Supervised Hierarchical Pitman-Yor Process for Natural Scene Segmentation

Alex Shyr
UC Berkeley

Trevor Darrell
ICSI, UC Berkeley

Michael Jordan
UC Berkeley

Raquel Urtasun
TTI Chicago

Abstract

From conventional wisdom and empirical studies of annotated data, it has been shown that visual statistics such as object frequencies and segment sizes follow power law distributions. Using these two as prior distributions, the hierarchical Pitman-Yor process has been proposed for the scene segmentation task. In this paper, we add label information into the previously unsupervised model. Our approach exploits the labelled data by adding constraints on the parameter space during the variational learning phase. We evaluate our formulation on the LabelMe natural scene dataset, and show the effectiveness of our approach.

1. Introduction

The problem of image segmentation and grouping remains one of the important challenges in computer vision, as segmenting a scene into semantic categories is one of the key steps towards scene understanding. As evidenced by the PASCAL VOC challenge [7], segmentation is still an unsolved problem - the accuracy of existing approaches is still insufficient for integration into real-world applications, e.g., robotics.

In the past few years, approaches based on Markov random fields (MRF) have been popular for segmentation [13] [11] [15]. In these approaches, the image is modelled as an undirected graphical model, with nodes being pixels and/or superpixels. Node potentials are defined in terms of the local evidence, and edge potentials are defined to encourage smoothness in the segmentation. The resulting inference problem is either solved by graph-cuts [4] or message-passing algorithms [9].

While very effective for certain tasks, these probabilistic models do not reflect the underlying statistics of natural images. Recent studies show that a wide range of natural image statistics are distributed according to heavy-tailed distributions. This problem has been noticed not only for segmentation, but also for optical flow (denoising) [25], intrinsic images [33] and layer extraction [1]. Moreover, long range dependencies are difficult to capture with MRFs.

The gPb method of [17] computes long-range interactions by building an affinity matrix from local cues via the Pb response[18] and computing gradients of the corresponding eigenvectors. These gradients are then combined with local feature gradients to obtain the final gPb function. [2] applies the oriented watershed transform (OWT) of the gPb response to form regions, and subsequently construct the ultrametric contour map (UCM), defining a hierarchical segmentation. We adopt the gPb function as a basic boundary model, and we demonstrate in our experiments that a probabilistic model with a prior which succinctly describes segment statistics achieves better performance than the OWT-UCM model.

Sudderth and Jordan [30] proposed an unsupervised probabilistic model for segmentation that is based on the Hierarchical Pitman-Yor process (HPY), which is a non-parametric Bayesian prior over infinite partitions. The HPY process is a generalization of the hierarchical Dirichlet processes (HDP), with heavier-tailed power law prior distributions. Confirming the findings of Sudderth et al., we show that the distribution over the size of natural segments as well as the frequencies that objects appear in an image follow a power law distribution. Long range dependencies are introduced in their framework via thresholded Gaussian processes. Their approach, however, is unsupervised, and does not leverage the ever growing abundance of annotations and ground truth data, e.g., LabelMe. As a consequence, the inferred segmentations are not always accurate and have room for improvement.

In this paper we propose a novel supervised discriminative Hierarchical Pitman-Yor process (DHPY) approach to segmentation. In particular, we frame the learning as a regularized constrained optimization problem, where we maximize a variational lower bound on the log likelihood while imposing the inferred labels at the segment and object levels agree with the ground truth annotations. We borrow intuitions from the literature of cutting plane and subgradient optimization methods, and derive an efficient method to train the HPY. At every step of the algorithm the most violated constraint is introduced into the optimization via Lagrange multipliers. While we leverage the additional annotations, we also inherit the nice properties of [30]; more specifically, we are able to capture long range dependencies

via thresholded Gaussian process, and we retain the natural power-law priors.

We demonstrate the effectiveness of our approach in a dataset composed of 8 different types of scenes and 100 object categories taken from the LabelMe dataset [26]. Our approach outperforms normalized cuts [28], the Ultrametric Contour Map approach of [2], as well as the unsupervised HPY [30]. In the remainder of the paper we first review the related work. We then introduce the HPY process, derive our supervised DHPY formulation, show empirical results, and conclude with avenues of future research.

2. Related Work

Markov Random Fields (MRFs) have become a popular approach to segmentation, as demonstrated by the large body of work [13, 11, 15]. In these approaches the image is modeled as an undirected graphical model at the level of pixels and/or superpixels. Node potentials describe local evidence and edge potentials usually encourage smoothness for neighboring pixels/superpixels with the same label. Several approaches to inference have been proposed for the MRF, such as graph cuts [4] and belief propagation [9].

One particular form of MRF that directly defines a discriminative distribution of the latent states is the Conditional Random Field (CRF) [29]. Inference is made easier since the conditional probabilities of the latent labels given the observations are modelled directly. Another way that supervision is used is through the fusing of contextual information [15]. Context can be added in the form of global constraints which usually specify class-co-occurrence and/or conditional dependence in the form of clique structure.

While effective for segmentation, MRFs have been shown to be inadequate for modelling the visual statistics of natural scenes [27]. In this paper, we build on top of the Hierarchical Pitman-Yor processes, which accurately model the power law prior distributions, such as the distributions over the number of objects per image as well as that of the size of natural segments.

