A Time and Space Dependent Topic Model for Unsupervised Activity Perception in Video

Eric Wang
Duke University
Durham, NC
ew28@duke.edu

Lawrence Carin
Duke University
Durham, NC
lcarin@ee.duke.edu

Abstract

We present a novel unsupervised topic model applicable to video, performing activity perception with spatial and temporal dependence. The proposed model employs a hierarchical Bayesian framework in which a set of topics are globally shared over a set of video segments, or tasks, and the topic probabilities evolve with time. Each task contains quantized motion features ("words") and corresponding spatial information extracted from a short segment of video. We characterize each task as a mixture of topics drawn from a spatially dependent mixture model. Variational inference is performed, and we show sample results on a complex fixed-camera traffic scene.

1 Introduction

We consider nonparametric statistical analysis of a contiguous series of video segments, each containing movement features with corresponding spatial information. In describing observed actions in a video sequence, we follow the recent work of [14] and break a video sequence into short segments (tasks or "documents"), extract quantized local motion features ("visual worlds") within each task, and describe the tasks using a Bayesian hierarchical topic model. However, unlike [14] we do not embed spatial information in the features. Advantages of our approach include maintaining only a small dictionary and robustness to camera shifts.

Bayesian hierarchical mixture models commonly adopt the Dirichlet process (DP) [7] as a convenient prior on infinite mixture models. Dirichlet process mixture models, like many others (*e.g.* [3, 8]), assume data exchangeability, or a "bag-of-words". Recently, there has been interest in non-exchangeable mixture models, imposing either temporal [2, 10] or spatial dependence [4, 6]. Since both temporal and spatial relationships exist in video applications, it is reasonable to build these dependencies into our model.

The logistic stick breaking process (LSBP) [5, 11] was proposed as an image segmentation model that encourages spatially contiguous clusters. Following [13], [11] also proposed the hierarchical LSBP (H-LSBP), extending LSBP to a multi-task setting, although tasks are still assumed exchangeable. To address the latter issue, our proposed model extends the H-LSBP by adopting the time-dependent framework from the dynamic Dirichlet topic model (d-DTM) [10].

In the next section, we present the proposed dynamic hierarchical LSBP (dH-LSBP). Section 3 discusses example results on a fixed-camera traffic scene. Finally, we present conclusions and discuss future work in Section 4.

2 The Dynamic Hierarchical Logistic Stick Breaking Process

Assume a set of sequentially time-stamped tasks $\{D_m\}_{m=1:M}$ where $D_m = \{x_{mi}, \boldsymbol{s}_{mi}\}_{i=1:N_m}$. In our case, the x_{mi} are observations that take integer values from 1 to 8 (see Fig. 1a), denoting a particular quantized direction, and the $\boldsymbol{s}_{mi}$ are two dimensional vectors indicating the pixel location of x_{mi}. Although multiple tasks can have the same time stamp value, to simplify the following discussion, we assume each task has a unique time stamp.

2.1 Single task spatial clustering

We first briefly discuss the spatially dependent segmentation of a single task, D_m, using LSBP. Assume an infinite set of "layers" with corresponding model parameters $\{\boldsymbol{\phi}^*_{mk}\}_{k=1:\infty}$ drawn iid from a base distribution G_m. A vector of binary gating variables $\boldsymbol{Z}_{mi} = \{z_{mi1}, ..., z_{mi\infty}\}$ is associated with each x_{mi} such that x_{mi} is assigned to layer k if $k = \min\{k : z_{mik} = 1\}$.

Each gating variable z_{mik} is drawn from $Bernoulli(\sigma(g_{mk}(\boldsymbol{s}_{mi})))$ where $\sigma(.)$ is the logistic function and $g_{mk}(\boldsymbol{s}_{mi})$ denotes a kernel-based linear regression for layer k with a sparseness promoting Student-t prior [1] on the layer weights. The kernel $K(\boldsymbol{s}_{mi}, \hat{\boldsymbol{s}}_p, \psi_{mk}) = \exp(-\psi_{mk}||\boldsymbol{s}_{mi} - \hat{\boldsymbol{s}}_p||_2^2)$ is the radial basis function where $\{\hat{\boldsymbol{s}}_p\}_{p=1:N_b}$ are fixed basis locations and ψ_{mk} is the kernel width.

Sparse layer weights encourage spatially proximate data to occupy the same layer (and share the same set of model parameters). It was shown in [11] that the probability of x_{mi} being associated with layer k follows a stick breaking construction analogous to [12]. To ease inference, we truncate the LSBP to a finite (but sufficiently large) number of topics K.

2.2 Time dependent multi-task spatial clustering

In [11], simultaneous spatially dependent segmentation of multiple images was considered via the H-LSBP, where a top-level $DP(\alpha, G_0)$ was introduced in the same manner as HDP [13] to provide a discrete base distribution $G = \sum_{l=1}^{\infty} \pi_l \delta_{\phi_l}$ shared across all tasks. As in HDP, tasks in H-LSBP are assumed exchangeable. We remove task-level exchangeability by imposing that the top level mixing weights $\boldsymbol{\pi}$ should evolve with time by adopting the time-dependent construction from [10].

