HIDDEN MARKOV TREE MODELING OF COMPLEX WAVELET TRANSFORMS

Hyeokho Choi, Justin Romberg, Richard Baraniuk

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

Nick Kingsbury

Dept. of Engineering
University of Cambridge, Trinity College
Cambridge, CB2 1TQ, UK

ABSTRACT

Multiresolution signal and image models such as the *hidden Markov tree* aim to capture the statistical structure of smooth and singular (edgy) regions. Unfortunately, models based on the orthogonal wavelet transform suffer from shift-variance, making them less accurate and realistic. In this paper, we extend the HMT modeling framework to the *complex wavelet transform*, which features near shift-invariance and improved angular resolution compared to the standard wavelet transform. The model is computationally efficient (with linear-time computation and processing algorithms) and applicable to general Bayesian inference problems as a prior density for the data. In a simple estimation experiment, the complex wavelet HMT model outperforms a number of high-performance denoising algorithms, including redundant wavelet thresholding (cycle spinning) and the redundant HMT.

1. INTRODUCTION

The compact representation of the discrete wavelet transform (DWT) has led to many successful signal and image processing algorithms, including wavelet-based estimation (denoising), detection and classification, segmentation, fusion, and synthesis. The transform's function as a multiscale edge detector provides the DWT coefficients with two key properties for a large class of signals and images [1]:

2Populations: Smooth signal/image regions are represented by small wavelet coefficients, while edges, ridges, and other singularities are represented by large coefficients.

Persistence: Large and small wavelet coefficient values cascade along the branches of the wavelet tree.

Interpreting these properties statistically, 2Populations implies that wavelet coefficients have non-Gaussian marginal statistics, with a large peak at zero (many small coefficients) and heavy tails (a few large coefficients), while Persistence implies that wavelet coefficient values are statistically dependent along the branches of the wavelet tree.

The energy compaction implied by the 2Populations property has inspired many simple schemes for removing additive white Gaussian noise from signals and images. Since most signal energy will concentrate in a few (large) wavelet coefficients while the noise energy will disperse over all coefficients, a simple thresholding operation in the wavelet domain will remove most of the noise [2].

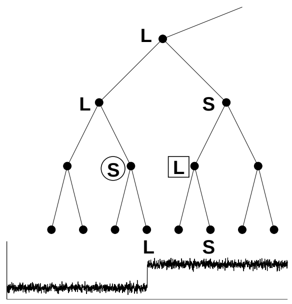


Figure 1: *Multiscale DWT wavelet tree for a 1-d noisy step edge. Ideally, an edge yields a cascade of large wavelet coefficients across scale (down the tree). The squared* L *and circled* S *coefficients result due to the additive noise. For good denoising performance, the circled* S *coefficient should be kept while the squared* L *coefficient should be discarded.*

While highly successful, wavelet thresholding algorithms suffer from two subtle problems. First, the noise can cause large coefficients from edge regions to become small, and not retaining them blurs the edges of the denoised image. Second, the noise can cause small coefficients from smooth regions to become large, and retaining them produces wavelet-shaped ripples in otherwise smooth regions of the denoised data (see Fig. 1).

Classical wavelet thresholding algorithms treat each wavelet coefficient independently. By capturing the dependencies between the coefficients espoused by the Persistence property, we can improve denoising performance significantly. Among the many different approaches to modeling the dependencies [3,4], the wavelet-domain *hidden Markov tree* (HMT) model [1] is most appropriate for our purposes.

The HMT models the joint probability density function (pdf) $f(\mathbf{w})$ of the wavelet coefficients $\mathbf{w}$ as a tree-structured Gaussian mixture. To capture the 2Populations property, we model each wavelet coefficient w_i as a two-state Gaussian mixture; a controlling hidden state s_i records the magnitude of the coefficient as either S (small) or L (large). To capture the Persistence property, we connect the hidden states across scale (down the wavelet tree) in a Markov-1 chain in such a way that the unbroken progressions L-L-L-L and S-S-S-S are more likely than broken ones like L-L-S-L and S-S-L-S). These latter two progressions lead to edge blurring and ripple artifacts, respectively (recall Fig. 1.)