Recently, [5, 6] show that segmentation can be framed as a two step process. First, candidate segments which can be part of an object are identified - [5] employs a graph-cut optimization framework, while [6] uses a CRF model. The segmentation problem is then formulated as a ranking problem over the candidate segments, which involves computing segment-level features such as segment area, perimeter and shape statistics.

Sudderth and Jordan [30] proposed an unsupervised approach to image segmentation that models segments of visual scenes with a hierarchical Pitman-Yor process (HPY). Thresholded Gaussian process are utilized to capture spatial coherence among regions. Moreover, this captures long-range dependences among the observations, which are difficult to achieve with MRFs. In this paper we propose a novel supervised HPY process framework for segmentation, and show that the use of annotations significantly improves the performance over the unsupervised HPY.

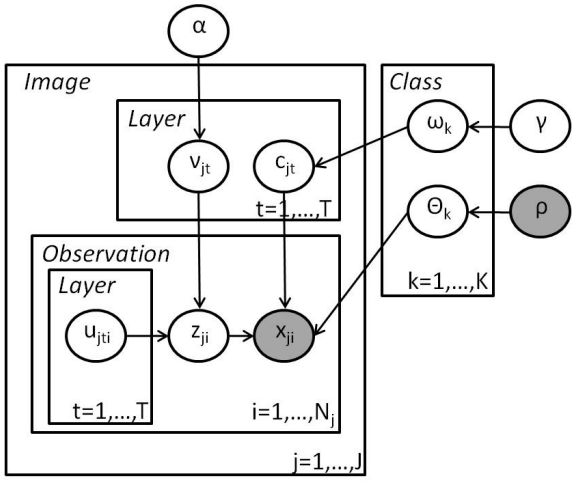


Figure 1: Graphical model of the Hierarchical Pitman-Yor Process (HPY), applied to natural scene segmentation [30]. Each observation x_{ji} is assigned to layer z_{ji}, an indicator variable. The assignment depends on the thresholded GPs $\mathbf{u}_{jt}$ and layer probabilities v_{jt}, which are generated from the PY stick-breaking prior $GEM(\alpha_a, \alpha_b)$. Each layer is assigned to class c_{jt}, another indicator variable, which follow the PY prior $GEM(\gamma_a, \gamma_b)$ where w_k are the stick lengths. Each class has an associated appearance model θ_k.

Discriminative nonparametric Bayesian models have been proposed in the context of latent variable models, e.g., Latent Dirichlet Allocation (LDA) [3, 14]. These approaches modify the graphical model, by adding either a generative distribution of the label given the latent state [3] or a discriminative distribution of the latent state given the label [14]. However, it is not easy to pick a suitable distribution, such as generalized linear models (GLM). There is also the question of how much discriminative power the modified likelihood can exert on the latent states. Unlike these approaches, we propose to maximize a variational lower bound on the log likelihood while imposing the inferred label assignments coincide with the ground truth annotations. We derive an efficient method to train the HPY model. In contrast to other supervised approaches to nonparametric Bayesian models, our discriminative approach makes use of supervised data directly resulting in significant performance improvements over the unsupervised model.

3. A Review of Unsupervised HPY Processes

We now describe the hierarchical Pitman-Yor (HPY) model for visual scenes [30], which is a generalization of the hierarchical Dirichlet process (HDP)[31]. In the next section we will introduce our new supervised HPY

model. The Pitman-Yor process [22], denoted by $\phi \sim GEM(\gamma_a, \gamma_b)$, places a prior distribution over partitions with hyperparameters γ_a, γ_b satisfying $0 \leq \gamma_a < 1$ and $\gamma_b > -\gamma_a$. It can be defined using the stick-breaking construction as

$$\begin{aligned} \phi_k &= w_k \prod_{l=1}^{k-1}(1-w_l) = w_k(1-\sum_{l=1}^{k-1}\phi_l), \qquad \text{with} \\ w_k &\sim Beta(1-\gamma_a, \gamma_b + k\gamma_a). \end{aligned} \tag{1}$$

The $\{\phi_k\}$ are the partition probabilities, while the $\{w_k\}$ are the stick lengths. Note that we recover a Dirichlet process, specified by a single concentration parameter γ_b, when $\gamma_a = 0$. When $\gamma_a > 0$, the partition probabilities follow a power-law distribution with a heavy tail. While the PY process is a prior on infinite partitions, only a finite subset of partitions will have positive probabilities greater than a threshold ϵ. Hence, the PY process implicitly imposes a prior on the number of partitions.

