ter of imagery. Recently Li et al. (2010) employed a nested Chinese restaurant process (nCRP) to infer a hierarchical tree representation for a corpus of images and (if available) accompanying text; however, in that work the VQ step was still employed, and therefore a precomputation of features was as well. Further, in Li et al. (2010), while the tree width was inferred, the depth was set. Finally, the nCRP construction in Li et al. (2010) has the disadvantage of only updating parent-child-transition parameters from one node of the tree at a time, in a sampler, yielding poor mixing relative to a stick-breaking Dirichlet process (DP) implementation (Ishwaran & James, 2001). Related but distinct dictionary learning with the nCRP was considered in Zhang et al. (2011).

Motivated by these recent contributions, and the limitations of most existing topic models of imagery and text, this paper makes the following contributions:

- A nested DP (nDP) model is developed to learn a hierarchical tree structure for a corpus of imagery and text, with a stick-breaking construction employed; we infer both the tree depth and width, using a retrospective stick-breaking construction (Papaspiliopoulos & Roberts, 2008).

- A beta-Bernoulli dictionary learning framework (Zhou et al., 2011b) is adapted to such a hierarchical model, removing the VQ step, and allowing one to perform topic modeling directly on image patches, thereby integrating feature design and topic modeling. However, if desired, the dictionary learning may also be applied to features pre-computed from the image, using any existing method for feature design, and again removing the limitations of VQ.

2 Modeling Image Patches

We wish to build a hierarchical model to arrange $M$ images and their associated annotations (when available); the vocabulary of such annotations is assumed to be of dimension $N_v$. The vector $x_{mi}$ represents the pixels or features associated with the $i$th patch in image $m$, and $y_m = (y_{m1}, \ldots, y_{mN_v})^T$ represents a vector of word counts associated with that image, when available ($y_{mn}$ represents the number of times word $n \in \{1, \ldots, N_v\}$ is present in the annotation).

The $m$th image is divided into $N_m$ patches (or super-pixels (Li et al., 2010)), and the data for the $i$th patch is denoted $x_{mi} \in \mathbb{R}^P$ with $i = 1, \ldots, N_m$. The vector $x_{mi}$ may represent raw pixel values, or a feature vector extracted from the pixels (using any available method of image feature extraction, e.g., SIFT (Lowe, 1999)). Each $x_{mi}$ is represented as a sparse linear combination of learned dictionary atoms. Further, each patch is assumed associated with a “topic”; the probability of which dictionary atoms are employed for a given patch is dictated by the topic it is associated with.

Specifically, each patch is represented as $x_{mi} = D(z_{mi} \odot s_{mi}) + e_{mi}$, where $\odot$ represents the element-wise/Hadamard product, $D = [d_1, \ldots, d_K] \in \mathbb{R}^{P \times K}$, $K$ is the truncation level on the possible number of dictionary atoms, $z_{mi} = [z_{m1}, \ldots, z_{miK}]^T$, $s_{mi} = [s_{m1}, \ldots, s_{miK}]^T$, $z_{mik} \in \{0, 1\}$ indicates whether the $k$th atom is active within patch $i$ in image $m$, $s_{mik} \in \mathbb{R}^+$, and $e_{mi}$ is the residual. Note that $z_{mi}$ represents the specific sparseness pattern of dictionary usage for $x_{mi}$. The hierarchical form of the model is

$$
\begin{align*}
    x_{mi} &\sim \mathcal{N}(D(z_{mi} \odot s_{mi}), \gamma_e^{-1}I_P) \\
    d_k &\sim \mathcal{N}(0, \frac{1}{L_k}I_P) \\
    s_{mi} &\sim \mathcal{N}_+(0, \gamma_s^{-1}I_K) \\
    z_{mi} &\sim \prod_{k=1}^K \text{Bernoulli}(\pi_{hmik}) \tag{1}
\end{align*}
$$

where gamma priors are placed on both $\gamma_e$ and $\gamma_s$. Positive weights $s_{mi}$ (truncated normal, $\mathcal{N}_+(\cdot)$) are imposed, which we have found to yield improved results.

The indicator variable $h_{mi}$ defines the topic associated with $x_{mi}$. The $K$-dimensional vector $\pi_h$ defines the probability that each of the $K$ columns of $D$ is employed to represent topic $h$, where the $k$th component of $\pi_h$ is $\pi_{hk}$. These probability vectors are drawn

$$
\pi_h \sim G_0, \quad G_0 = \prod_{k=1}^K \text{Beta}(a_0/K, b_0(K-1)/K) \tag{2}
$$

where $\pi_{hk}$ represents the probability of using $d_k$ for object type $h$, and the introduction of $G_0$ is for discussions below. This representation for $\pi_h$ corresponds to an approximation to the beta-Bernoulli process (Thibaux & Jordan, 2007; Paisley & Carin, 2009; Zhou et al., 2011a,b), which also yields an approximation to the Indian buffet process (IBP) (Griffiths & Ghahramani, 2005; Teh et al., 2007).

3 Tree Structure via nDP

The nested Dirichlet process (nDP) tree construction developed below is an alternative means of constituting the same type of tree manifested by the nested Chinese restaurant process (Blei et al., 2003a; Li et al., 2010). We emphasize the nDP construction because of the stick-breaking implementation we employ, which