TO APPEAR AT THE 2007 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION

Learning Multiscale Representations of Natural Scenes Using Dirichlet Processes

Jyri J. Kivinen
Helsinki University of Technology
Espoo, Finland
International Computer Science Institute
Berkeley, CA, USA
kivinen@cs.berkeley.edu

Erik B. Sudderth[†] and Michael I. Jordan[†⋆]
University of California, Berkeley
Computer Science Division[†]
and Department of Statistics[⋆]
Berkeley, CA, USA
sudderth,jordan@cs.berkeley.edu

Abstract

We develop nonparametric Bayesian models for multiscale representations of images depicting natural scene categories. Individual features or wavelet coefficients are marginally described by Dirichlet process (DP) mixtures, yielding the heavy-tailed marginal distributions characteristic of natural images. Dependencies between features are then captured with a hidden Markov tree, and Markov chain Monte Carlo methods used to learn models whose latent state space grows in complexity as more images are observed. By truncating the potentially infinite set of hidden states, we are able to exploit efficient belief propagation methods when learning these hierarchical Dirichlet process hidden Markov trees (HDP-HMTs) from data. We show that our generative models capture interesting qualitative structure in natural scenes, and more accurately categorize novel images than models which ignore spatial relationships among features.

1. Introduction

Psychophysical experiments have shown that, in a manner analogous to object categorization, images of natural environments are often identified as members of certain "basic level" scene categories [24]. In this paper, we develop nonparametric statistical methods which learn multiscale representations of natural scenes, and use these models to accurately categorize images of new environments. This kind of semantic recognition is directly useful in applications such as image annotation and retrieval [25]. More generally, the global identity and structure of natural scenes provide important contextual cues for the detection and recognition of objects [23, 24].

A variety of statistical cues are associated with scene perception, including local structural elements [6, 12, 13], color patterns, and global spatial attributes such as openness, roughness, and naturalness [15, 24]. Given human–annotated training data, these global scene properties can be inferred from spectral representations of images, and used to classify new scenes [15]. Alternatively, semantic labelings of local image regions have been proposed as an intermediate representation for global scene identification [25]. One drawback of these approaches is that hundreds of training images must be manually labeled according to their structural properties [15] or constituent objects [25].

To reduce the degree of supervision required during training, several recent papers have adapted textual topic models [2, 8] to categorize natural scenes [3, 6, 17]. These approaches focus on clustering characteristic local textures, transforming images to "bags of features" and ignoring global spatial structure. In this paper, we generalize these approaches in two complementary ways. First, rather than choosing a fixed number of latent topics by cross-validation, we adapt the hierarchical Dirichlet process (HDP) [22] to define nonparametric models whose complexity grows as additional scenes are observed. Second, we use a tree–structured graphical model [5, 20, 26, 27] to couple the visual topic assignments at nearby positions and scales. Our results confirm that the resulting *hierarchical Dirichlet process hidden Markov tree* (HDP-HMT) more accurately captures natural scene statistics, and leads to improved categorization performance.

The HDP-HMT was first proposed as a model for the joint statistics of wavelet decompositions [11]. This earlier work demonstrated that Dirichlet process mixtures provide a suitable model for the heavy–tailed statistics of wavelet coefficients, and used the HDP-HMT to develop an effective image denoising algorithm. In this paper, we build hierarchical models for a family of natural scenes rather than single images, and use them for semantic categorization rather than low–level image processing. We compare the discriminative power of two different image representations: the multiscale oriented edge responses of steerable pyramids [18], and a discrete vocabulary of vector quantized SIFT descriptors [14, 19]. To allow learning from large databases of scenes, we also adapt finite approxima-

tions of the Dirichlet process [9, 10] to develop a *truncated* Gibbs sampler with significant computational advantages over existing methods [11, 22].

We begin in Sec. 2 by reviewing previous models for wavelet coefficients based on Gaussian scale mixtures. Sec. 3 describes a complementary family of topic models adapted to unstructured, feature–based image representations. We then integrate these research themes in the HDP-HMT (Sec. 4), develop efficient Monte Carlo methods for learning from training images (Sec. 5), and evaluate its suitability as a model of natural scene categories.

2. Wavelet Representations of Natural Images

Images of natural scenes typically contain large, homogeneously textured regions, as well as localized intensity changes caused by occlusion boundaries. Their statistics are thus most simply characterized in representations which are jointly localized in spatial position and frequency [24, 27]. *Wavelet* transforms decompose images at multiple scales by recursively filtering with a scaled, band–pass kernel function. This invertible linear transform produces a set of low–pass *scaling* coefficients x_{t0}, and a forest of multiscale trees containing higher frequency *detail* coefficients $\mathbf{x}_t = \{x_{ti}\}$. As illustrated in Fig. 2, we let x_{ti} denote the vector of detail coefficients (of different orientations) at location i beneath scaling coefficient t.