HC, JR, RB supported by NSF grants MIP–9457438 and CCR–9973188, ONR grant N00014–99–1–0813, DARPA/AFOSR, grant F49620–97–1–0513, and Texas Instruments. Email: {choi, jrom, richb}@ece.rice.edu; Internet: www.dsp.rice.edu

NK supported by the Rice-Trinity Exchange Program and the UK Engineering and Physical Sciences Research Council (EPSRC). Email: ngk@eng.cam.ac.uk; Internet: www.eng.cam.ac.uk/∼ngk

When employed in a Bayesian framework as a prior density on the wavelet coefficients of the noise-free data, the HMT model plays the rôle of a statistical *grammar* that rules that unbroken strings of L and S states are preferable to broken strings. Numerous experiments with real data have verified that the HMT is an accurate and effective model for a large class of real-world signals and images [1, 5, 6].

Unfortunately, due to the strong shift-variance of the orthogonal DWT, the 2Populations and Persistence properties are strongly affected by the placement of the smooth and edgy regions in the data. In particular, the coefficient magnitudes near an edge can swing wildly as the alignment between the edge and the wavelet basis functions changes slightly (even by one sample). Clearly this will lead to both blurred edges and ripple artifacts in denoised data.

The redundant (undecimated) wavelet transform [7] comprises the coefficients of the orthogonal DWTs of each shift of the signal or image. Redundant wavelet denoising algorithms (including [7], "cycle spinning" [8], and the undecimated HMT approach [5]) are the current performance standards, both in mean-square-error and visual metrics. The price paid is an increase in computational complexity from $O(N)$ to $O(N \log N)$ for an N-point data set.

In this paper we take a different approach based on the dual tree *complex wavelet transform* (CWT) [9, 10]. The CWT expands a signal or image in terms of a set of complex wavelets with complementary real and imaginary parts. The most important property for the present study is the near shift invariance of the CWT coefficient *magnitudes*, which leads to strong persistence of large magnitudes around edges, independent of their alignment. As an added bonus, in 2-d the CWT offers six angle-selective subbands rather than the standard DWT's three. Finally, the computational cost of the CWT remains $O(N)$.

We will develop an HMT model for the CWT that outperforms the standard HMT model and undecimated HMT model (and at a significantly reduced cost). After reviewing the HMT and CWT in Sections 2 and 3, we derive the new model in Section 4. In Section 5, we demonstrate a denoising algorithm based on the new model that, while simple, outperforms most techniques in the literature. Section 6 concludes with a discussion of other potential applications of the new model, including detection/classification, segmentation, and synthesis.

2. WAVELET HMT MODEL

For concreteness, we will emphasize images in this paper, though the ideas apply to data of any dimension.

2.1 Model: The 2-d orthogonal DWT consists of three subbands of coefficients [11], each of which has a quad-tree structure like that shown in Fig. 2(a). Denote the *parent* of wavelet coefficient w_i as $w_{\rho(i)}$.

To statistically match the 2Populations property of the wavelet transform, we consider the collection of small/large wavelet coefficients as outcomes of a Gaussian pdf with small/large variance. This approximates the marginal pdf $f(w_i)$ of each wavelet coefficient as a *Gaussian mixture*. To each wavelet coefficient w_i, we associate a discrete hidden state s_i that takes on values $m = \mathsf{S}, \mathsf{L}$ with probability mass function (pmf) $p(s_i)$. Conditioned on $s_i = m$, w_i is Gaussian with mean $\mu_{i,m}$ and variance

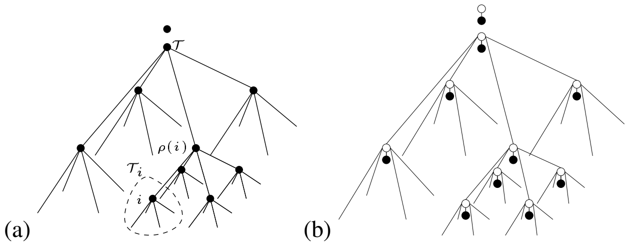


Figure 2: *(a) Quad-tree of wavelet coefficients w_i (black nodes) for one subband of the 2-d DWT. (b) 2-d HMT model (white nodes represent hidden state variables s_i).*

