Image Understanding in a Nonparametric Bayesian framework

Soumya Ghosh
sghosh@cs.brown.edu

Abstract

This project proposes new nonparametric Bayesian models and develops corresponding inference schemes for parsing natural images. The primary focus of the proposed work is to develop models for unsupervised image segmentation, using spatially dependent Pitman-Yor Processes. The project will also explore developing Expectation Propagation (EP) based inference schemes for the proposed models. Finally, various mechanisms for incorporating object-category level supervision will be explored.

1 Introduction

Image Understanding, or interpreting images by locating and characterizing their content, is arguably the holy grail of Computer Vision. A general Image Understanding system must flexibly deal with "stuff" (material) and "things" (objects) [1]. Forsyth et al. [7] define stuff as "a homogeneous or repetitive pattern of fine-scale properties, but no specific or distinctive spatial extent or shape" while a thing is defined as "an object with specific shape and size". For instance, trees, sky and gravel are good examples of stuff, while cars,tigers and boats are examples of objects (Figure 1). Traditionally, statistical models have dealt with either stuff (under the umbrella of image segmenta-

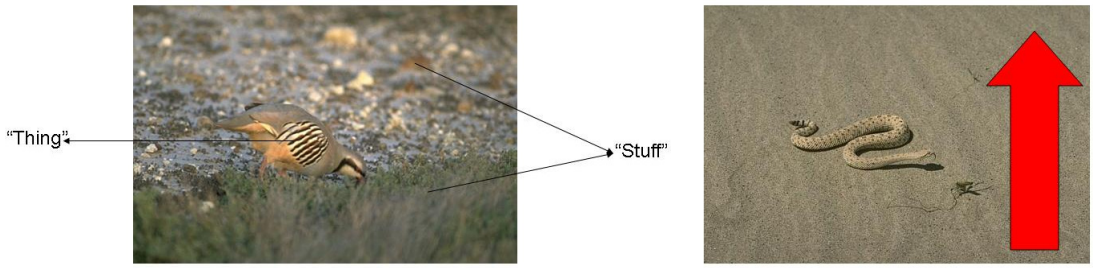

Table 1: Left: Distinction between Stuff and Things. Right: Smooth gradient in the direction of the arrow.

tion [4],[6],[12]) or things(object detectors) [15] but rarely both. More recently, some progress has been made in leveraging one model to better learn the other. Typically, object models for a fixed number of object categories are specified and learnt from training data. These are then used to detect potential objects in an image. These predictions are then combined in a coherent fashion using the "stuff" models. For instance, Heitz et al. [8] use "stuff" based clusters (segments) to prune away false positives from the predictions of a sliding window based car ("thing") detector.

However, such models are difficult to generalize. They scale poorly with the number of object categories. Adding new object categories, involves a significant overhead of specifying and training new object models. To deal with these issues, various nonparametric approaches have been suggested. These approaches come in two popular flavors: nearest neighbor and nonparametric Bayesian. The