In the HPY model, two Pitman-Yor process priors are placed over the distributions of global class categories and segment proportions. Fig. 1 shows the directed graphical model. Each image is segmented into superpixels, which are from now on treated as the observed data units x_{ji}. Each data point is then assigned to a layer with probability

$$P[z_{ji} = t | z_{ji} \neq t-1, \ldots, 1] = P[u_{jti} < \Phi^{-1}(v_{jt})] = v_{jt},$$

where we have introduced a zero mean Gaussian process (GP) $\mathbf{u}_{jt}$ for each layer t. These thresholded GPs completely determine the layer assignment of each superpixel, with the assignment rule being

$$z_{ji} = \min\{t | u_{jti} < \Phi^{-1}(v_{jt})\}. \tag{2}$$

Each layer is associated with a global object class c_{jt} with an appearance model θ_k. The emission probability is then

$$p(x_{ji} | z_{ji} = t, c_{jt} = k, \boldsymbol{\theta}) = Mult(x_{ji} | \theta_k). \tag{3}$$

with $\boldsymbol{\theta} = \{\theta_1, \cdots\}$ To place PY priors on the distributions over global class categories and segment proportions, the class assignments c_{jt} are sampled from $\boldsymbol{\phi} \sim GEM(\gamma_a, \gamma_b)$, which is the stick-breaking prior described above, with w_k the stick length. Similarly, the layer assignment probabilities v_{jt} are sampled from $\boldsymbol{\pi} \sim GEM(\alpha_a, \alpha_b)$.

4. Supervised Hierarchical Pitman-Yor Model

In this section we present our supervised hierarchical Pitman-Yor process model for image segmentation. We first derive our variational learning approach and show how to incorporate supervision by solving a constrained optimization problem. We then derive a cutting plane method to efficiently learn the model.

Following [30], we train the HPY model with a mean field variational approximation. A completely factorized variational posterior is introduced as follows

$$q(\mathbf{u}, \mathbf{v}, \mathbf{c}, \mathbf{w}, \boldsymbol{\theta}) = \prod_{k=1}^{K} q(w_k|\omega_k) q(\theta_k|\eta_k) \times$$
$$\times \prod_{j=1}^{J}\prod_{t=1}^{T} q(v_{jt}|\nu_{jt}) q(c_{jt}|\kappa_{jt}) \prod_{i=1}^{N_j} q(u_{jti}|\mu_{jti}),$$

where the distributions are, with $\bar{v}_{jt} = \Phi^{-1}(v_{jt})$,

$$\begin{aligned} q(\theta_k|\eta_k) &= Dir(\eta_k) \\ q(\mathbf{c}_j|\boldsymbol{\kappa}_j) &= Mult(\mathbf{c}_j|\boldsymbol{\kappa}_j) \\ q(w_k|\omega_{k,a}, \omega_{k,b}) &= Beta(w_k|\omega_{k,a}, \omega_{k,b}) \\ q(\bar{v}_{jt}|\nu_{jt}, \delta_{jt}) &= N(\bar{v}_{jt}|\nu_{jt}, \delta_{jt}) \\ q(u_{jti}|\mu_{jti}, \lambda_{jti}) &= N(u_{jti}|\mu_{jti}, \lambda_{jti}). \end{aligned}$$

We truncate the variational posterior by setting $q(v_{jT} = 1) = 1$ and $q(w_K = 1) = 1$. We then train the model by optimizing the lower bound on the marginal likelihood

$$\log p(\mathbf{x}|\alpha, \gamma, \rho) \geq H(q) + E_q[\log p(\mathbf{x}, \mathbf{z}, \mathbf{u}, \mathbf{v}, \mathbf{c}, \mathbf{w}, \theta|\alpha, \gamma, \rho)]$$
$$= E_q[\log p(\mathbf{x}, \mathbf{z}, \mathbf{u}, \mathbf{v}, \mathbf{c}, \mathbf{w}, \theta|\alpha, \gamma, \rho)] - E_q[\log q(\mathbf{u}, \mathbf{v}, \mathbf{c}, \mathbf{w}, \theta)] \equiv \mathcal{L}$$

This is equivalent to minimizing the KL-divergence between p and q. The optimization is done through a combination of closed-form updates and gradient descent.

Inference in the unsupervised HPY model produce layer-level and class-level segmentations using the variational marginals $\arg\max_t P_q(z_{ji} = t)$ and $\arg\max_k P_q(c_{jt} = k)$, where

$$P_q(z_{ji} = t) = \Phi\left(\frac{\nu_{jt} - \mu_{jti}}{\sqrt{\delta_{jt} + \lambda_{jti}}}\right) \prod_{\tau=1}^{t-1}\left(1 - \Phi\left(\frac{\nu_{j\tau} - \mu_{j\tau i}}{\sqrt{\delta_{j\tau} + \lambda_{j\tau i}}}\right)\right)$$
$$P_q(c_{jt} = k) = \kappa_{jtk}$$
$$= \exp\{\sum_{i,l} x_{ji,l} P_q(z_{ji} = t)\, E_q \log \eta_{k,l} + \sum_{k'=1}^{k} E_q \log \omega_{k',l}\},$$

with $E_q \log \eta_{k,l} = \Psi(\eta_{k,l}) - \Psi(\sum_l \eta_{k,l})$ and $E_q \log \omega_{k',l}$ $= \Psi(\omega_{k',t}) - \Psi(\omega_{k',t} + \omega_{k',b})$ from the Dirichlet posterior assumption. The last equation stems from the closed-form update for κ_{jtk}.

Let $\mathcal{A}$ be the set of annotations. We are provided with two types of annotations:

$$\mathcal{A} = \left\{(a^s_{ji}, a^c_{ji}) | a^s_{ji} \in \{1, \ldots, T\}, a^c_{ji} \in \{1, \ldots, K\}\right\}$$

where a^s_{ji} : segment-level annotation, and
a^c_{ji} : class-level annotation.