$\sigma^2_{i,m}$. Thus, its overall marginal pdf is given by[1]

$$f(w_i) = \sum_{m=\mathsf{S},\mathsf{L}} p(s_i = m)\, f(w_i | s_i = m), \qquad (1)$$

with $f(w_i | s_i = m) \sim N(\mu_{i,m}, \sigma_{i,m})$. While each wavelet coefficient w_i is conditionally Gaussian given its state variable s_i, overall it has a non-Gaussian density due to the randomness of s_i.

To match the Persistence property of large and small coefficients, we set up a Markov-1 dependency structure on the hidden states across scale (a hidden Markov model) [1]. Using links to denote dependencies, the HMT model takes the form of the tree in Fig. 2(b). The state transition probabilities between the connected states model the persistence of large/small coefficient magnitudes across scale. The overall HMT model consists of an HMT for each of the three DWT subbands.

The HMT model parameters consist of the Gaussian mixture means and variances $\mu_{i,m}, \sigma^2_{i,m}$, the transition probabilities $p(s_i | s_{\rho(i)})$, and the pmf $p(s_0)$ for the root state s_0. Group these into the parameter vector $\mathbf{\Theta}$. In practice, we usually practice *tying*, where we share the same HMT parameters within each scale band [1]. A reduced-parameter HMT that exploits the self-similarity of image data is derived in [5].

The HMT can be trained to capture the wavelet-domain features of the image(s) of interest using the iterative expectation-maximization (EM) algorithm [1]. For a given set of training data, the trained model $f(\mathbf{w}|\mathbf{\Theta})$ approximates the joint pdf of the wavelet coefficients. See [1, 5, 6] for more details.

2.2 Model-based Estimation: With a wavelet-domain HMT model at hand, many statistical image processing tasks, such as estimation, classification, and segmentation, are straightforward [1, 5, 6]. For example, the HMT can provide the prior for an empirical Bayesian image estimation algorithm.

Consider the problem of estimating the noise-free wavelet coefficients $\mathbf{w}$ given the coefficients $\mathbf{y}$ corrupted by additive white Gaussian noise of variance σ_n. It is straightforward to train an HMT for the clean coefficients using the noisy coefficients [1]. Because of the simple structure of the Gaussian mixture prior, the conditional mean estimate for the noise-free wavelet coefficient w_i is then given by

$$\widehat{w}_i := \mathsf{E}\left[y_i | \mathbf{w}, \mathbf{\Theta}\right] = \sum_{m=\mathsf{S},\mathsf{L}} p(s_i = m | \mathbf{w}, \mathbf{\Theta}) \frac{\sigma^2_{i,m}}{\sigma^2_n + \sigma^2_{i,m}} w_i, \qquad (2)$$

[1] The number of states can be increased beyond two to match the marginal distribution with arbitrary precision. However, we have found two states a fine accuracy/complexity balance.

a linear combination of two Wiener shrinkage filters. The all-important posterior state probabilities $p(s_i|\mathbf{w},\mathbf{\Theta})$ are produced as byproducts of the EM training algorithm.

HMT-based denoising typically outperforms standard thresholding techniques, because the degree of coefficient shrinkage is determined based not only on the magnitude of the coefficient but also on its relationship with its neighbors across scale.

Proper determination of the posterior state probabilities $p(s_i|\mathbf{w},\mathbf{\Theta})$ is the key to the performance of the algorithm. To avoid variations due to shift variance, these quantities should ideally be estimated based on a number of different shifts of the data. By considering all possible shifts, we obtain the shift-invariant, undecimated HMT model of [5]. However, the complexity rises with the number of incorporated shifts. The dual tree CWT provides a computationally attractive alternative.