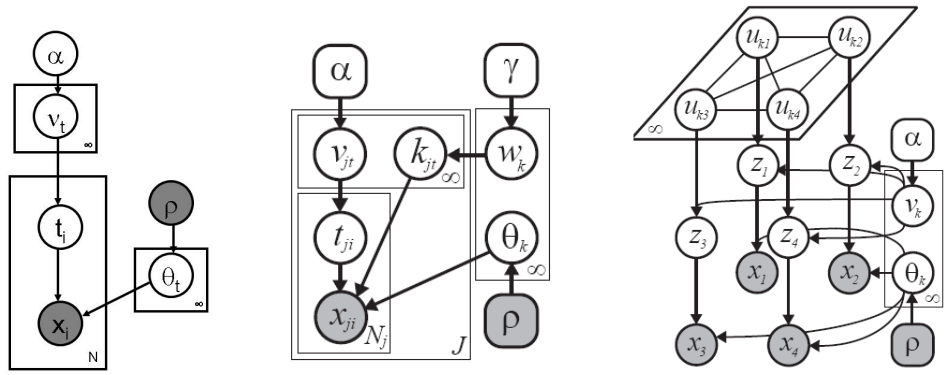

Table 2: Left: A Pitman-Yor mixture model for independent image segmentation. Center:HPY model for shared segmentation of J images. For image j draw segment proportions $\boldsymbol{\pi}_j \sim GEM(\alpha)$, equivalently $\upsilon_{jt} \sim Beta(1-\gamma_a, \gamma_b + k\gamma_a)$ and $\pi_{jt} = \upsilon_{jt}\prod_{l=1}^{t-1}(1-\upsilon_{jl})$. For segment t draw a global object category $k_{jt} \sim \boldsymbol{\psi}$, where $\boldsymbol{\psi} \sim GEM(\gamma)$. For each super-pixel i in image j draw a region $t_{ji} \sim Mult(\boldsymbol{\pi}_j, 1)$. Finally, generate the observed features $x_{ji} \sim Mult(\theta_{k_{jt_{ji}}})$. Right: Spatially Dependent Pitman-Yor mixture. For every segment k, $\boldsymbol{u}_k \sim GP(\mathbf{0}, \boldsymbol{K})$ and for a super-pixel i, $z_i = min\{k|u_{ki} < \Psi^{-1}(\nu_k)\}$ where $\Psi(u_{ki})$ is the Normal CDF

nearest neighbor based approaches involve searching over a gigantic database of images, finding a set of closest matches and transferring image content information from this set to a given test image [9]. The intuition here is that only those object categories which appear amongst an image's closest neighbors are likely to appear in the image. These models are limited by the availability and the ability to search over millions of images.

An alternate strategy exploits the Nonparametric Bayesian (NPB) machinery and places nonparametric priors over the distribution of object-categories and segments. Various NPB models have been proposed to explain natural images [2], [5],[13]. Our proposed research, builds on previous work by Sudderth and Jordan [13], where nonparametric models have been used for unsupervised image segmentation. The authors model two well known, but difficult to model effects in natural images, a) Segment sizes and Object frequencies follow a power law distribution and b) image patches (super-pixels) are not exchangeable but spatially dependent.

By placing a Pitman-Yor process prior on segment assignments, the authors induce heavy tailed distributions over segment sizes. The resulting model is a simple Pitman-Yor mixture model (Figure 2 *(left)*) and corresponds to a "stuff" model. It can be generalized to share information amongst images and simultaneously segment a collection of J images using the hierarchical Pitman Yor (HPY) process model. The generalized model shares information through objects (which themselves are drawn from another PY prior) which share segments (Figure 2 *(center)*). Such a generalization corresponds to a simple synergistic combination of the "stuff" and "thing" models.

Spatial dependence amongst image patches is induced by placing a Gaussian process(GP) prior over the latent variables generating the patch features (Figure 2 *(right)*). The covariance kernel of the Gaussian process models the dependence between the latent variables. In particular, a GP is associated with each of the infinite possible segments (t). Thus, every super-pixel i has an infinite collection of Gaussian random variables $\kappa = \{u_{ti}\}_{t=1\ldots\infty}$. The PY process prior then determines an appropriate threshold to chop κ at. Super-pixel i is then assigned to the first segment t to cross the threshold.

2 Proposed Research

Our proposed research plans to extend [13] in a number of ways. First, we propose more flexible models for image segmentation and cleverer inference schemes for those models. Next, we propose a "thing" model for leveraging available object-category labels and finally we plan to investigate ways

of synergistic combinations of the "stuff" and "thing" models. We now describe these extensions in greater detail:

2.1 Intelligent Modeling of "stuff"

The primary goal of this project is image segmentation. Specifically, we are planning two major extensions to the segmentation model proposed in [13].

Learning Segmentation Resolution. In the spatially dependent model, the segmentation resolution is governed by two factors, the number of segments assigned to the image by the PY prior and the covariance kernel of the Gaussian process prior. Thus, by learning the PY hyper-peramters and the GP covariance kernel we can learn the desired segmentation resolution.

