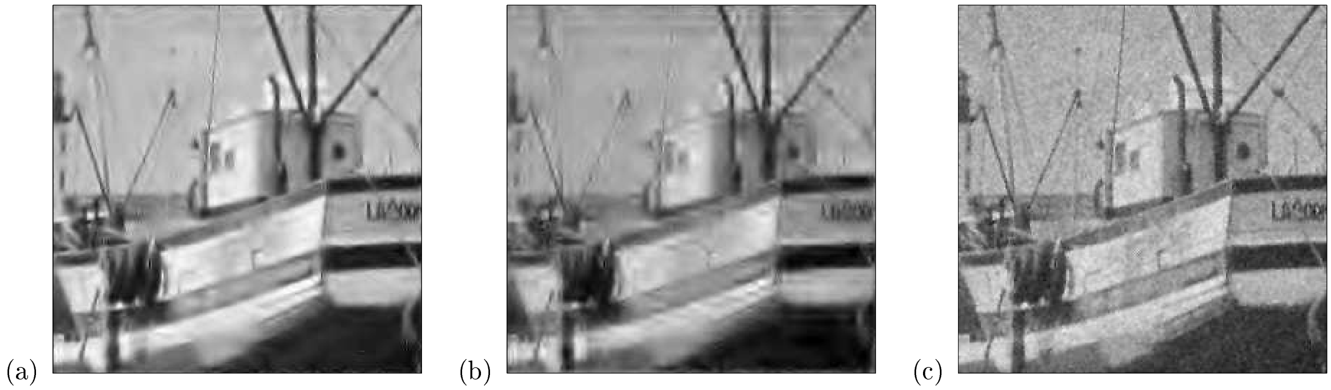


Figure 5. Boats image denoised using (a) hard thresholded redundant wavelet transform (RDWT) with threshold chosen to minimize the MSE, MSE=26.3dB; (b) soft thresholded discrete wavelet transform with empirical best threshold, MSE=22.5dB; (c) spatial domain 3x3 Wiener filter, MSE=26.1dB.

algorithm, which at $O(n)$ computational complexity for each iteration, can be very time consuming. This makes the HMT inappropriate for applications requiring rapid processing.

To address these problems, we need to reduce the number of parameters it takes to specify an HMT model. In doing this, the HMT model will become less accurate; two images that have different parameterizations with the general form of the HMT may have the same parameterization in a reduced-parameter model. What the model gains is a reduction in complexity; less a priori information is needed to specify model parameters and training becomes more robust.

In Crouse et al.[4] and in Section 3, our modeling paradigm was to assume that every image has a different HMT model, with the $4L$ parameters being specified by training on an observation. In this section we take a different approach. We specify a new HMT model, called the iHMT, with a drastically reduced set of parameters (only 9), that incorporates properties common to all images in a class.

An iHMT model with a given set of parameters specifies a class of images that is more general than that specified by an HMT model. Images in an iHMT class have similar smoothness as characterized by the rate at which the wavelet coefficients decay, and similar "edge strength," as characterized by the rate at which the state transition matrix goes to its asymptotic form. Different images in an iHMT class are not distinguished between a priori; we consider them statistically similar.

As fewer parameters are used to specify a model, the model becomes less image-specific; the class of images a given parameter set models becomes larger. The amount of parameter reduction that is appropriate depends on the application and the amount 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 "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 or 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 be described by more parameters.

4.1. Tertiary Properties of the Wavelet Coefficients

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 models, we will leverage the following *tertiary* properties of wavelet transform:

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 fine scales: The persistence of large/small wavelet coefficient magnitudes become stronger at finer scales.

The exponential decay property (**T1**) stems from the overall smoothness and self-similarity 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.

We can obtain intuition behind property **T2** by considering the simple yet powerful image model of Cohen and D'Ales.[15] 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 $j_{\rm crit}$ such that for all $j \geq j_{\rm crit}$, 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 resolution of the wavelet coefficients get finer exponentially implies that the probability that every edge is isolated goes to 1 exponentially.

By **P4**, for fine scales such that $j \gg j_{\rm crit}$ 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.

