ABSTRACT

We develop a hierarchical, nonparametric statistical model for wavelet representations of natural images. Extending previous work on Gaussian scale mixtures, wavelet coefficients are marginally distributed according to infinite, Dirichlet process mixtures. A hidden Markov tree is then used to couple the mixture assignments at neighboring nodes. Via a Monte Carlo learning algorithm, the resulting hierarchical Dirichlet process hidden Markov tree (HDP-HMT) model automatically adapts to the complexity of different images and wavelet bases. Image denoising results demonstrate the effectiveness of this learning process.

Index Terms— hidden Markov trees, hierarchical Dirichlet processes, nonparametric Bayesian methods, wavelet transforms, image denoising.

1. INTRODUCTION

Wavelet decompositions of natural images exhibit significant, highly non-Gaussian residual dependencies. Empirical results indicate that Gaussian scale mixtures often provide excellent models for the heavy-tailed distributions of individual wavelet coefficients [1]. Several previous papers have used hidden Markov trees to couple scale mixtures in a coherent global model [1–3]. Such models are used in a range of tasks, including image compression, denoising, and classification.

In this article, we adapt the hierarchical Dirichlet process [4] to design a nonparametric Bayesian model for wavelet coefficients. This model employs a potentially infinite set of hidden states to capture dependencies among observed wavelet coefficients. Through an appropriate prior distribution, however, only a finite, data-driven subset of these states is used to make predictions. In the following sections, we develop a collapsed Gibbs sampler which learns model parameters from training data, and validate its accuracy via wavelet histograms and image denoising performance.

2. WAVELET-BASED STATISTICAL MODELS

Natural images exhibit sharply localized intensity changes due to occlusion boundaries, as well as more homogeneously textured regions. For these reasons, their statistics are most simply described by representations which are jointly localized in spatial position and frequency. Wavelet decompositions provide a family of widely used image bases designed to achieve these two competing goals [5].

Orthogonal wavelet transforms decompose images at multiple scales by recursively filtering with a scaled, band-pass kernel function. This invertible linear operator produces a set of low-pass scaling coefficients \( x_{00} \), and a forest of multiscale trees containing higher frequency detail coefficients \( x_t \). As illustrated in Fig. 1, we let \( x_t \) denote the vector of detail coefficients (of different orientations) at location \( i \) beneath scaling coefficient \( t \).

Because orthogonal wavelet coefficients are nearly decorrelated, they lead to effective compression algorithms. However, such critically sampled decompositions are not translationally invariant, and exhibit instability and aliasing artifacts in the presence of noise. Steerable pyramids address these issues via an overcomplete basis, or frame, optimized for increased orientation selectivity [6]. While the statistics of such non-orthogonal transformations are more complex, our results demonstrate their advantages in image analysis.

2.1. Mixture Models for Heavy-Tailed Marginals

The marginal distributions of wavelet coefficients are typically highly kurtotic, with "heavy tails" indicating that extreme values occur frequently compared to Gaussian distributions. This behavior is captured by Gaussian scale mixtures, which model \( x_t \) as the product of two independent variables:

\[
x_{t_i} = v_{t_i} u_{t_i} \quad v_{t_i} \geq 0, \quad u_{t_i} \sim \mathcal{N}(0, \Lambda)
\]

Marginalizing the scalar multiplier \( v_{t_i} \) mixes Gaussians of varying scales. A variety of continuous mixing distributions provide good models of wavelet statistics [1]. In many cases, however, simple two-component mixtures are also effective:

\[
x_{t_i} \sim \pi \mathcal{N}(0, \Lambda_0) + (1 - \pi) \mathcal{N}(0, \Lambda_1)
\]

Here, \( \pi \) is the probability that \( x_{t_i} \) is drawn from an "outlier" component with large variance \( \Lambda_0 \), and \( \Lambda_1 \) is smaller to capture the many near-zero coefficients. Such discrete mixtures have important computational advantages, and have been successfully used for image denoising [7].
Fig. 1. Two levels of an HDP-HMT in which hidden discrete states $z_{ti}$ generate wavelet coefficients $x_{ti}$. Each of the infinitely many states has an output covariance $\Lambda_k$ and transition distribution $\pi_k$, which are coupled when learning via a global random measure, $\beta$.