While orthogonal wavelets approximately decorrelate natural images, and thus lead to effective compression algorithms, their lack of translational invariance may lead to 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 [18]. While the statistics of such non–orthogonal transformations are more complex, they are advantageous for image analysis [11, 16].

2.1. Mixture Models for Heavy–Tailed Marginals

Wavelet coefficients typically have *kurtotic* marginal distributions, with "heavy tails" indicating that extreme values occur frequently compared to Gaussian distributions. This behavior is captured by *Gaussian scale mixtures*, which model x_{ti} as the product of two independent variables:

$$x_{ti} = v_{ti} u_{ti} \qquad v_{ti} \geq 0,\ u_{ti} \sim \mathcal{N}(0, \Lambda) \tag{1}$$

Marginalizing the scalar multiplier v_{ti} mixes Gaussians of varying scales. While continuous mixing distributions provide good models of wavelet statistics [26], in many cases two–component discrete mixtures are also effective:

$$x_{ti} \sim \pi \mathcal{N}(0, \Lambda_0) + (1 - \pi)\mathcal{N}(0, \Lambda_1) \tag{2}$$

Here, π is the probability that x_{ti} is drawn from an "outlier" component with large variance Λ_0, and Λ_1 is smaller to capture the many near–zero coefficients. Discrete mixtures have important computational advantages, and have been successfully used for image denoising [4].

2.2. Modeling Wavelets with Markov Trees

The statistical structure of natural scenes is highly non–stationary [24], and wavelet coefficients typically retain significant non–Gaussian dependencies. For example, large magnitude coefficients often *persist* across multiple scales, and *cluster* at nearby spatial locations [5, 26]. Motivated by image denoising tasks, *local* Gaussian scale mixtures have been used to relate wavelet coefficients at neighboring locations and scales [16]. This paper instead develops a *global* graphical model of multiscale image decompositions.

Due to their scale–recursive construction, wavelet decompositions suggest models defined on *Markov trees* [27]. For images, these graphical models associate detail coefficient x_{ti} with a single coarser scale *parent* $x_{\mathrm{Pa}(ti)}$, and four finer scale *children* $\{x_{tj} \mid tj \in \mathrm{Ch}(ti)\}$ (see Fig. 2). Tree–structured Gaussian random fields have been used to capture correlations among wavelet coefficients [27], and to model the latent multipliers underlying a global Gaussian scale mixture [26]. Alternatively, the discrete mixture of eq. (2) has been generalized to define a binary *hidden Markov tree* (HMT) [5]. In HMTs, the mixture component z_{ti} generating detail coefficient x_{ti} is influenced by the corresponding parent coefficient:

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

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 its own *transition distribution* π_k, HMTs also capture dependencies among nearby coefficients.

The earliest applications of HMTs defined independent graphical models for each orientation subband, and tied model parameters within each scale to avoid overfitting [5]. It would be preferable to capture dependencies between orientations, and reduce boundary artifacts from the latent tree structure [27], by using higher–order discrete models to generate vectors of wavelet coefficients. However, 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 [20] shares parameters within each scale, and optimizes K via a minimum description length (MDL) criterion. This paper instead extends the nonparametric approach of [11] to *learn* latent states whose complexity grows as new training images are observed.

3. Feature–Based Image Representations

In vision applications involving geometric correspondence or semantic categorization, image representations based on distinctive local features are often effective [3, 6, 13, 14, 19]. By making individual features discriminative, these methods can make reasonably accurate predictions with very simple global scene models. In this paper, we show that coupling local features with global spatial models can further boost recognition performance [12, 21].

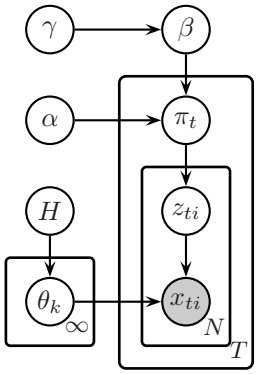


Figure 1. Graphical representation of a bag–of–features hierarchical DP. Observed data points x_{ti} in each of the T groups are drawn from group–specific mixtures, whose mixing proportions π_t are coupled via a global random measure β. Hidden states z_{ti} index which of the infinitely many shared mixture components, or topics, generate observed feature $x_{ti} \sim F(\theta_{z_{ti}})$.