In 2-D, the situation is similar except that instead of a discontinuities at points, we now have discontinuities along curves. At $j_{\rm crit}$, 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 small children.

Similar ideas were used to construct the *lacunary wavelet series* model of Jaffard[16] and Baraniuk.[17]

4.2. The iHMT model

Based on the tertiary properties of the wavelet transforms of real-world images, we can specify the HMT model parameters in a hyper-parametric form. The coefficient decay and the 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 wavelet-domain.

4.2.1. Modeling wavelet coefficient decay

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 thus require that they decay exponentially across scale as well. We write

$$\begin{aligned} \sigma^2_{\mathsf{S};j} &= C_{\sigma_\mathsf{S}} 2^{-j2\alpha_\mathsf{S}}, && (16) \\ \sigma^2_{\mathsf{L};j} &= C_{\sigma_\mathsf{L}} 2^{-j2\alpha_\mathsf{L}}. && (17) \end{aligned}$$

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 a $1/f$ power spectrum.

4.2.2. Modeling coefficient persistence

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 are isolated, 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 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. To help understand, consider a wavelet coefficient z lying over an isolated edge ($S_z = \mathsf{L}$) in the 1-D slice at scale $j \geq j_{\mathrm{crit}}$. Call z's children a and b. Since the edge is perfectly localized in space, one and only one of a and b will be large. This means that

$$
\begin{aligned}
p(S_a = \mathsf{L}, S_b = \mathsf{S} | S_z = \mathsf{L}) &= 1/2, & (18)\\
p(S_a = \mathsf{S}, S_b = \mathsf{L} | S_z = \mathsf{L}) &= 1/2, & (19)\\
p(S_a = \mathsf{S}, S_b = \mathsf{S} | S_z = \mathsf{L}) &= 0, & (20)\\
p(S_a = \mathsf{L}, S_b = \mathsf{L} | S_z = \mathsf{L}) &= 0. & (21)
\end{aligned}
$$

This condition is slightly problematic. 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 a and b is large. In fact, given S_z, under the HMT S_a and S_b are independent. Instead of modeling (18)–(21) exactly, the HMT only models the marginals

$$
\begin{aligned}
p(S_a = \mathsf{L} | S_z = \mathsf{L}) = p(S_a = \mathsf{L}, S_b = \mathsf{S} | S_z = \mathsf{L}) + p(S_a = \mathsf{L}, S_b = \mathsf{L} | S_z = \mathsf{L}) &= 1/2, & (22)\\
p(S_b = \mathsf{L} | S_z = \mathsf{L}) = p(S_a = \mathsf{S}, S_b = \mathsf{L} | S_z = \mathsf{L}) + p(S_a = \mathsf{L}, S_b = \mathsf{L} | S_z = \mathsf{L}) &= 1/2. & (23)
\end{aligned}
$$

As a result, the HMT persistency probability $p_j^{\mathsf{L}\to\mathsf{L}} \to 1/2$ as $j \to \infty$. This is far from a perfect model, since for all values of j, there is a chance that the edge will disappear (since $p(S_a = \mathsf{S}, S_b = \mathsf{S} | S_z = \mathsf{L}) = 1/4$ under the HMT) or bifurcate[2] (since $p(S_a = \mathsf{L}, S_b = \mathsf{L} | S_z = \mathsf{L}) = 1/4$).

Extension of this analysis to 2-D is not exact, except for horizontal, vertical, and diagonal edges. In 2-D, edges lie on curves in space, and the curve could intersect the spatial support of the basis functions of one, two, or three children of a coefficient that has isolated the curve.

The localization probability going to 1 exponentially means that the asymptotic values for persistency and novelty parameters are approached exponentially. This gives the state transition matrix (see (9) specified by four parameters:

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

The transition matrix has asymptotic form

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

4.2.3. iHMT 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 have reduced the number of parameters that specify the iHMT model to 9:

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

4.3. The iHMT and Besov Space

In the last section, we provided the intuition behind a certain form for the HMT parameters of real-world images. In this section, we will relate this form to the Besov space characterization of the smoothness of images by showing that realizations of the iHMT will lie in a certain Besov space with probability 1.