3. COMPLEX WAVELET TRANSFORM

The dual tree CWT comprises two parallel wavelet filter bank trees that contain carefully designed filters of different delays that minimize the aliasing effects due to decimation [9, 10]. Each tree produces a valid set of *real* DWT coefficients u_i and v_i; together they form the *complex coefficients* $c_i = u_i + jv_i$.[2] The CWT is near shift invariant [9] and, in 2-d, achieves good directional selectivity with six oriented subbands per scale [10].

In 2-d, the CWT basis functions closely approximate modulated complex exponentials (Gabor functions) of the form

$$h(x, y) = a(x, y)\, e^{j(\omega_x x + \omega_y y)} \tag{3}$$

with $a(x, y)$ a slowly varying Gaussian-like real window function centered on $(0, 0)$, and (ω_x, ω_y) the center frequency of the corresponding subband. The real and imaginary parts of the CWT basis functions are symmetric and antisymmetric respectively about $(0, 0)$ and are therefore orthogonal. This means that the real and imaginary parts of each CWT coefficient are statistically uncorrelated.

Furthermore, since $a(x, y)$ is slowly varying, the *magnitude* $|c_i| = \sqrt{u_i^2 + v_i^2}$ of each CWT coefficient is insensitive to small image shifts. It therefore forms a more accurate estimate of image activity at a given location and scale than the corresponding coefficient of a real DWT, which will suffer from shift variance.

The directional properties of the CWT arise from the fact that $h(x, y)$ has constant phase along lines such that $\omega_x x + \omega_y y$ is constant, giving filters selective to edges parallel to this direction. For the six subbands at each scale, $\omega_y/\omega_x \approx \{-3, -1, -\frac{1}{3}, \frac{1}{3}, 1, 3\}$ which produces bands oriented at approximately all odd multiples of $15°$.

4. COMPLEX WAVELET HMTS

Due to its near shift invariance, the progressions of large and small magnitudes in the CWT coefficients $\mathbf{c}$ are more reliable than in either the real part $\mathbf{u}$ or imaginary part $\mathbf{v}$ alone. In this section, we will leverage this reliability into a robust HMT model for the CWT.

In the simplest method, we associate with each CWT coefficient c_i a hidden state q_i taking value S,L depending on whether $|c_i|$ is small or large. Consider each complex coefficient c_i as a 2-d random vector (u_i, v_i), and recall that since the real and imaginary filter banks are orthogonal, the real and imaginary parts are uncorrelated random variables. Thus we approximate the marginal density $f(c_i)$ as a two-state, 2-d Gaussian mixture

$$f(c_i|q_i = m) = \frac{1}{\sqrt{2\pi}\sigma_{i,m}} \exp\left(-\frac{u_i^2 + v_i^2}{2\sigma_{i,m}^2}\right). \tag{4}$$

Factoring (4) into

$$f(c_i|q_i = m) = \frac{1}{\sqrt{2\pi}\sigma_{i,m}}\, e^{-u_i^2/2\sigma_{i,m}^2}\, e^{-v_i^2/2\sigma_{i,m}^2} \tag{5}$$

reveals that we model the real and imaginary parts as two-state, 1-d Gaussian mixtures with *shared state variable* q_i. This is quite reasonable: Small $|c_i|$ result only when both $|u_i|$ and $|v_i|$ are small simultaneously (smooth region). This is captured by $q_i = \mathsf{S}$. On the contrary, large $|c_i|$ result when either $|u_i|$ is large, $|v_i|$ is large, or both are large (edge region). This is captured by $q_i = \mathsf{L}$. (Note that, as in the DWT HMT case, small outcomes from the L-state Gaussian mixture density are entirely possible – and in fact likely – since this component is zero mean.)

This approach models each $|c_i|$ as a two-state Rayleigh mixture model. The above formulation (coupled Gaussian mixtures for the real and imaginary parts) is preferable, however, as direct Rayleigh modeling leads to unmanageable posterior distribution expressions.