The hyper-parameters will be learnt through ML estimation over a training set of user provided image segmentations. We plan to learn the covariance kernel discriminatively from the training set. To learn the covariance between two super-pixels we propose using discriminative cues, which highlight the difference between them. This information is complimentary to the generative features used to model the super-pixels (For instance, in [13] each super-pixel is modeled by it's color and texture histogram). Preliminarily, we plan to experiment with two features: the probability of an edge and the distance between two super-pixels. These features will be pumped through a logistic regression classifier, the output of which will be calibrated to give valid correlation values (between +1 and -1). It is worth noting that the proposed model is able to capture both positive and negative correlations between super-pixels. Over the course of the project, we will concentrate on both expanding the set of discriminative features and on the details of logistic regression (number of basis functions, bandwidth etc.).

Flexible Segments. Feature clustering based segmentation techniques like [13] are poorly suited for dealing with smooth texture and color gradients. Such gradients, common in natural images due to perspective and/or shading effects (Figure 1*(right)*), cause patches (super-pixels) to appear dissimilar despite belonging to the same segment and result in over-segmentation. We propose to deal with such smooth gradients by modeling each segment with a range of smoothly varying color and texture histograms, instead of just one. This added flexibility allows super-pixels with different appearances to belong to the same segment. Intuitively, we achieve this by specifying the mean parameters of the multinomial describing a segment and then varying them smoothly in some low dimensional space, while ensuring that the latent variables generating neighboring super-pixels remain correlated. Formally, the following generative story holds for every segment (Figure 3*(left)*):

1. For every segment t, we have $\mathbf{\Phi}_t \in \mathcal{R}^{W \times D}$ and $\boldsymbol{\theta}_t \in \mathcal{R}^{W \times 1}$, where W is the dimensionality of the histograms used to describe the segment and D is the dimensionality of the low dimensional space responsible for capturing the within segment smooth variations.

2. For $d = 1..D; \boldsymbol{u}_{td} \sim GP(\mathbf{0}, K_d)$ with $\boldsymbol{u}_{td} \in \mathcal{R}^{N \times 1}$ where N is the number of super-pixels in the image.

3. For every super-pixel i, $\boldsymbol{z}_{ti} = \mathbf{\Phi}_t[u_{t1_i}...u_{tD_i}]^T + \boldsymbol{\theta}_t$

4. $X_i \sim Mult(\sigma(\boldsymbol{z}_{ti}))$ where σ is the softmax function.

Evaluation We will evaluate the proposed models by comparing them to models proposed in [13] and more traditional image segmentation algorithms, Normalized cuts [12], Mean Shift [4] and Graph Based image segmentation [6]. They will be compared on the basis of their performance on the Berkeley Image Segmentation Dataset (BSDS) [10] dataset. The BSDS dataset has emerged as a standard benchmark [3] for evaluating unsupervised image segmentation results. It contains a set of 300 natural images with 200 training and 100 test images. All 300 images have a set of manual ("ground-truth") segmentations available. The machine generated segmentations will be quantitatively evaluated using two popular metrics Probabilistic Rand Index [14] and segmentation cover [3].