Roughly speaking, a Besov space $B_q^s(L^p)$ contains functions with s derivatives measured in L^p, with q making finer smoothness distinctions.[18] For $s < 1$, $B_q^s(L^p)$ contains functions that are uniformly regular but have isolated discontinuities.[2] These properties are similar to those of real-world images; we expect images to be members of a Besov space.

The fact that wavelets form an unconditional basis for a Besov space $B_q^s(L^p)$ means the Besov norm can be computed equivalently as a simple sequence norm on the wavelet coefficients[19]:

$$||x||_{B_q^s(L^p)} \asymp c_0 ||u_{j_0,k,m}||_p + \left[\sum_{j \geq j_0} 2^{js'q} \left(\sum_{k,m} |w_{j,k,m}|^p \right)^{q/p} \right]^{1/q} \tag{27}$$

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

The following theorem (proved in Romberg et al.[6]) tells us how smooth we can expect realizations of the iHMT model to be, that is, to which Besov spaces do the images modeled by the iHMT belong.

THEOREM 4.1. *Let* $\mathbf{w} \sim f(\mathbf{w}|\mathbf{\Theta})$ *be a realization of an independent Gaussian mixture model (see Section 3.1) with parameters* $\mathbf{\Theta} = \{p_{\mathsf{L};j}, \sigma^2_{\mathsf{S};j}, \sigma^2_{\mathsf{L};j}\}$. *Let* $\mathbf{x}$ *as be the inverse wavelet transform of* $\mathbf{w}$. *Let* $\sigma_{\mathsf{L};j} = C_{\sigma_\mathsf{L}} 2^{-\alpha_\mathsf{L} j}$, $\sigma_{\mathsf{S};j} = C_{\sigma_\mathsf{S}} 2^{-\alpha_\mathsf{S} j}$, *and* $p_j^\mathsf{L} \to 0$ *exponentially. Then* $\mathbf{x} \in B_q^s(L^p)$ *with probability* 1 *for* $p, q < \infty$ *if and only if* $\alpha > s + 1/2$.

If we have knowledge that an image we wish to model is in a particular Besov space, we can use this information to determine the α_L and α_S. However, as we can see from the Theorem, membership in a Besov space does not imply any kind of dependency structure. The Besov norm does not account for persistence across scale; it is invariant to wavelet coefficients at a given scale being "shuffled." Therefore, an image being in a Besov space does not tell us anything about the state transition matrix A_j. An image being modeled by the iHMT is a stronger condition than membership in a Besov space; the iHMT accounts for the persistency properties **S2** and **T3**, while the Besov norm (27) does not. By specifying persistency probabilities, we are constraining the likely realizations of our model to a subset of the Besov space that encompasses "real-world images" more tightly.

4.4. A "Universal" iHMT: The uHMT

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

Another possibility is to fix the parameters directly. This yields an iHMT model for a class of images, with each member in the class being treated as statistically equivalent.

To see how much variation in iHMT parameters there is across photograph-like images, we trained HMT models for a set 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 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 we can use a specific, "universal" set of iHMT parameters to reasonably characterize photograph-like images.

Although we clearly lose accuracy by viewing all images we are interested in as statistically equivalent, we have totally eliminated the need for training. This can save us a tremendous amount of computation. For example, the EM algorithm on a 512×512 image can take anywhere from minutes to hours to converge on a workstation (depending on the amount of noise).

4.5. Application: Bayesian Estimation with the uHMT

With the "universal" iHMT parameters, we have a prior on the w_i and the estimation problem in Section 3.5 is approached from a purely Bayesian standpoint. To find the conditional mean vector, the state probabilities $p(S_i = q|\mathbf{y}, \mathbf{\Theta}_{\mathrm{i}})$ are calculated using the upwards-downwards algorithm and used to evaluate (15). Since we have eliminated training, the estimation algorithm is truly $O(n)$ and takes only a few seconds to run on a workstation.