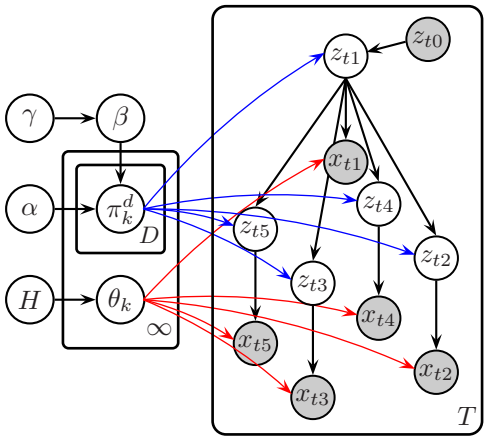


Figure 2. Two levels of an HDP-HMT in which hidden discrete states z_{ti} generate observed features x_{ti}. In contrast with the HDP of Fig. 1, the states at neighboring locations and scales are coupled by direction dependent transition distributions π_k^d. As before, a global measure β is used to couple these transitions when learning, encouraging reuse of hidden states.

3.1. Feature Detection and Extraction

Previous comparative studies have shown that for scene categorization tasks, the best performance is attained by computing features on a dense, regular grid [3, 6], rather than at sparse interest points [14, 17]. The intuitive explanation for this phenomenon is that the presence of open, textureless regions is highly indicative of certain scene categories [15, 24]. In this paper, we extract features from overlapping patches spaced on a two-pixel grid. To provide further discriminative power, we also rescale the input image and extract dense features at three coarser scales [3].

Following several recent papers [3, 6, 12, 17], we use SIFT descriptors [14] to describe the appearance of each feature. SIFT descriptors provide effective estimates of the local orientation cues characteristic of natural scenes [24], as well as some invariance to lighting and viewpoint. To reduce dimensionality, we use K–means clustering to vector quantize the SIFT descriptors observed in training images, producing a dictionary of "visual words." This approach is quite similar to texton approaches to texture recognition [13], except that SIFT descriptors aggregate filter responses over a local region rather than at a single point.

Fig. 4 illustrates the visual words extracted from several natural scenes. To visualize our unordered SIFT dictionary, we compute a PCA decomposition of the observed features, and sort codebook vectors according to their projection on the first principal direction. For SIFT descriptors, this procedure arranges words by their dominant orientation. Note that this one–dimensional PCA projection is used only for visualization; the actual SIFT codebook contains higher–order information about local appearance patterns.

3.2. Topic Models for Bags of Features

Several recent approaches to discovering semantic or spatial structure in visual scenes have been inspired by methods for learning the topics discussed in a corpus of text documents. Hierarchical probabilistic models like *probabilistic latent semantic analysis* (pLSA) [8] and *latent Dirichlet allocation* (LDA) [2] provide one effective framework for unsupervised topic discovery. These models begin by representing documents as "bags of words," discarding the sentence structure and underlying syntax. Semantic content is then analyzed via a two–level hierarchical clustering, in which topics are associated with typical words and documents with a distribution over topics.

Topic models have previously been used to discover objects in simple scenes [19] or web search results [7], parse presegmented captioned images [1], decompose object categories via a shared set of parts [21], and categorize natural scenes [3, 6, 17]. Following an initial stage of segmentation [1] or feature extraction [3, 6, 17, 19], most such models discard location information, retaining an unstructured bag of features. However, because image patches are far less distinctive than words in human languages, such approximations ignore potentially valuable global scene attributes. Image-based coordinate frames have been used to model the internal structure of objects [7, 21], but seem too rigid for natural scene categories.

Another limitation of many such models is that the *number* of latent topics, parts, or objects must be specified. This choice is known to significantly impact predictive performance [2, 3, 6, 21, 22], and computationally expensive cross–validation procedures are often required. The following section describes the hierarchical Dirichlet pro-

cess (HDP) [22], a nonparametric alternative which avoids model selection via a *probabilistically* constrained infinite state space. We then integrate the HDP with the hidden Markov trees of Sec. 2.2, and thus augment visual topic assignments with spatial dependencies.

4. Nonparametric Models of Image Features

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 [20]. Nonparametric Bayesian methods avoid these issues by defining priors on *infinite* models. Learning algorithms then produce robust predictions by *averaging* over model substructures whose complexity is justified by the observed data [21, 22].

4.1. Dirichlet Process Mixtures

Let H denote a prior on some space Θ of appearance distributions F. For example, we later use inverse–Wishart H to construct zero–mean Gaussian models of continuous wavelet coefficients, and Dirichlet H for multinomial models of vector quantized SIFT descriptors. A *Dirichlet process* (DP) with concentration parameter $\gamma > 0$, denoted by $\mathrm{DP}(\gamma, H)$, then defines a prior over *infinite* mixtures:

$$\beta_k = \beta'_k \prod_{\ell=1}^{k-1} (1 - \beta'_\ell) \qquad \beta'_\ell \sim \mathrm{Beta}(1, \gamma) \tag{4}$$

$$p(x_{ti} \mid \beta, \theta_1, \theta_2, \ldots) = \sum_{k=1}^{\infty} \beta_k f(x_{ti} \mid \theta_k) \tag{5}$$

Component appearance parameters are independently sampled as $\theta_k \sim H$. The *stick–breaking construction* of eq. (4), which we denote by $\beta \sim \mathrm{GEM}(\gamma)$, 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 [21], and by the availability of many efficient computational methods [9, 21, 22].