2.2. Wavelet Cascades on Markov Trees

Although natural images often lead to uncorrelated wavelet coefficients, they retain important non–Gaussian dependencies. In particular, large magnitude coefficients tend to cluster at nearby spatial locations, and persist across multiple scales [1,2]. One of the most effective image denoising algorithms employs local Gaussian scale mixtures relating each wavelet coefficient to its nearest neighbors in location and scale [8]. In this paper, we instead develop a global graphical model of multiscale image decompositions.

The scale–recursive operations underlying wavelet decompositions suggest models defined on Markov trees [5]. For images, these graphical models associate detail coefficient $x_{ti}$ with a single coarser scale parent $x_{Pa(ti)}$, and four finer scale children $\{x_{tj} \mid t_j \in Ch(ti)\}$ (see Fig. 1). Tree–structured Gaussian random fields have been used to capture correlations among wavelet coefficients [5], and to model the latent multipliers underlying a global Gaussian scale mixture [1]. Alternatively, the discrete mixture of eq. (2) has been generalized to define a binary hidden Markov tree (HMT) [2]. In HMTs, the mixture component $z_{ti}$ generating detail coefficient $x_{ti}$ is influenced by the corresponding parent coefficient:

$$z_{ti} \mid z_{Pa(ti)} \sim \pi_{z_{Pa(ti)}} \quad x_{ti} \mid z_{ti} \sim \mathcal{N}(0, \Lambda_{z_{ti}}) \quad (3)$$

As before, detail coefficient $x_{ti}$ may be generated via states $z_{ti}$ of low or high variance. However, by associating each parent state $k$ with a different transition distribution $\pi_k$, HMTs also capture dependencies among nearby coefficients.

Although the HMT originally defined separate graphical models for each orientation subband, states may alternatively generate vectors of wavelet coefficients [3]. Dependencies among orientations are then better captured by higher–order discrete models. To do this, one must select an appropriate number of hidden states $K$, as well as the pattern used to share states among different coefficients. For example, the hierarchical image probability (HIP) model [3] shares parameters within each scale, and optimizes $K$ via a minimum description length (MDL) criterion. In the following section, we propose an alternative nonparametric approach which learns such model structures from training images.

3. NONPARAMETRIC HIDDEN MARKOV TREES

Many model selection criteria, including MDL, have asymptotic justifications which are poorly suited to small datasets. When applied to hierarchical models, they may also lead to combinatorial problems requiring greedy approximations [3]. Nonparametric Bayesian methods avoid these issues by defining priors on infinite models. Learning algorithms then produce robust predictions by averaging over model structures whose complexity is justified by the observed data.

3.1. Dirichlet Process Mixtures

Let $H$ denote a prior distribution on the space of zero–mean Gaussian distributions, such as the inverse–Wishart [9]. A Dirichlet process (DP) with concentration parameter $\gamma > 0$, denoted by $DP(\gamma, H)$, then defines a prior distribution over infinite Gaussian mixtures:

$$\beta_k = \beta_k' \prod_{\ell=1}^{k-1}(1 - \beta_{\ell}') \quad \beta_k' \sim \text{Beta}(1, \gamma) \quad (4)$$

$$p(x_{ti} \mid \beta, \Lambda_1, \Lambda_2, \ldots) = \sum_{k=1}^{\infty} \beta_k \mathcal{N}(x_{ti}; 0, \Lambda_k) \quad (5)$$

Component variances are independently sampled as $\Lambda_k \sim H$. The stick-breaking construction [4, 9] of eq. (4), which we denote by $\beta \sim \text{GEM}(\gamma)$, defines mixture weights using beta random variables. In contrast with finite mixtures, DPs favor simple models given few observations, but also create low–probability clusters to capture details revealed by large, complex datasets. Practically, DP mixtures are motivated both by their attractive asymptotic guarantees [9], and by the availability of efficient computational methods [4, 9].