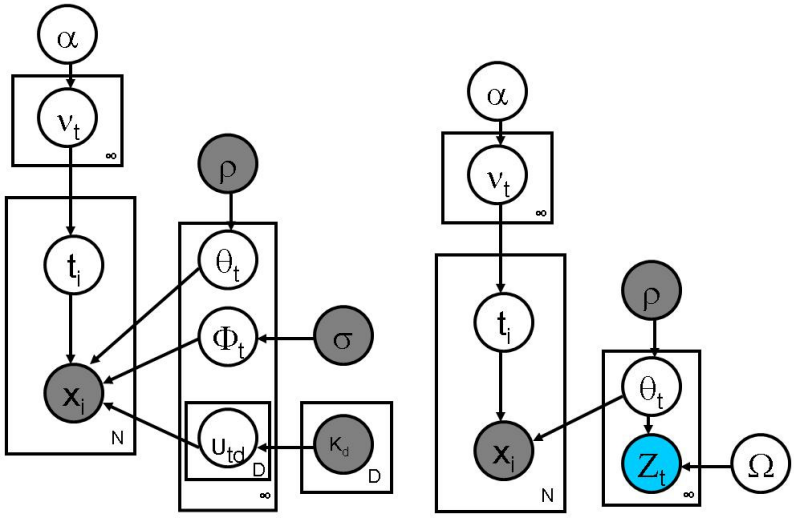

Table 3: We only present the non spatially dependent single image models here. These models can be easily extended to both the spatially dependent and shared versions. Left: Flexible Segments Model. Right: A simple model incorporating object-category level supervision. The gray nodes are visible, the blue nodes are visible only during training. A categorical response variable $Z_t \in 1...C$ is now attached with each segment, with $P(Z_t = c|t, \theta_t) = \frac{exp(\boldsymbol{\omega}_c{}^T \theta_t)}{\sum_j exp(\boldsymbol{\omega}_j{}^T \theta_t)}$ and $\Omega = [\omega_1 ... \omega_C]$

2.2 Modeling "things"

A medium term goal of the project is to incorporate object-category level supervision to model "things". In the literature there exist models which augment unsupervised latent variable models with response variables. These response variables are observed during training and need to be predicted during testing. However, typically such models incorporate document (image) level supervision. In our context, it makes more sense to incorporate finer grained segment level supervision. One concrete way of incorporating such supervision is presented in Figure 3 *(right)*. There probably exist other interesting ways of incorporating such supervision and will be further explored. Also, "things" tend to be made up of specific "parts" arranged in a particular configuration (for instance, cars are made up of wheels,lights,windows etc.). A "thing" model will probably benefit from explicitly modeling the constituent parts of things. Coming up with such models is a long term goal of the project.

Finally, we note that most existing models do not combine "stuff" and "thing" models in any interesting way. Usually, the two models are learnt independently(or sequentially at best) thus throwing away useful information about each other. A long term goal of this project is to investigate more sophisticated mechanisms of combining these models in a synergistic fashion.

2.3 Inference Scheme

In [13], Sudderth and Jordan develop Mean Field variational inference for their shared image segmentation models. However, Mean Field methods are fraught with local optima problems. We anticipate such local optima problems to be more acute for the more complicated models proposed in this project. This necessitates exploring other techniques for performing inference. In particular, one such alternate inference scheme called Expectation Propagation(EP)[11] is known to work well for Gaussian process classification. Our spatially dependent models are similar and we will start by deriving and implementing EP for the models proposed in [13] as well as their extensions proposed here.

References

[1] E. H. Adelson. On seeing stuff: the perception of materials by humans and machines. In B. E. Rogowitz & T. N. Pappas, editor, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 4299 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 1–12, June 2001.

[2] Q. An, C. Wang, I. Shterev, E. Wang, L. Carin, and D. B. Dunson. Hierarchical kernel stick-breaking process for multi-task image analysis. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 17–24, New York, NY, USA, 2008. ACM.

[3] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:2294–2301, 2009.

[4] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[5] L. Du, L. Ren, D. Dunson, and L. Carin. A bayesian model for simultaneous image clustering, annotation and object segmentation. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 486–494. 2009.

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2):167–181, 2004.

[7] D. Forsyth, J. Malik, M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler. Finding pictures of objects in large collections of images. Technical report, Berkeley, CA, USA, 1996.

[8] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 30–43, Berlin, Heidelberg, 2008. Springer-Verlag.

[9] C. Liu, J. Yuen, and A. B. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, pages 1972–1979, 2009.

[10] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

[11] T. P. Minka. Expectation propagation for approximate bayesian inference. In *UAI*, pages 362–369, 2001.

[12] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE TRANS. ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 22(8), 2000.

[13] E. B. Sudderth and M. I. Jordan. Shared segmentation of natural scenes using dependent pitman-yor processes. In *NIPS*, pages 1585–1592, 2008.

[14] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward objective evaluation of image segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):929–944, June 2007.

[15] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.