4.2. Hierarchical Dirichlet Processes

The *hierarchical Dirichlet process* (HDP) provides a flexible framework for sharing mixture components, or topics, among *groups* of related data [22]. In visual recognition tasks, for example, these groups could correspond to images of similar scenes, or categories of related objects [21]. As in eq. (4), we begin by sampling *global* weights β for an infinite set of shared mixture components $\{\theta_k\}_{k=1}^{\infty}$. Each of the T groups (see Fig. 1) then reuses these same components in different proportions $\pi_t = (\pi_{t1}, \pi_{t2}, \ldots)$:

$$\pi_t \sim \mathrm{DP}(\alpha, \beta) \qquad \beta \sim \mathrm{GEM}(\gamma) \tag{6}$$

Here, β determines the mean frequency of each topic, while α controls the variability of topic weights across groups [22]. Fixing these parameters, observations are then independently sampled as in finite topic models:

$$z_{ti} \sim \pi_t \qquad x_{ti} \,|\, z_{ti} \sim F(\theta_{z_{ti}}) \tag{7}$$

Rather than strictly constraining the number of latent topics, the HDP's stick–breaking prior places a softer bias towards the simplest models which explain observed data. As we demonstrate in Sec. 6, this leads to rich models whose complexity grows as additional data are observed.

4.3. Hierarchical DP Hidden Markov Trees

Hierarchical Dirichlet processes have been previously used to define an HDP-HMM which learns the structure of a countably infinite hidden Markov chain from training data [22]. In this section, we develop an *HDP hidden Markov tree* (HDP-HMT) which captures the global statistics of wavelet decompositions or locally extracted image features. Our approach generalizes earlier work [11] by sharing hidden states among multiple images, and allowing distinct transition densities for each child node.

Consider a hidden Markov tree, as in Fig. 2, where each node has a countably infinite state space $z_{ti} \in \{1, 2, \ldots\}$. Each value k of the current state indexes a different transition distribution $\pi_k^d = (\pi_{k1}^d, \pi_{k2}^d, \ldots)$ over child states in different directions d. We couple these transitions via a shared DP prior:

$$\pi_k^d \sim \mathrm{DP}(\alpha, \beta) \qquad \beta \sim \mathrm{GEM}(\gamma) \tag{8}$$

The simplest approach ties all four children of each parent to follow the same transition distribution [11]. However, we have found that allowing a distinct transition distribution π_k^d for each of the four child directions d more accurately models the asymmetries present in natural scenes. Given these infinite transition distributions, visual features are generated via the following coarse–to–fine recursion:

$$z_{ti} \,|\, z_{\mathrm{Pa}(ti)} \sim \pi_{z_{\mathrm{Pa}(ti)}}^{d_{ti}} \qquad x_{ti} \,|\, z_{ti} \sim F(\theta_{z_{ti}}) \tag{9}$$

By defining β to be a *discrete* probability measure, we ensure with high probability that a common set of child states are reachable from each parent state [22].

Analogously to the standard HDP of Fig. 1, this hierarchical construction encourages reuse of states when learning. However, the group associated with each observation is now dynamically determined by the state of its parent node, rather than being fixed *a priori*. This allows the HDP-HMT to learn complex patterns characteristic of multiscale observation sequences, and avoids the need to specify a fixed scheme for sharing states among observations. Furthermore, by defining a prior on infinite models, the HDP-HMT avoids the model selection issues considered by previous applications of Markov trees [20] and topic–based visual scene models [3, 6, 17].

4.4. Mapping Image Features to HDP-HMTs

Our scene categorization results compare two sets of visual features: the steerable pyramids of Sec. 2, and vector quantized SIFT descriptors of Sec. 3.1. For steerable pyramids, we define a separate quadtree t for the detail coefficients x_{ti} located beneath each scaling coefficient x_{t0}. While the sequence of hidden states z_{ti} within each quadtree is independently sampled, the trees from all images of a given scene category share a common set of observation and transition distributions, and are thus coupled when learning. The coarsest–scale detail coefficient z_{t1} in each tree is sampled from a distribution indexed by a fixed *root state* z_{t0} (see Fig. 2). In our experiments, we divide the *rows* of scaling coefficients into eight sets, and associate each row with a different root state. This construction is designed to capture the vertically layered structure of common natural scenes [15, 23, 24].

When using discrete features, we take a similar approach: SIFT descriptors are computed at several scales, with a 50% reduction in image resolution between successive scales. A separate quadtree is then used to link the finer-scale features beneath each feature at the coarsest scale. This structure is heuristically similar to a recently proposed kernel–based method for spatial feature matching [12]. However, that work treats the features at different scales independently, while we introduce hidden states which explicitly couple feature appearance across scales. In addition, while kernel methods must typically retain a large proportion of the training images for later testing, our learning algorithms produce a compact set of latent states which can be used to quickly categorize novel scenes.