3.2. Hierarchical Dirichlet Processes

The hierarchical Dirichlet process (HDP) provides a flexible framework for sharing mixture components among groups of related data [4]. It has been previously used to define an HDP-HMM which learns the structure of a countably infinite hidden Markov chain from training data. Extending this work, we develop an HDP hidden Markov tree (HDP-HMT) to model the global statistics of wavelet coefficients.

Consider a hidden Markov tree with a countably infinite state space $z_{ti} \in \{1, 2, \ldots\}$. Each value $k$ of the current state indexes a different transition distribution $\pi_k = (\pi_{k1}, \pi_{k2}, \ldots)$ over child states, which we couple via a shared DP prior

$$\pi_k \sim \text{DP}(\alpha, \beta) \quad \beta \sim \text{GEM}(\gamma) \quad (6)$$

By defining $\beta$ to be a discrete probability measure, we ensure with high probability that a common set of child states are reachable from each parent state [4]. This hierarchical
construction encourages reuse of states when learning. Given these infinite transition distributions, wavelet coefficients are generated as in the coarse–to–fine recursion of eq. (3).

By defining a prior on infinite models, the HDP-HMT avoids the model selection issues considered by [3]. Importantly, we also do not need to specify a fixed pattern by which states are shared among different coefficients. While the HDP-HMT biases coefficients to share states, it learns detailed transition structures from training images.

**4. MODELLING IMAGES USING THE HDP-HMT**

Extending the direct assignment Gibbs sampler of [4], we now develop a Monte Carlo learning method for HDP-HMTs. For simplicity, we first consider models for the wavelet coefficients in a single noise–free image. Sec. 5 then extends this sampler to develop an image denoising algorithm.

**4.1. Collapsed Gibbs Sampling**

Given a training image containing wavelet coefficients $x_{ti}$, we would like to infer the posterior distribution of the HDP-HMT’s parameters. We do this via a Gibbs sampler which alternates between sampling assignments $z_{ti}$ to hidden states and global transition probabilities $\beta$. Given fixed values for these variables, the state–specific transition distributions $\pi_k$ and covariances $\Lambda_k$ can be marginalized in closed form. Such Rao–Blackwellization improves the efficiency and accuracy of MCMC methods [9].

Given fixed assignments $z = \{ z_{ti} \}$ of coefficients to hidden states, $\beta$ can be resampled as described in [4]. However, in contrast with standard HDP models, the HDP-HMT dynamically regroups wavelet coefficients as parent states are resampled. From Fig. 1, the posterior distribution of $z_{ti}$ given all other state assignments $z_{t \neq i}$ factors as

$$p(z_{ti} \mid z_{t \neq i}, \beta, x) \propto p(z_{ti} \mid z_{t \neq i}, \beta) p(x_{ti} \mid x_{t \neq i}, z)$$

(7)

The second term is the predictive likelihood of $x_{ti}$, which for inverse–Wishart priors is multivariate Student–t [9]. The form of the first term depends on the position $i$ of the sampled coefficient, and the states of its neighbors. Let $n_{ti}(k, \ell)$ denote the number of transitions from parent state $k$ to child state $\ell$ instantiated by $z_{t i}$, and $n_{ti}(k, \cdot)$ the total number of outgoing transitions from state $k$. For finest scale coefficients, $p(z_{ti} \mid z_{Pa(ti)}) = k, z_{t i}, \beta) = \int \pi_k(z_{ti}) p(\pi_k \mid z_{t i}, \beta) d\pi_k$

$$= \left( \frac{n_{ti}(k, z_{ti}) + \alpha(\beta_{z_{ti}})}{n_{ti}(k, \cdot) + \alpha} \right)$$

(8)

The form of this ratio follows from the properties of Dirichlet distributions [4]. When evaluating eq. (8), we consider candidate states $z_{ti}$ corresponding to every state which is used at least once elsewhere in the wavelet tree, as well as a potential new state. This predictive rule allows HDP-HMTs to determine state space cardinality in a data–driven fashion.

