2.2 Object Label Frequencies

Pitman–Yor processes have been previously used to model the well-known power law behavior of text sequences [15, 16]. Intuitively, the labels assigned to segments in the natural scene database have similar properties: some (like *sky*, *trees*, and *building*) occur frequently, while others (*rainbow*, *lichen*, *scaffolding*, *obelisk*, etc.) are more rare. Fig. 1(b) plots the observed frequencies with which unique text labels, sorted from most to least frequent, occur in two scene categories. The overlaid quantiles correspond to the best fitting DP and PY processes, with parameters $(\hat{\gamma}_a, \hat{\gamma}_b)$ estimated via maximum likelihood. When $\hat{\gamma}_a > 0$, $\log \mathbb{E}[\widetilde{\varphi}_k \mid \hat{\gamma}] \approx -\hat{\gamma}_a^{-1} \log(k) + \Delta(\hat{\gamma}_a, \hat{\gamma}_b)$ for large k [11], producing power law behavior which accurately predicts observed object frequencies. In contrast, the closest fitting DP model $(\hat{\gamma}_a = 0)$ significantly underestimates the number of rare labels.

We have quantitatively assessed the accuracy of these models using bootstrap significance tests [17]. The PY process provides a good fit for all categories, while there is significant evidence against the DP in most cases. By varying PY hyperparameters, we also capture interesting differences among scene types: urban, man-made environments have many more unique objects than natural ones.

2.3 Segment Counts and Size Distributions

We have also used the natural scene database to quantitatively validate PY priors for image partitions [17]. For natural environments, the DP and PY processes both provide accurate fits. However, some urban environments have many more small objects, producing power law area distributions (see Fig. 1(c)) better captured by PY processes. As illustrated in Fig. 1(d), PY priors also model uncertainty in the *number* of segments at various resolutions.

While power laws are often used simply as a descriptive summary of observed statistics, PY processes provide a consistent generative model which we use to develop effective segmentation algorithms. We do not claim that PY processes are the only valid prior for image areas; for example, log-normal distributions have similar properties, and may also provide a good model [18]. However, PY priors lead to efficient variational inference algorithms, avoiding the costly MCMC search required by other segmentation methods with region size priors [18, 19].

3 A Hierarchical Model for Bags of Image Features

We now develop *hierarchical Pitman–Yor* (HPY) process models for visual scenes. We first describe a "bag of features" model [1, 2] capturing prior knowledge about region counts and sizes, and then extend it to model spatially coherent shapes in Sec. 4. Our baseline bag of features model directly generalizes the stick-breaking representation of the hierarchical DP developed by Teh et al. [12]. N-gram language models based on HPY processes [15, 16] have somewhat different forms.

3.1 Hierarchical Pitman–Yor Processes

Each image is first divided into roughly 1,000 *superpixels* [18] using a variant of the normalized cuts spectral clustering algorithm [13]. We describe the texture of each superpixel via a local texton histogram [20], using band-pass filter responses quantized to $W_t = 128$ bins. Similarly, a color histogram is computed by quantizing the HSV color space into $W_c = 120$ bins. Superpixel i in image j is then represented by histograms $x_{ji} = (x^t_{ji}, x^c_{ji})$ indicating its texture x^t_{ji} and color x^c_{ji}.

Figure 2 contains a directed graphical model summarizing our HPY model for collections of local image features. Each of the potentially infinite set of global object categories occurs with frequency φ_k, where $\boldsymbol{\varphi} \sim \mathrm{GEM}(\gamma_a, \gamma_b)$ as motivated in Sec. 2.2. Each category k also has an associated appearance model $\theta_k = (\theta^t_k, \theta^c_k)$, where θ^t_k and θ^c_k parameterize multinomial distributions on the W_t texture and W_c color bins, respectively. These parameters are regularized by Dirichlet priors $\theta^t_k \sim \mathrm{Dir}(\rho^t)$, $\theta^c_k \sim \mathrm{Dir}(\rho^c)$, with hyperparameters chosen to encourage sparse distributions.

Consider a dataset containing J images of related scenes, each of which is allocated an infinite set of potential segments or *regions*. As in Sec. 2.3, region t occupies a random proportion π_{jt} of the area in image j, where $\boldsymbol{\pi}_j \sim \mathrm{GEM}(\alpha_a, \alpha_b)$. Each region is also associated with a particular global object category $k_{jt} \sim \boldsymbol{\varphi}$. For each superpixel i, we then *independently* select a region $t_{ji} \sim \boldsymbol{\pi}_j$, and sample features using parameters determined by that segment's global object category:

$$p\big(x^t_{ji}, x^c_{ji} \mid t_{ji}, \mathbf{k}_j, \boldsymbol{\theta}\big) = \mathrm{Mult}\big(x^t_{ji} \mid \theta^t_{z_{ji}}\big) \cdot \mathrm{Mult}\big(x^c_{ji} \mid \theta^c_{z_{ji}}\big) \qquad z_{ji} \triangleq k_{jt_{ji}} \tag{2}$$

As in other adaptations of topic models to visual data [8], we assume that different feature channels vary independently within individual object categories and segments.