Segment-level annotations describe the layer assignment of each observation, while class-level annotation describes the class assignment of each layer. Note that we have imposed an absolute ordering on the layers, due to the stick-breaking

construction of the layer model; in practice, we sort the different layers in decreasing order of their sizes.

We apply supervision constrains that are added to the variational program. Learning the supervised HPY can then be formulated as the following maximization problem

$$\begin{aligned} \max \quad & \mathcal{L} \\ s.t. \quad & \forall (j,i) \in \mathbf{A}, P_q(z_{ji} = a^s_{ji}) \geq \max_t P_q(z_{ji} = t) \\ & \forall (j,i) \in \mathbf{A}, P_q(c_{ja^s_{ji}} = a^c_{ji}) \geq \max_k P_q(c_{ja^s_{ji}} = k) \end{aligned}$$

with respect to μ_{jti}, λ_{jti}, ν_{jt} and δ_{jt}. Note that the above probabilities have been defined perviously in terms of these variables.

We transform the above optimization problem into a single objective by adding slack variables

$$\begin{aligned} \max \quad & \mathcal{L} - \textstyle\sum_{(i,j)\in\mathcal{A}}(C^s\zeta^s_{ji} + C^c\zeta^c_{ji}) \\ s.t. \quad & \forall (j,i) \in \mathbf{A}, P_q(z_{ji} = a^s_{ji}) + \zeta^s_{ji} \geq \max_t P_q(z_{ji} = t) \\ & \forall (j,i) \in \mathbf{A}, P_q(c_{ja^s_{ji}} = a^c_{ji}) + \zeta^c_{ji} \geq \max_k P_q(c_{ja^s_{ji}} = k) \end{aligned}$$

The Lagrangian ℓ can then be defined as

$$\begin{aligned} \ell = \mathcal{L} - C^s \sum_{(i,j)\in\mathcal{A}} (\max_t P_q(z_{ji} = t) - P_q(z_{ji} = a^s_{ji})) \\ -C^c \sum_{(i,j)\in\mathcal{A}} (\max_k P_q(c_{ja^s_{ji}} = k) - P_q(c_{ja^s_{ji}} = a^c_{ji})) \end{aligned}$$

Maximizing the Lagrangian defines the optimization problem we solve to learn the discriminative HPY model. The coefficients C^s and C^c determine the relative weighting the model puts on minimizing the KL divergence and minimizing the segmentation error.

Learning: In the unsupervised model gradient descent is carried out independently for each image and layer. With the segment-level and class-level constraints, the layers and images are now dependent, complicating the optimization. Since adding all constraints to the optimization is computationally expensive, we derive a cutting plane type algorithm that selects, during each iteration, the most violated constraint and adds it to the optimization. We now explain the learning process in detail (see also Table 1).

First, we initialize the appearance model multinomial parameters, setting them according to the class-level annotations. Note that it is possible that our truncated value for the number of classes, K, is less than the number of global class categories. We first sort the classes in descending frequency, and lump the truncated classes into a "background" class. For classes that have not been observed in the training data, we randomly sample their appearance model from the background class, which may exhibit extensive intra-class variability. We then initialize all the other variational posteriors.

Within each image, we first train the variational parameters in the unsupervised fashion via gradient descent on the objective $\mathcal{L}$. Next, we permute the layers to minimize the number of violated constraints. This is necessary since we imposed an absolute ordering on the layers. The problem of computing the optimal permutation can be formulated as a bipartite matching on a graph where the nodes are the assignment labels and the edge weights are the number of agreements in the layer assignments - or more simply, the confusion matrix. The intuition behind this reformulation is that a permutation is an independent edge set (or a *matching*) since each assignment id can only be permuted to one other id, and vice versa. The matching is carried out with the Hungarian algorithm [12].

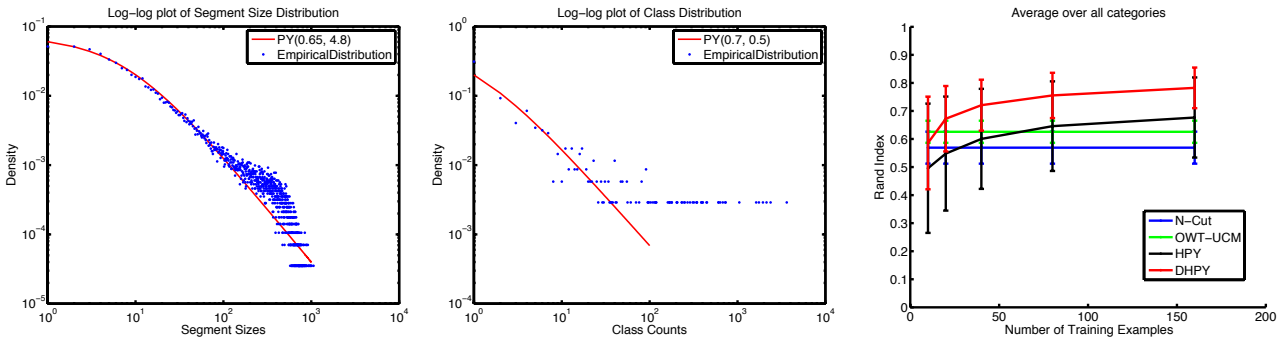
Once the layers are properly permuted, we iteratively identify the pair of layers with the most violated constraints, as motivated by the cutting-plane method in [10]. This corresponds to finding the largest off-diagonal entry in the new confusion matrix. Given this pair of layers, we find the set of observations which violate these segment-level constraints, as well as the subset of observations which violate