We now make concrete the base distribution G_m for the LSBP of task D_m. Let $G_m = \sum_{l=1}^{L} \tau_{ml} \delta_{\boldsymbol{\phi}_l}$, where L is a finite truncation level and each $\boldsymbol{\phi}_l$ is a globally shared topic, then at time m, the mixture weights $\boldsymbol{\tau}_m$ can be written as $\boldsymbol{\tau}_m = (1 - w_m)\boldsymbol{\tau}_{m-1} + w_m \boldsymbol{\pi}_m$, where $\boldsymbol{\pi}_m$ are innovation weights, $\{\boldsymbol{\tau}_1, \boldsymbol{\pi}_2, ..., \boldsymbol{\pi}_M\} \sim Dir(\alpha)$. The innovation parameter w_m is drawn from $Beta(a_0, b_0)$ and governs the degree to which the innovation weights $\boldsymbol{\pi}_m$ are considered in the construction of $\boldsymbol{\tau}_m$. Note that while the top level topic probabilities evolve with time, the topics are shared explicitly over all tasks and all times.

Based on the above definitions, the dynamic hierarchical LSBP (dH-LSBP) can be expressed as

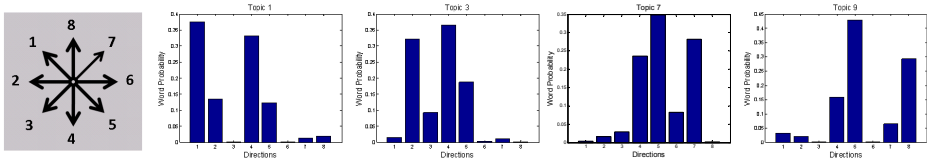
$$
\begin{aligned}
x_{mi} &\sim Mult(\boldsymbol{\phi}^*_{m\min\{k:z_{mik}=1\}}), \;\; i = \{1, ..., N_m\} \\
z_{mik'} &\sim Bernoulli(\sigma(g_{mk'}(s_{mi}))), \;\; z_{miK} = 1, \;\; k' = \{1, ..., K-1\} \\
g_{mk'}(s_{mi}) &= \sum_{p=1}^{N_b} \theta_{mk'p} K(s_{mi}, \hat{s}_p, \psi_{mk'}) + \theta_{mk'0}, \\
\boldsymbol{\theta}_{mk'} &\sim \prod_{p=0}^{N_b} N(\theta_{mk'p}|0, \lambda_{mk'p}) Gamma(\lambda_{mk'p}|c0, d0), \\
\boldsymbol{\phi}^*_{mk} &\sim G_m, \;\; m = \{1, ..., m\}, \;\; k = \{1, ..., K\} \\
G_m &= \sum_{l=1}^{L} \tau_{ml} \delta_{\boldsymbol{\phi}_l}, \;\; \boldsymbol{\tau}_m = (1 - w_m)\boldsymbol{\tau}_{m-1} + w_m \boldsymbol{\pi}_m, \\
\boldsymbol{\pi}_m &\sim Dir(\boldsymbol{\alpha}), \;\; w_m \sim Beta(a_0, b_0), \;\; m = \{1, ..., M\} \\
\boldsymbol{\phi}_l &\sim Dir(\boldsymbol{\eta}), \;\; l = \{1, ..., L\}.
\end{aligned}
$$

In our proposed approach, the set of topics which are available to task m (*i.e.* those topics with high probability in G_m) change smoothly as a function of time. Note that if $w_m \to 0$ for all m,

then the dH-LSBP is equivalent to H-LSBP with the top level truncated stick breaking construction replaced by a finite dimensional Dirichlet distribution. We obtain analytical variational posteriors on all parameters except $\psi_{mk'}$, which we sample. For brevity, we omit the update equations in this paper.

3 Example Results and Discussion

We ran our model on a 4000 frame (approx. 4 min) fixed camera video of a busy intersection, available from `http://www.ngsim.fhwa.dot.gov/`. Optical flow [9] is used to extract local motion features on moving pixels, defined as pixels whose normalized intensities changed by more than 0.5 between successive frames. To reduce data load, each 640×480 frame is divided into non-overlapping 5×5 patches. If a patch contains more than 10 moving pixels and the average magnitude of motion across all moving pixels in the patch is greater than 0.5 , then the average motion of the patch is quantized into one of 8 possible words corresponding to directions (see Fig. 1a). The location of each patch is taken at its center. To form tasks, we collect the quantized directions and



(a) (b) (c) (d) (e)

Figure 1: (a) shows the word indices corresponding to the 8 directions. We also example inferred topics that are pertinent to turning situations (*i.e.* topics with high probability of words 1, 3, 5, and 7).