To test this new Bayesian estimator, we denoised the above set of test images using the uHMT with parameters: $\alpha_L = \alpha_S = 5/4$, $C_{\sigma_{\mathsf{S}}} = 2^7$, $C_{\sigma_{\mathsf{L}}} = 2^{13}$, $\gamma_{\mathsf{S}} = \gamma_{\mathsf{L}} = 1$, $C_{\mathsf{SS}} = C_{\mathsf{LL}} = 32/5$, and $p_0^{\mathsf{L}} = 1/2$.

The estimation results are summarized in Tables 1–3, and an example is given in Fig. 4(b). The results are almost identical to the much more complicated empirical Bayes HMT approach, suggesting that we have lost almost nothing by totally eliminating training.

5. SHIFT-INVARIANT HMT IMAGE ESTIMATION

An image estimate obtained using an orthogonal wavelet transform (DWT) often exhibits visual artifacts, usually in the form of ringing around the edges. These artifacts result from the DWT not being shift-invariant. As we mentioned before, two different shifts of an image can have very different wavelet transforms. In particular, a singularity at two different shifts can have very different characteristics.

For a shift-invariant wavelet representation, we turn to the redundant wavelet transform (RDWT). Ideally, we would like to model the RDWT coefficients 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 graph that would be hard or impossible to do Bayesian inference on.

Another way to make the image estimate shift-invariant is to follow the "cycle-spinning" programme proposed by Coifman and Donoho.[20] The estimation algorithm is applied to all shifts of the noisy image, and the results are averaged. The shift-invariant estimate of an image $\mathbf{x}$ that has been corrupted by noise, $\mathbf{v} = \mathbf{x} + \mathbf{n}$, is given by

$$\widehat{x} = \mathrm{Average}(\mathbf{S}_{-k,-m}(\mathbf{D}(\mathbf{S}_{k,m}(v))))_{0 \le k,m \le N-1} \tag{28}$$

where $\mathbf{S}_{k,m}(v) = v(s-k, t-m)$ is the 2-D shift operator and $\mathbf{D}$ denotes the estimator in Section 3.5 or Section 4.5.

This approach fits into the Bayesian framework quite nicely. The shift (K, M) can be viewed as an unknown random variable. Since we have no a prior information about (K, M) except that $0 \le K, M \le N-1$, 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[13]

$$\widehat{x} = \sum_{k,m} p(k,m|y)\mathbf{S}_{-k,-m}(\mathbf{D}(\mathbf{S}_{k,m}(q))). \tag{29}$$

Because it is not clear how to infer any information about the underlying shift given the observed data, $p(k,m|y)$ is also taken as uniform, and (28) follows.

The algorithm (28), if implemented directly, would be computationally expensive, $O(n^2)$. However, the complexity can be reduced substantially by realizing that the same wavelet coefficients appear in multiple shifts, and assuming that the HMT model is same for all shifts. This assumption is implicit in the definition of the uHMT parameters. If empirical methods are used, models from multiple shifts must be combined. Since there are only $n \log n$ unique coefficients between all possible shifts, the resulting algorithm is $O(n \log n)$.

The RDWT coefficients can be arranged in such a way that the coefficients shared between the DWT trees at two different shifts occupy an entire subtree. A 1-D example of this overlap is shown in Fig. 6. By looking at the figure, we see that each node now has two children and two parents, each parent coming from a different DWT tree (in the 2-D case, each node has 4 parents and 4 children). Unfortunately, this dependency graph is not singly connected, so it is hard, maybe even impossible, to find exact inference[11] or training algorithms.[21] The graph still has significant structure, however, and it is this structure that will allow us to quickly compute a cycle-spin estimate (28).