Algorithm 1: DHPY

for each k = observed, non-background class
  Set $\eta_k = \frac{\sum_{j,i} \mathbf{1\{a^c_{ji}=k\}x_{j,i}}}{\sum_{j,i} \mathbf{1\{a^c_{ji}=k\}}}$
for each k = unobserved class
  Initialize η_k randomly from background class
Initialize table assignments
Initialize class assignments κ_{jt} from a^c_{ji}
Initialize $\mu_{jti}, \lambda_{jti}, \nu_{jt}, \delta_{jt}$
for each image j = 1 to J
  Run unsupervised HPY training
  for each i = 1 to N_j
    Set assign(i) = $P_q(\min\{t|u_{jti} < \Phi^{-1}(v_{jt})\} = t)$
  Set M = getOptimalPermutation(assign, a^s_{ji})
  Permute layers according to M
  Construct new confusion matrix C
  Do while $\sum_{s\neq t} C(s,t)$ stabilizes:
    Set s^*, t^* = $\arg\max_{s\neq t} C(s,t)$
    Set $k^* = \arg\max_k P_q(c_{j,s^*} = k)$
    Set $\mathbf{I}^* = \{i|P_q(z_{ji} = t^*) < P_q(z_{ji} = s^*)\}$
    Set $\mathbf{I}^{**} = \{i \in \mathbf{I}^*|P_q(c_{j,s^*} = a^c_{ji}) < P_q(c_{j,s^*} = k^*)\}$
    Set $\partial\mathcal{L}^s = \partial\mathbf{1}_{\{i\in\mathbf{I}^*\}}(P_q(z_{ji} = t^*) - P_q(z_{ji} = s^*))$
    Set $\partial\mathcal{L}^c = \partial\mathbf{1}_{\{i\in\mathbf{I}^{**}\}}(P_q(c_{j,s^*} = k^*) - P_q(c_{j,s^*} = a^c_{ji}))$
    Run gradient descent with $\partial\mathcal{L} - C^s\partial\mathcal{L}^s - C^c\partial\mathcal{L}^c$
end

function getOptimalPermutation(assign, assign_gt)
Construct confusion matrix C:
for n = Range(assign)
  for m = Range(assign_gt)
    $C(n,m) = |\{(i,j)|\text{assign}(i) = n, \text{assign\_gt}(j) = m\}|$
Run Hungarian algorithm with weight matrix C
return matching M

Table 1: Discriminative HPY (DHPY) learning algorithm using Gradient Descent



(a) Segment Sizes (b) Class Frequencies (c) Rand Index

Figure 2: (a,b) Power-law Empirical Distributions across all categories and their fitted Pitman-Yor processes. (c) Performance in terms of Rand Index as a function of the number of training examples aggregated across categories

the class-level constraints. We now perform gradient descent on these sets of nodes, using the subgradient method suggested in [23] which approximates

$$\frac{\partial}{\partial x}\max f(x) \quad \approx \quad \max \frac{\partial f(x)}{\partial x} \in \mathcal{S},$$

where $\mathcal{S}$ is the set of subgradients of $\max f(x)$. We proceed until there are no more violated constraints. In practice, however, this is never accomplished; hence, we iterate until the number of violated constraints stabilizes.

Modeling Spatial Dependencies: We employ a few mechanisms to capture spatial dependencies. First, a bottom-up approach preprocess the data into superpixels [8, 19, 24, 20, 16], over-segmenting each image into locally consistent regions. In our experiments, we use TurboPixels [16]. Recall that in the HPY model, each layer is associated with a zero mean GP over $\mathbf{u}_{jt}$. If the GPs have diagonal covariance functions, the model is spatially independent. More general covariances can encode affinities among pairs of data features. In particular, for an image j, we employ a covariance that incorporates intervening contour cues based on the gP_b detector [2],

$$W_j(x_i, x_{i'}) = \exp\{-\frac{||\bar{x_i} - \bar{x_{i'}}||^2}{2\sigma_{sp}^2}\}(1 - gP_b(x_i, x_{i'}))^{\sigma_{gpb}},$$

where σ_{sp} and σ_{gpb} are constants, $\bar{\mathbf{x}}_i$ is the centroid of the i-th superpixel and $gP_b(x_i, x_{i'})$ is the maximal gP_b response along the line between the two superpixels' centroids. To induce sparsity, we also included a neighborhood parameter, ϵ_n; the covariance entry is zero if the corresponding superpixel centroids are more than ϵ_n pixels apart; otherwise, the value is the same as above. Since W_j is a covariance matrix, it is required to be positive semi-definite (PSD). To ensure that the covariance is PSD, we compute the eigendecomposition of W_j and retain only the eigenvalues that are at least ϵ_{eig} times the maximal eigenvalue. This is done for robustness and computational reasons (the feature dimension becomes smaller).