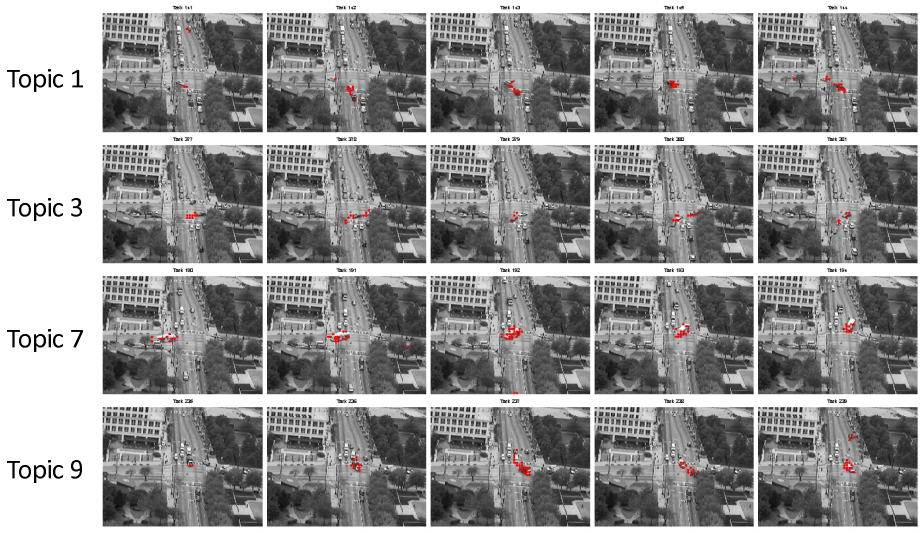


Figure 2: Sample segmentation results for the four topics from above. For each topic, we show representative frames from 5 consecutive tasks. The red dots denote centers of patches which have high posterior probability to the corresponding topic (at left) in the task. These are best viewed by zooming in on the electronic version of the paper.

their locations from every 10 frames, without overlap. Finally, we assigned the same time stamp value to every 10 consecutive tasks, without overlap. By collecting data in this manner, we ensure a rich set of features and locations across all tasks and all times. The processed dataset consists of $M = 400$ tasks, with 40 unique time stamps. The truncation levels are set to $L = 10$ and $K = 5$, and the LSBP sparseness hyperparameters are set as $c_0 = d_0 = 10^{-6}$.

In Fig. 1b-e, we show example topics learned by the dH-LSBP. In the interest of space, we focus on topics which relate to more-interesting turning situations. In Fig. 2, we show some typical segmentation results for our scene. The red dots correspond to the centers of patches with high posterior probability to the listed topic. We note generally good topic spatial contiguity in turning situations, a result which is not entirely intuitive from Fig. 1b-e because the topics do not focus on a single word, but tend to have high probability for several correlated words. For example, topic 1 (Fig. 1b) relates to cars making two types of turns: 1) traveling diagonally up and left, then proceeding left out of the frame; and 2) traveling down from the top of the frame and then diagonally down and right. This is reasonable since these two types of turns tend to happen at the same time.

Finally, by decoupling the words from their spatial locations, our topics are robust to camera shifts; additionally, it allows us to train on a particular video and perform retrieval on a similar but different video. For these and other extended results, we refer the interested reader to our website, available at `http://ericxwang.moonfruit.com/research`.

4 Conclusion and Future Work

We propose a novel spatially and temporally dependent topic model called dH-LSBP for nonparametrically learning activities in video. We have shown example results from a complex intersection video sequence. Future work includes more in-depth study and comparison of dH-SLBP against other activity models and an improved initialization scheme using factor models to build a library of actions. Finally, another application of our model is on time-evolving documents as considered in [10], since dH-LSBP can impose dependencies both across and within documents.

References

[1] C. Bishop and M. Tipping. Variational relevance vector machines. In *Proceedings UAI 2000*, pages 46–53, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the ICML 2006*, pages 113–120, New York, NY, USA, 2006. ACM.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[4] Y. Chung and D. B. Dunson. Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 2009.

[5] L. Du, L. Ren, D. B. Dunson, and L. Carin. A Bayesian model for simultaneous image clustering, annotation and object segmentation. In *Proceedings of NIPS 2009*, 2009.

[6] D. B. Dunson and J. Park. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.

[7] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[8] L. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Proceedings of CVPR 2009*, 2009.

[9] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision, 1981.

[10] I. Pruteanu-Malinici, L. Ren, J. Paisley, E. Wang, and L. Carin. Hierarchical Bayesian modeling of topics in time-stamped documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), 5555.

[11] L. Ren, D. B. Dunson, and L. Carin. Logistic stick breaking process. *submitted to J. Machine Learning Research*, 2009.

[12] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[13] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101, 2003.

[14] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(3):539–555, 2009.