5. Learning Hierarchical Scene Models

In this section, we propose a novel Monte Carlo method for learning HPD-HMT parameters from training images. Dirichlet processes have several complementary analytic representations, which have led to the development of many different sampling algorithms [9, 21, 22]. In previous work, posterior inference for the HDP-HMT was accomplished via a *direct assignment* Gibbs sampler [11], adapted from a related approach to HDP-HMMs [22]. This sampler explicitly instantiates the assignments z_{ti} of features to hidden states, as well as global mixture weights β_k for states with at least one assigned observation. Given these variables, the state–specific transition distributions π_k and observation parameters θ_k can be marginalized analytically.

While the direct assignment sampler desirably employs *Rao–Blackwellization* [21] to avoid explicitly sampling some latent variables, it can exhibit slow mixing because it only updates one hidden state assignment at a time. In addition, the recursive updates of sufficient statistics needed to marginalize parameters can be costly when performed after every feature reassignment. To address these issues, we propose an alternative *truncated* sampler which uses finite approximations of the Dirichlet process to allow joint resampling of entire trees of state assignments.

5.1. Truncations of Dirichlet Processes

There are two basic methods for producing finite approximations to DP models. The first truncates the stick–breaking construction of eq. (4) by setting $\beta'_L = 1$ for some sufficiently large L. In this paper, we instead explore an alternative, "weak limit" approximation which samples β from a K–dimensional finite Dirichlet distribution with symmetric parameters:

$$\beta = (\beta_1, \ldots, \beta_K) \sim \mathcal{D}(\gamma/K, \ldots, \gamma/K) \qquad (10)$$

We then take β as the weight vector for a finite, K–component mixture model with parameters $\theta_k \sim H$ as before. It can then be shown that the predictions based on this finite model converge in distribution to those of a corresponding Dirichlet process $\mathrm{DP}(\gamma, H)$ as $K \to \infty$ [9, 10]. A similar finite approximation exists for the HDP [22] of Fig. 1, in which β is sampled as in eq. (10) and group–specific mixture weights are drawn according to

$$\pi_t = (\pi_{t1}, \ldots, \pi_{tK}) \sim \mathcal{D}(\alpha\beta_1, \ldots, \alpha\beta_K) \qquad (11)$$

The next section extends this approximation to the HDP-HMT to develop a truncated Gibbs sampling algorithm.

It is important to note that the truncation level K is *not* taken to be the number of mixture components observed in the data, but rather a loose upper bound on that number. As we show in Sec. 6, the Dirichlet priors of eqs. (10, 11) cause the sampler to explain observations via a dynamically chosen *subset* of the pool of available mixture states. Theoretical results are available which characterize the mixture size needed for accurate posterior approximations [10].

5.2. Truncated Gibbs Sampling

Given a truncation level K, our truncated Gibbs sampler alternates between blocked resampling of trees of state assignments $\mathbf{z}_t$, global mixture weights β, and state-specific model parameters and transition distributions $\{\theta_k, \pi_k\}_{k=1}^K$. The following sections briefly sketch the details of these resampling steps. The truncation level K can be either chosen larger than the number of expected states to ensure a good approximation to the underlying HDP, or set smaller to control computational complexity with large datasets.

Sampling Assignments via Belief Propagation We begin by conditioning on each state's transition distribution π_k and observation distribution θ_k. Given these fixed parameters, the joint distribution of the hidden states $\mathbf{z}_t$ and observations $\mathbf{x}_t$ can be represented by a forest of tree–structured, directed graphical models (see Fig. 2). For such models, the belief propagation (or sum–product) algorithm can be used to efficiently resample *all* of the latent assignments in closed form [27]. Messages are first passed from the

leaves to the root of each tree to collect summary statistics, which can also be used to evaluate the marginal likelihood $p\big(\mathbf{x}_t \mid \{\pi_k, \theta_k\}_{k=1}^K\big)$ in closed form. A top–down recursion is then used to resample each node z_{ti} given its parent $z_{\mathrm{Pa}(ti)}$. The computational cost of resampling the assignments for N observed features is thus $\mathcal{O}(NK^2)$.

Sampling Model Parameters In the second stage of the truncated sampler, we condition on the assignments z of observations to hidden states. It is then straightforward to resample the observation distributions θ_k by aggregating statistics of the observations $\{x_{ti} \mid z_{ti} = k\}$ assigned to each state [21, 22]. To resample state–specific transition distributions π_k^d, we first count the number $n^d(k, \ell)$ of transitions from parent state k to child state ℓ, in direction d, instantiated by z. The posterior is then Dirichlet:

$$\pi_k^d \sim \mathcal{D}\left(n^d(k,1) + \alpha\beta_1, \ldots, n^d(k,K) + \alpha\beta_K\right) \quad (12)$$

Finally, the global mixture weights β, as well as the HDP concentration parameters γ and α, can be resampled via auxiliary variable methods [22].