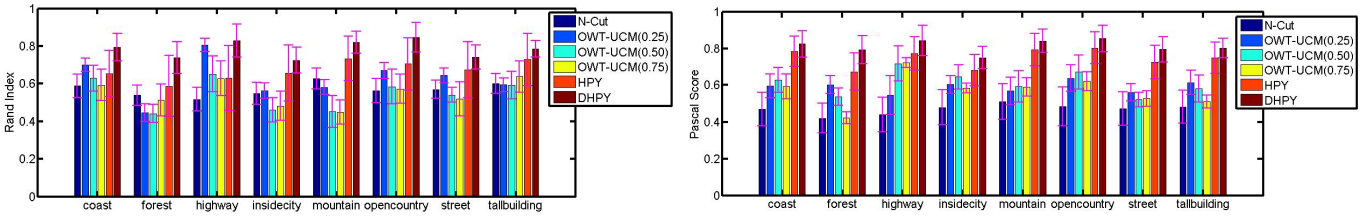
5. Experimental Evaluation

We validate our approach on the natural scene dataset of [21, 30], which is a subset of the LabelMe [26] database. The dataset consists of eight categories and a total of 2,688 images, each comprising a number of manually segmented polygons with a semantic text label. For all experiments, we use $\sigma_{sp} = 300$, $\sigma_{gpb} = 0.3$, $\epsilon_n = 200$, and $\epsilon_{eig} = 0.001$, which are estimated via cross-validation.

Each image is first preprocessed into roughly 1,000 superpixels. We use a local texton histogram quantized to 64 bins and a color histogram quantized to 100 bins as features. For each image, the segments are sorted in decreasing order, with the largest segment assigned to be layer 1. The segment label for each superpixel is then computed as the majority segment id among the encompassing pixels. The class label is taken to be the most frequent unigram, accounting for plurals and ignoring labels marked as occluded. The empirical distributions of the segment sizes (in terms of superpixels) and class counts are shown in Fig. 2. The asymptotic linearity in the loglog plot is evidence of a power law distribution; fitted Pitman-Yor priors are shown.

A single HPY and DHPY model is trained for each category. In our experiments, we set the number of global classes K to be 100 (across all scene categories); we bundle the truncated classes into a background class. For computation reasons, we set the number of segments T to be the same as the number of global classes. The crucial class-level PY hyperparameters γ_a, γ_b are set to their fitted values (0.7 and 0.5 respectively). Similarly, the segment-level PY hyperparameters α_a, α_b are set to the globally fitted values (0.65 and 4.8).

The segmentation is computed as the class with the maximal posterior probability of assignment. Results are reported with respect to the ground truth (manual segmentation) in terms of two metrics: the Rand index [32] and the Pascal score. The Rand index measures the similarity between two data clustering schemes. Given two cluster assignments X and Y, it is defined as



(Rand Index) (Pascal Score)

Figure 3: Performance across the 8 categories for our approach and the baselines. Best viewed in color.

$$RandIndex(X,Y) = \frac{\text{\# Agreements between X,Y}}{\text{All possible pairs}}$$

The Pascal score, on the other hand, measures the number of correctly labelled segments. Note that the Rand index does not require a particular assignment permutation, while the Pascal score depends on the permutation. The problem of computing the optimal Pascal score (with the optimal permutation) can be formulated as a bipartite matching problem, as explained in the previous section (see Table 1).

We compare our supervised model (DHPY) with the unsupervised HPY model, as well as Normalized Cuts (Ncut) and the thresholded Oriented Watershed Transform - Ultrametric Contour Map (OWT-UCM) [2]. The unsupervised HPY model is initialized with the same parameters as the DHPY. The Ncut baseline is performed on the covariance matrix used in our thresholded Gaussian processes, which makes use of the discriminative gPb detector as well as an Euclidean smoothness metric. For each image, we compute Ncut assuming the number of clusters is the number of segments in the ground truth annotation. The OWT-UCM baseline is built on top of the state-of-the-art gPb contour detector; given a threshold value, closed regions can be obtained by computing the connected components in the image. Since we do not know the optimal threshold a priori, we run three OWT-UCM baselines, with the threshold being 0.25, 0.50 and 0.75.

The results across the 8 categories are shown in Fig. 3, in terms of the Rand index and the Pascal score respectively. Our supervised approach improves the performance of the unsupervised model while reducing the variance. Across all categories, our model achieves a Rand index of 0.7848 ± 0.0696 and a Pascal score of 0.8125 ± 0.0686. In terms of the Rand Index averaged over all categories, our method achieves an improvement of 0.1136 over the closest competitor with a p-value of 0.0223, which is statistically significant at the 5% level.

We also evaluate the Rand index performance as a function of the number of training examples. Fig. 2(c) depicts the Rand index averaged among all categories, while Fig. 4 depicts the index for each individual category. The OWT-UCM baseline is chosen to be the best one out of the three thresholded versions. Our model converges to the asymptotic performance more quickly, while incurring a smaller variance with fewer training data. Finally, we show segmentation outputs from the different models in Fig. 5.