The CWT HMT (CHMT) corresponding to (4) with scale-to-scale Markov transitions has a structure identical to the real DWT HMT of Section 2 and [1, 5]. The only differences will be the substitution of (4) for (1) and the use of six subband trees instead of three. The result is an HMT that supports $O(N)$ coefficient computation and training and inference algorithms and features better angle selectivity.

As an alternative to the above, we could model the $\mathbf{u}$ and $\mathbf{v}$ DWTs separately using standard 2-state, 1-d HMTs. The result is a 4-state, 2-d HMT for each $|c_i|$ that depends on the four combinations of large/small u_i and v_i. Tying the $\mathbf{u}$ and $\mathbf{v}$ HMTs node by node by forcing them to share the same mixture means and variances (but not hidden state values) effectively models $\mathbf{u}$ and $\mathbf{v}$ as independent realizations of the same random process. Thus, we can apply the training and inference algorithms for multiple data observations described in [1].

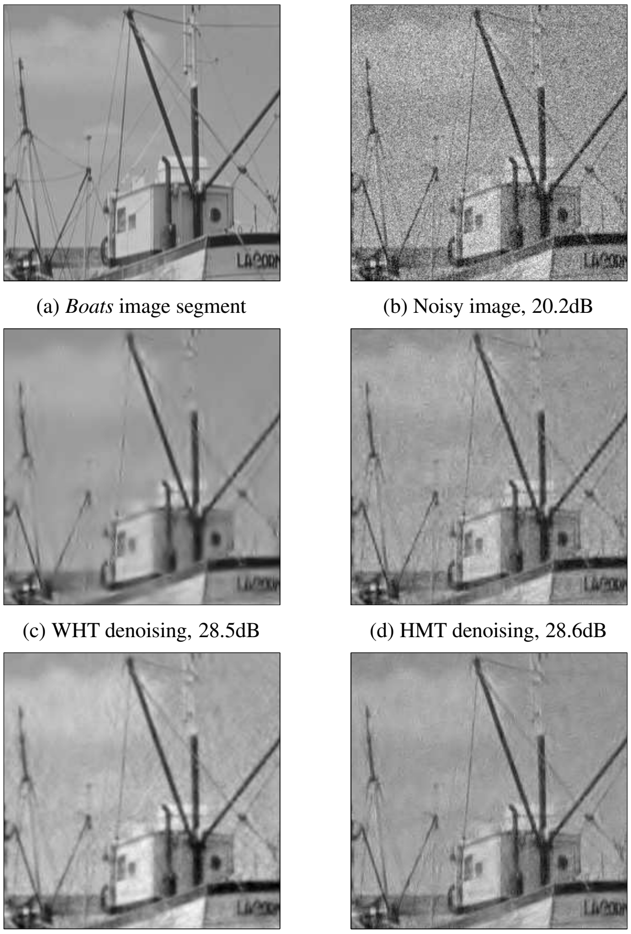
5. APPLICATION: BAYESIAN IMAGE DENOISING

CHMTs are applicable to the gamut of statistical signal and image processing tasks, from estimation and detection/classification to segmentation, fusion, and synthesis. To illustrate the modeling accuracy of the CHMT, we will continue the simple denoising experiment from Section 2.2.

Table 1 and Fig. 3 compare the performance of a number of different denoising algorithms. While the computational complexity of the CHMT is virtually the same as the real DWT HMT, we see that complex modeling buys us a significant performance improvement both in PSNR[3] and visual metrics (sharper edges and smoother backgrounds). We attribute this improvement to the near

[2] The CWT can be interpreted as a wavelet tight frame with a redundancy factor of two [11].

[3] Peak signal-to-noise ratio for an N-sample data set with values between 0 and 255 is defined as $\mathrm{PSNR} := -20\,\log_{10}\left(\frac{\|\mathbf{x}-\widetilde{\mathbf{x}}\|}{256\,N}\right)$.