6. Analysis of Natural Scenes

To evaluate the HDP-HMT, we compare steerable pyramid and SIFT descriptor representations of the eight natural scene categories defined by Oliva and Torralba [15]. Wavelet-domain features were extracted from 128×128 grayscale images, using 6 orientation, 4 level steerable pyramid transforms (sp5) [18]. Low–pass and high–pass residual bands were not modeled. SIFT descriptors were extracted on a dense grid from 256×256 grayscale images, at four resolutions produced by dyadic subsampling. We then used K–means clustering to create a 200–entry codebook from 50,000 randomly chosen training features.

6.1. Visualization of Learned Scene Statistics

In Fig. 3, we illustrate wavelet coefficient histograms [26] computed from images in two categories, "coast" and "tallbuilding." We compare this raw data to coefficients simulated from two models: the HDP-HMT, and a baseline bag–of–words (HDP-BOW) model (a nonparametric generalization of [6]). For the HDP-BOW, groups correspond to observed features at different scales, which are drawn from infinite mixtures whose components are shared across scales. For both models, we used 100 training images, and ran the Gibbs sampler for 500 iterations.

The HDP-HMT models the non–Gaussian "bow tie" shapes of wavelet histograms, and also accurately captures the complex orientation and scale relationships exhibited by steerable pyramids. However, it underestimates the strong positive correlations between adjacent horizontal and vertical coefficients in the horizontal and vertical finest scale subbands, respectively. This deficiency is partially due to the Markov tree boundaries which separate some pairs of fine scale coefficients [27].

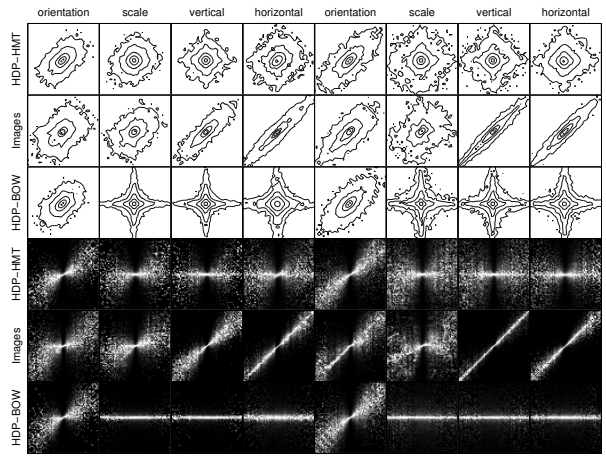


Figure 3. Pairwise histograms of steerable pyramid detail coefficients for 128×128 images from the "coast" (columns 1-4) and "tallbuilding" (columns 5-8) scene categories. Rows 2 & 5 are computed from observed images, while rows 3 & 6 and 1 & 4 summarize samples from bag–of–features (HDP-BOW) and hidden Markov tree (HDP-HMT) models, respectively. As in [26], we visualize log-contours of joint distributions (top) as well as normalized conditional distributions (bottom).

By construction, the HDP-BOW cannot capture any dependencies between coefficients at different locations or scales. It also captures qualitative differences among scene categories less accurately than the HDP-HMT. In particular, note that the contours for tallbuilding images are more elongated than those for coastal scenes, which contain less high-frequency content. Also, the vertically layered structure of large–scale environments [24] can be seen in the greater frequency of horizontal gradients in coast images. The inability of the HDP-BOW to capture scale and location correlations is also evident in the much less coherent *maximum a posteriori* (MAP) assignments of features to topics in test images (see Fig. 4). MAP assignments for the HDP-HMT, which were computed efficiently via the max-product algorithm [27], suggest that it captures interesting feature dependencies. For example, note the restoration of SIFT descriptors in coarse scale regions of the tallbuilding image corrupted by aliasing artifacts.

To further illustrate the nonparametric properties of the truncated HDP, we have trained models for two categories with varying numbers of training images. We counted the number of states with at least one assigned observation over 400 MCMC iterations, after discarding 100 burn-in iterations. Figure 5 plots the posterior mean number of hidden states versus training set size. Note that the complexity of this nonparametric model grows as the number of training images increases, adapting *automatically* to observed data. With additional sampling iterations, these growth rates would become smoother. In this experiment, the truncation level of $K = 200$ did not limit model expressiveness.