6. Conclusion

We have proposed a novel supervised Discriminative Hierarchical Pitman-Yor (DHPY) model, and demonstrated its effectiveness on a natural scene dataset with manual human annotations. Our approach outperforms several baselines, notably the unsupervised HPY model and the OWT-UCM algorithm. Moreover, we have formulated a new constrained optimization framework for non-parametric Bayesian models which can directly and discriminatively train the likelihood distribution, while incorporating the power law priors that are appropriate for generating the visual statistics in natural scenes. In the future we plan to investigate other ways of introducing annotations. [1]

References

[1] N. Apostoloff and A. Fitzgibbon. Bayesian video matting using learnt image priors. In *CVPR*, 2005. 2281

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *CVPR*, 2009. 2281, 2282, 2285, 2286

[3] D. Blei and J. Mcauliffe. Supervised topic models. In *NIPS*, 2007. 2282

[4] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26, 2004. 2281, 2282

[5] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010. 2282

[6] I. Endres and D. Hoiem. Category independent object proposals. In *ECCV*, 2010. 2282

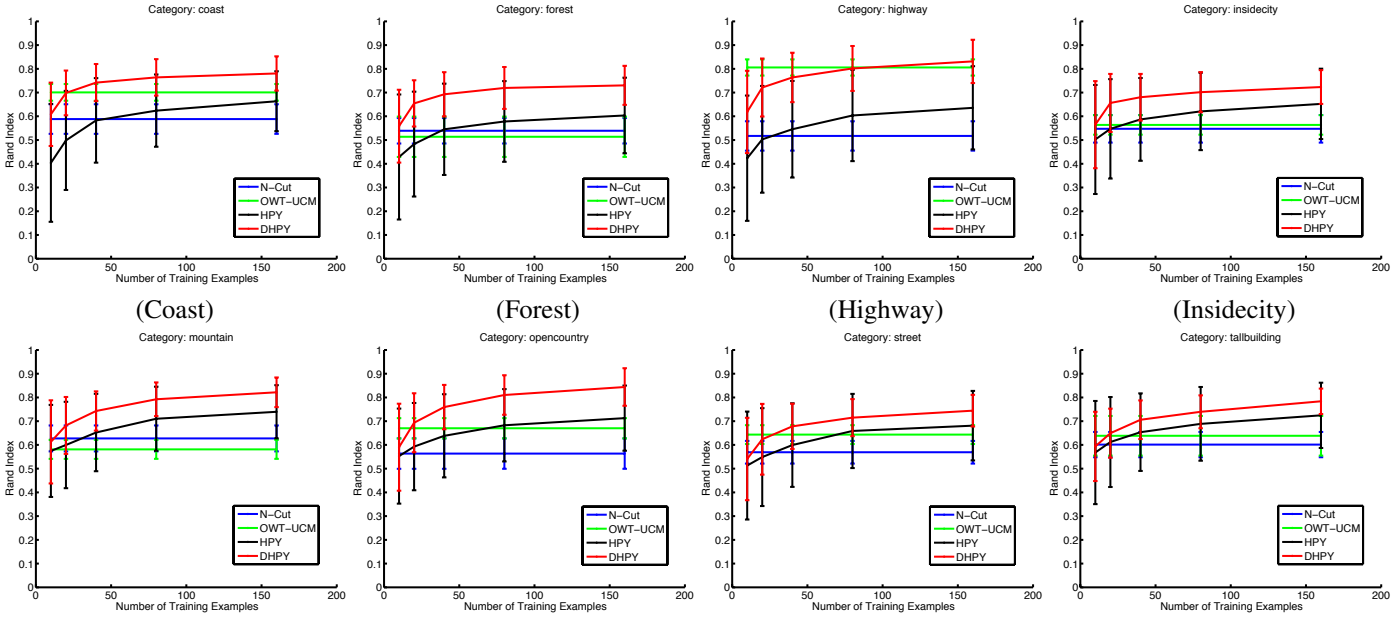
[7] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88, 2010. 2281

[8] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59, 2004. 2285

[9] P. Felzenszwalb and D. Huttenlocher. Efficient belief propagatino for early vision. *IJCV*, 41, 2006. 2281, 2282

[10] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77, 2009. 2284

[1] Various co-authors of this work have been supported in part by awards from the US DOD and DARPA, including contract W911NF-10-2-0059, by NSF awards IIS-0905647 and IIS-0819984, and by Toyota and Google.



(Coast) (Forest) (Highway) (Insidecity)

(Mountain) (Opencountry) (Street) (TallBuilding)

Figure 4: Rand Index as a function of the number of training examples for the different categories.

[11] P. Kohli, M. Pawan, K. Philip, and H. S. Torr. p^3 & beyond: Solving energies with higher order cliques. In *CVPR*, 2007. 2281, 2282

[12] H. W. Kuhn. The hungarian method for the assignment problem. In *50 Years of Integer Programming 1958-2008*, pages 29–47. 2010. 2284

[13] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Obj cut. In *CVPR*, 2005. 2281, 2282

[14] S. Lacoste-julien, F. Sha, and M. I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2008. 2282

[15] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010. 2281, 2282

[16] A. Levinstein, A. Stere, K. Kutulakos, D. Fleet, S. Dickinson, and K. Siddiqi. Turbopixels: Fast superpixels using geometric flows. *PAMI*, 31, 2009. 2285

[17] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik. Using contours to detect and localize junctions in natural images. In *CVPR*, 2008. 2281

[18] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *PAMI*, 26, 2004. 2281

[19] A. Moore, S. Prince, J. Warrell, U. Mohammed, and G. Jones. Superpixel lattices. In *CVPR*, 2008. 2285