For non–leaf nodes, $p(z_{ti} \mid z_{Pa(ti)}, \beta)$ is also influenced by its children’s states $z_{Ch(ti)} = \{ z_{tj} \mid \ell_j \in Ch(ti) \}$. In candidate states where $z_{ti} \neq z_{Pa(ti)}$, this conditional distribution factors into two terms: one as in eq. (8), and a second corresponding to the likelihood $p(z_{Ch(ti)} \mid z_{ti}, \beta)$. When $z_{ti} = z_{Pa(ti)}$, a correction is needed to avoid double–counting information.

**4.2. Validation Through Wavelet Histograms**

In Fig. 2, we illustrate wavelet coefficient histograms [1] computed from the grayscale peppers image. We compare this raw data to coefficients simulated from HDP-HMTs learned using 1,000 Gibbs sampling iterations. We correctly model the non–Gaussian “bow tie” shapes of wavelet histograms, and also accurately capture the complex orientation and scale relationships exhibited by steerable pyramids. Our underestimate of the dependence between spatially adjacent coefficients is probably caused by the Markov tree boundaries which separate some pairs of fine scale coefficients [5].

![Fig. 2. Pairwise histograms of steerable pyramid detail coefficients from the 256 x 256 peppers image. Rows 1 & 3 are computed from the observed image, while rows 2 & 4 summarize samples from an HDP-HMT. As in [1], we visualize log-contours of joint distributions (top) as well as normalized conditional distributions (bottom).](image-url)
on the noisy wavelet tree. Discarding the first 1,000 “burn–in” iterations, we then collect 500 samples \( \theta^{(s)} = \{ \pi_k^{(s)}, \Lambda_k^{(s)} \} \) from the parameters’ posterior distribution. Note that each sample \( s \) instantiates a different number of states \( K_s \).

Given \( \theta^{(s)} \), the conditional mean of \( w_{ti} \) equals

\[
\mathbb{E}[w_{ti} \mid x, \theta^{(s)}] = \sum_{k=1}^{K_s} p(z_{ti} = k \mid x, \theta^{(s)}) \mathbb{E}[w_{ti} \mid x_{ti}, \Lambda_k^{(s)}]
\]

where the posterior state probabilities \( p(z_{ti} = k \mid x, \theta) \) may be efficiently computed via the belief propagation algorithm [2, 5]. Given \( z_{ti} \), denoising reduces to linear least squares:

\[
\mathbb{E}[w_{ti} \mid x_{ti}, \Lambda_k^{(s)}] = \Sigma_k^{(s)} \Sigma_k^{-1} x_{ti}
\]

As in [2, 8], we estimate \( \Sigma_k^{(s)} \) by subtracting \( \Sigma_n \) from \( \Lambda_k^{(s)} \), and setting any negative eigenvalues to zero. When the mean of the variance prior \( H \) is sufficiently large, such truncations are rarely necessary. The denoised image is then determined via an inverse wavelet transform combining observed scaling coefficients with the posterior mean of each detail coefficient.

### 5.2. Results: Denoising Peppers

In Fig. 3, we compare the HDP-HMT’s denoising performance to three other methods. Using identical Daubechies-4 orthogonal wavelets, the larger state space of the HDP-HMT consistently improves on the binary HMT [2] in sharpness, clarity, and peak signal-to-noise ratio (PSNR). Alternatively, a 5\( \text{th} \)-order steerable pyramid decomposition leads to an HDP-HMT providing clearer estimates with fewer aliasing artifacts. Compared to a state–of–the–art local GSM model [8], the HDP-HMT produces images with sharper details, but more artifacts (and hence lower PSNR) in smooth regions.

### 6. DISCUSSION

We have developed a nonparametric, tree–based model of joint wavelet statistics, and demonstrated its effectiveness in an image denoising task. This HDP-HMT generalizes existing finite–state models by allowing the number of hidden states, and the structure of their transitions, to be learned from training images. Future work will explore the degree to which these states generalize across natural image families.

### 7. REFERENCES