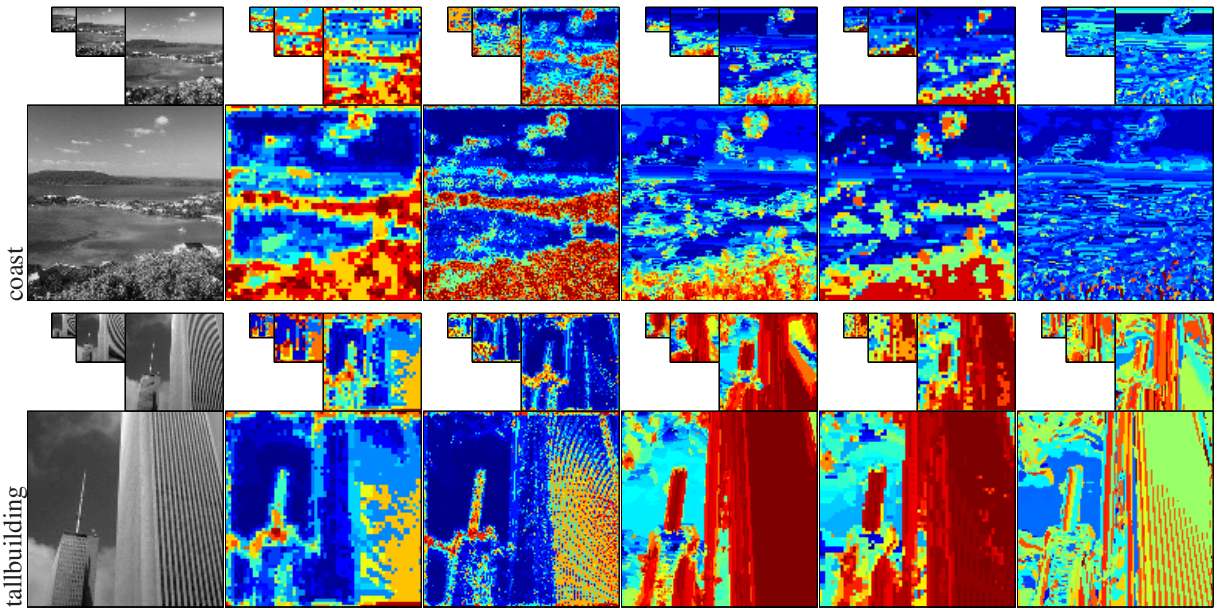


Figure 4. Maximum a posteriori joint hidden state assignments for HDP-BOW and HDP-HMT models of two test images. For steerable pyramid (sp5) features, states are sorted by the determinant of the emission distribution's covariance matrix. For SIFT features, states are sorted via a trimmed posterior mean of multinomial emission distributions (after sorting via the first principal component).

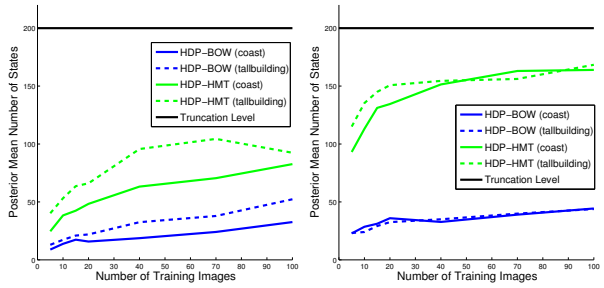


Figure 5. Posterior mean number of states used by the HDP-BOW and HDP-HMT models, for varying training set sizes. We compare steerable pyramid (sp5, left) and SIFT (right) models of the "coast" and "tallbuilding" scene categories. In all cases, the truncation level was set to $K = 200$.

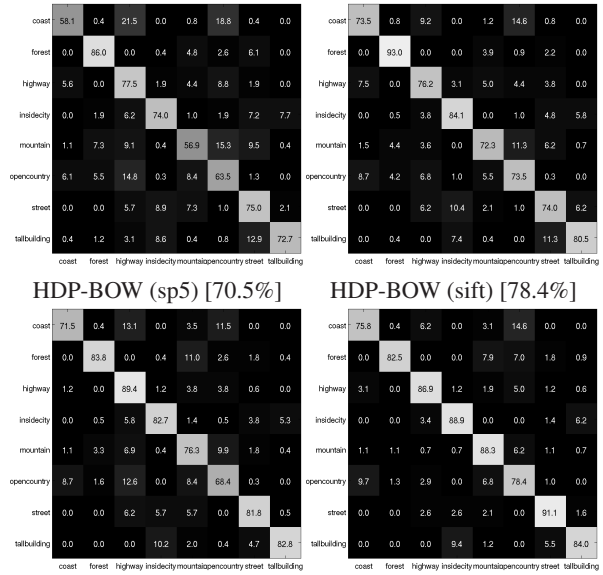
6.2. Scene Categorization Results

For our final experiments, we learned nonparametric models of both feature types for all eight scene categories [15]. We used 100 images of each category for training, and the remainder for test. For the HDP-HMT, we classified test images as the category which assigned the highest marginal likelihood to test features. These likelihoods can be efficiently computed in closed form with a single, coarse–to–fine belief propagation (BP) recursion [5, 27]. The confusion matrices of Fig. 6 compare the HDP-HMT's categorization performance to that of a baseline bag-of-words model (HDP-BOW) [3, 6]. Table 1 also summarizes average classification performance for the "natural" and "man-made" subsets of the scene categories [15]. We find that the more distinctive local features provided by SIFT descriptors lead to significant performance improvements. For both feature types, the HDP-HMT is much more accurate than the HDP-BOW, demonstrating the benefits of coupling local features with global spatial models. As in [3, 15], our HDP models have the most difficulty distinguishing the "coast" and "opencountry" categories.