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 the different trees that include it. Since each node still has two children, the

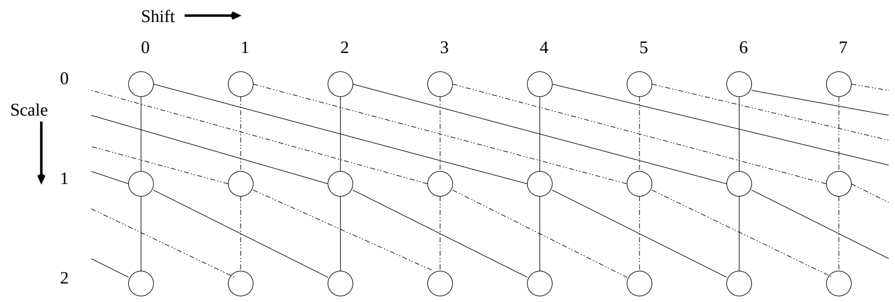


Figure 6. 1-D example of the overlapping of DWT trees in the RDWT domain. The shift of the signal at which the DWT tree was obtained is noted along the top of the figure, while the scale is noted along the side. Note that each node now has two parents as well as two children, and is included in 2^j different trees. In 2-d, the RDWT consists of overlapping quad-trees; each node has four parents and four children and is included in 4^j different trees.

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

Image	Cspin-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

downwards binary tree structure has been preserved, and an $O(n \log n)$ algorithm can be obtained by a modification to the upward-downward algorithm.[6]

Our results using the uHMT parameters from Section 4.5 in the shift-invariant estimator are summarized in the first column of Tables 1–3, with an example shown in Fig. 4(c). As can be seen from the figure, the shift-invariant transform has smoothed 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.

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. The resulting model, the iHMT, has only 9 parameters.

A set of "universal" parameters — the uHMT parameters that model a wide range of images accurately — arise naturally from the form of the iHMT. These nine numbers completely specify a model for a large class of real-world images, eliminating any need for training in the estimation algorithm.

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.

Acknowledgments: Thanks to Mithat "TBG" Gönen for giving us a favorable prior on Bayesian statistics.

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

Image	Cspin-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

REFERENCES

1. S. Geman and D. Geman, "Stochastic relaxation, gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. on Pattern Anal. Machine Intell.* **6**, pp. 721–741, 1984.
2. S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, 1998.
3. M. Vetterli and J. Kovačević, *Wavlets and Subband Coding*, Prentice Hall, Englewood Cliffs:NJ, 1995.
4. M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Proc.* **46**, pp. 886–902, April 1998.
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. J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden markov models," *preprint* .
7. I. Daubechies, *Ten Lectures on Wavelets*, SIAM, New York, 1992.
8. S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, pp. 674–693, July 1989.
9. C. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms: A Primer*, Prentice Hall, 1998.
10. L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE* **77**, pp. 257–285, Feb. 1989.
11. B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*, MIT Press, Cambridge, MA, 1998.
12. O. Ronen, J. R. Rohlicek, and M. Ostendorf, "Parameter estimation of dependence tree models using the EM algorithm," *IEEE Sig. Proc. Letters* , August 1995.
13. M. A. T. Figueiredo and R. D. Nowak, "Bayesian wavelet-based signal estimation using non-informative priors," in *Proc. 32nd Asilomar Conf.*, 1998.
14. H. A. Chipman, E. D. Kolaczyk, and R. E. McCulloch, "Adaptive Bayesian wavelet shrinkage," *J. Amer. Stat. Assoc.* **92**, 1997.
15. A. Cohen and J. P. D'Ales, "Nonlinear approximation of random functions," *SIAM J. Appl. Math.* **57**, April 1997.
16. S. Jaffard, "On lacunary wavelet series," *preprint* .
17. R. G. Baraniuk, "Tree-structured lacunary wavelet series," *preprint* .
18. R. A. DeVore, B. Jawerth, and B. J. Lucier, "Image compression through wavelet transform coding," *IEEE Trans. on Information Theory* **38**, pp. 719–746, March 1992.
19. Y. Meyer, *Wavelets and Operators*, Cambridge Univ. Press, Cambridge, 1992.
20. R. Coifman and D. Donoho, "Translation-invariant de-noising," in *Wavelets and Statistics*, Lecture Notes in Statistics, Springer-Verlag, 1995.
21. H. Lucke, "Which stochastic models allow Baum-Welch training?," *IEEE Trans. Signal Processing* **44**, November 1996.