[20] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR*, 2004. 2285

[21] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42, 2001. 2285

[22] J. Pitman and M. Yor. The two-parameter poissondirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25, 1997. 2283

[23] N. Ratliff, A. Bagnell, and M. Zinkevich. Subgradient methods for maximum margin structured learning. In *Workshop on Learning in Structured Outputs Spaces at ICML*, 2006. 2285

[24] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, 2003. 2285

[25] S. Roth and M. J. Black. On the spatial statistics of optical flow. In *ICCV*, 2005. 2281

[26] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77, 2008. 2282, 2285

[27] U. Schmidt, Q. Gao, and S. Roth. A generative perspective on mrfs in low-level vision. In *CVPR*, 2010. 2282

[28] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22, 2000. 2282

[29] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006. 2282

[30] E. Sudderth and M. Jordan. Shared segmentation of natural scenes using dependent pitman-yor processes. In *NIPS*, 2008. 2281, 2282, 2283, 2285

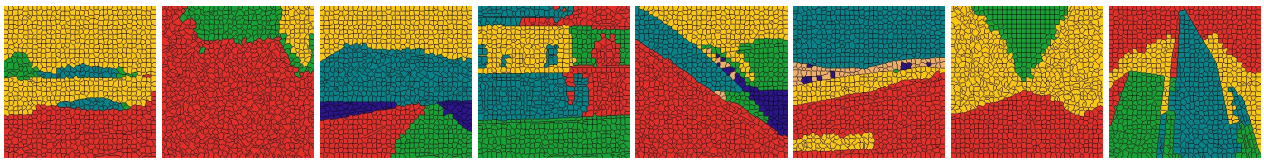
[31] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2004. 2282

[32] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward objective evaluation of image segmentation algorithms. *PAMI*, 29, 2007. 2285

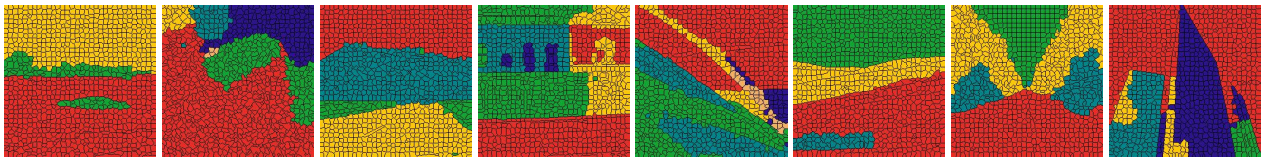
[33] Y. Weiss. Deriving intrinsic images from image sequences. In *ICCV*, 2001. 2281

(a) Raw Image

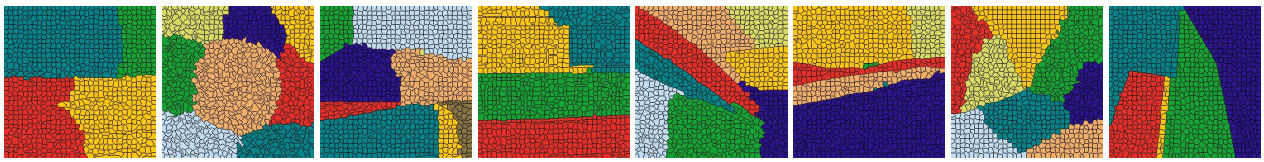
(b) Ground Truth



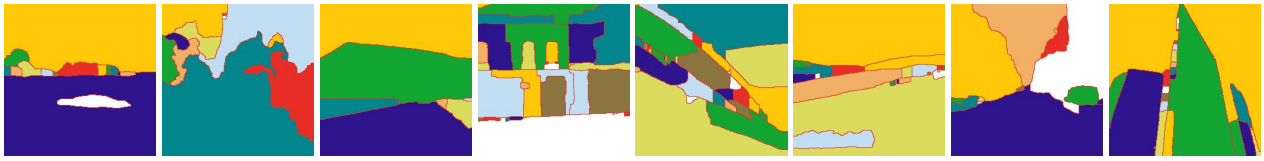
(c) HPY model



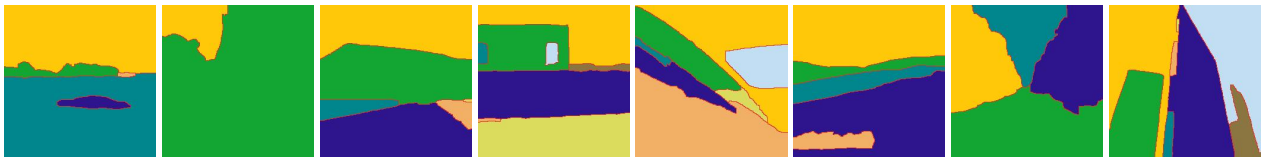
(d) DHPY model



(e) N-cut



(f) OWT-UCM(0.25)



(g) OWT-UCM(0.50)

(h) OWT-UCM(0.75)

Figure 5: Segmentation results for categories *coast*, *forest*, *highway*, *insidecity*, *mountain*, *opencountry*, *street*, *tallbuilding*. The different colors represent different segments. The superpixel boundaries are also displayed for rows (b)-(e). Best viewed in color.