(a) *Boats* image segment

(b) Noisy image, 20.2dB

(c) WHT denoising, 28.5dB

(d) HMT denoising, 28.6dB

(e) RHMT denoising, 28.6dB

(d) CHMT denoising, 29.3dB

Figure 3: *A segment of the* 512×512 Boats *image denoised using a variety of approaches, with PSNRs. (a) Original image with 256 gray levels. (b) Noisy image (*$\sigma_n = 25.5$*). (c) Redundant wavelet hard thresholding (WHT) using D8 wavelet [7, 8]. (d) HMT denoising (2) using D8 DWT [1]. (e) Redundant RHMT denoising using D8 [5]. (f) CHMT denoising using CWT from [10].*

shift invariance and superior angular resolution of the CWT. Indeed, the CHMT outperforms the state-of-the-art (and computationally expensive) redundant HMT denoiser of [5].

6. CONCLUSIONS

The ability of multiresolution signal and image models to capture the structures of signals and images is often hampered by the shift variance of the underlying wavelet transform. In this paper, we have seen that using complex wavelets, we can design statistical models that focus on the salient signal features (persistence across scale of edges and so forth) rather than artifacts due to shift variance. As our simple denoising example has demonstrated, the CHMT outperforms even the expensive but state-of-the-art redundant HMT model, due to the parsimony and better angular resolution of its underlying transform.

Table 1: *PSNRs in dB of images denoised using the wavelet algorithms of Fig. 3 plus* MATLAB*'s spatially adaptive Wiener filtering (Wiener2) for noise variances* $\sigma_n = 10$ *and* 25.5.

Image	*Boats*		*Lena*		*Bridge*	
σ_n	10	25	10	25	10	25
Noisy	28.1	20.2	28.1	20.0	28.2	20.2
Wiener2	31.2	27.9	32.7	28.7	26.0	24.6
WHT	33.2	28.5	33.9	29.6	27.7	24.2
HMT	33.5	28.6	33.9	29.3	29.7	25.4
RHMT	33.8	28.6	34.6	29.8	29.8	24.7
CHMT	34.4	29.3	34.9	29.9	30.8	25.8

While we have chosen to illustrate the application of the CHMT with an estimation (denoising) problem, this new model is generally applicable to Bayesian inference problems, including detection/classification, segmentation, synthesis, and fusion. The near shift invariance and improved angular resolution of the CWT should benefit these applications as well. Finally, the reduced-complexity universal HMT modeling framework of [5] should apply to the CHMT without modification, creating an extremely simple yet powerful signal and image modeling framework.

7. REFERENCES

[1] 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, no. 4, pp. 886–902, April 1998.

[2] D. Donoho and I. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

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

[4] M. Basseville, A. Benveniste, K. C. Chou, S. A. Golden, R. Nikoukhah, and A. S. Willsky, "Modeling and estimation of multiresolution stochastic processes," *IEEE Trans. on Info. Theory*, vol. 38, no. 2, pp. 766–784, Mar. 1992.

[5] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden Markov models," in *Proc. SPIE Math. Modeling, Bayesian Estimation, and Inverse Problems* Conf., Denver, CO, July 1999, vol. 3816, pp. 31–44. Extended version available at www.dsp.rice.edu/publications.

[6] H. Choi and R. Baraniuk, "Multiscale texture segmentation using wavelet-domain hidden Markov models," in *Proc. 32st Asilomar Conference*, Pacific Grove, CA, Nov. 1-4 1998. Extended version available at www.dsp.rice.edu/publications.

[7] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus, and R. O. Wells Jr, "Noise reduction using an undecimated discrete wavelet transform," *IEEE Signal Processing Letters*, vol. 3, no. 1, pp. 10–12, 1996.

[8] R. Coifman and D. Donoho, "Translation-invariant de-noising," in *Wavelets and Statistics*, Lecture Notes in Statistics. Springer-Verlag, 1995.

[9] N. G. Kingsbury, "Shift invariant properties of the dual-tree complex wavelet transform," in *Proc. ICASSP 99*, Phoenix, AZ, March 1999.

[10] N. G. Kingsbury, "Image processing with complex wavelets," *Phil. Trans. Royal Society London A*, vol. 357, pp. 2543–2560, September 1999.

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