Comparing our results to state-of-the-art methods employing grayscale features, the HDP-HMT performs comparably to the discriminative approach of [3], who in turn improved on [6] for a larger, thirteen category dataset. Note that we used only half (100 per category) as many training images as [3]. Furthermore, in contrast with nearest–neighbor [3, 15] and kernel [12] methods, our classifier does not need to store any training data for classification. Computation of one test image likelihood via BP, using fixed parameters for $K = 200$ states sampled during training, takes less than a second. As scene and object recognition systems are applied to larger datasets, such savings in storage and computation become increasingly important.

7. Discussion

We have developed a nonparametric, data–driven model for image features which captures spatial dependencies via a multiscale graphical model. Our results show that this HDP-HMT captures natural scene statistics more accurately than bag–of–feature models, and leads to improved catego-



HDP-BOW (sp5) [70.5%] HDP-BOW (sift) [78.4%]

HDP-HMT (sp5) [79.6%] HDP-HMT (sift) [84.5%]

Figure 6. Confusion matrices for the recognition of eight scene categories [15] using HDP-BOW (top) and HDP-HMT (bottom) models of wavelet (sp5, left) or SIFT (right) features. Average performance across all categories is shown in parentheses.

Dataset [15]	HDP-BOW (sp5, sift)		HDP-HMT (sp5, sift)	
All	70.46	78.38	79.58	84.49
Natural	75.58	82.90	80.89	84.42
Man-made	80.04	81.80	87.27	90.44

Table 1. Average scene categorization performance (mean of the confusion matrix's diagonal) for the groupings considered by [15].

rization performance. We are currently exploring the HDP-HMT's ability to learn richer appearance patterns from very large datasets, and model other families of visual scenes.

Acknowledgments

We acknowledge the support of ONR Grant N00014-06-1-0734 and the Defense Advanced Research Projects Agency under contract NBCHD030010. JJK was funded by the National Technology Agency of Finland and the Jenny & Antti Wihuri Foundation.

References

[1] K. Barnard et al. Matching words and pictures. *JMLR*, 3:1107–1135, 2003.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[3] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via pLSA. In *ECCV*, pages 517–530, 2006.

[4] H. A. Chipman, E. D. Kolaczyk, and R. E. McCulloch. Adaptive Bayesian wavelet shrinkage. *J. Amer. Stat. Assoc.*, 92(440):1413–1421, Dec. 1997.

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

[6] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, volume 2, pages 524–531, 2005.

[7] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google's image search. In *ICCV*, volume 2, pages 1816–1823, 2005.

[8] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[9] H. Ishwaran and M. Zarepour. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

[10] H. Ishwaran and M. Zarepour. Exact and approximate sum–representations for the Dirichlet process. *Can. J. Stat.*, 30:269–283, 2002.

[11] J. Kivinen, E. B. Sudderth, and M. I. Jordan. Image denoising with nonparametric hidden Markov trees. In *ICIP*, 2007.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2176, 2006.

[13] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001.

[14] D. G. Lowe. Distinctive image features from scale–invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[15] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[16] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Proc.*, 12(11):1338–1351, Nov. 2003.

[17] P. Quelhas et al. Modeling scenes with local descriptors and latent aspects. In *ICCV*, volume 1, pages 883–890, 2005.

[18] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multi-scale transforms. *IEEE Trans. Info. Theory*, 38(2):857–607, Mar. 1992.

[19] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *ICCV*, volume 1, pages 370–377, 2005.

[20] C. Spence, L. C. Parra, and P. Sajda. Varying complexity in tree–structured image distribution models. *IEEE Trans. IP*, 15(2):319–330, Feb. 2006.

[21] E. B. Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. PhD thesis, Massachusetts Institute of Technology, May 2006.

[22] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Amer. Stat. Assoc.*, 101(476):1566–1581, Dec. 2006.

[23] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):169–191, 2003.

[24] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Comp. Neural Sys.*, 14:391–412, 2003.

[25] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *IJCV*, 72(2):133–157, 2007.

[26] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky. Random cascades on wavelet trees and their use in analyzing and modeling natural images. *ACHA*, 11:89–123, 2001.

[27] A. S. Willsky. Multiresolution Markov models for signal and image processing. *Proc. IEEE*, 90(8):1396–1458, 2002.