DYNAMIC BAYESIAN ACTIVITY MODELING IN VIDEO VIA MULTI-FEATURE INTEGRATION

T. Scott Brandes

Signal Innovations Group, Inc.
4721 Emperor Blvd., ste 330,
Durham, NC 27703

Eric Wang *

United States Naval Academy
Department of Mathematics
572C Holloway Road, Chauvenet Hall
Annapolis, MD 21402-5002

ABSTRACT

We present a Bayesian approach to unsupervised clustering of activity within video imagery. Vehicles and pedestrians are tracked within the video imagery and their collective activity in each time frame is measured and categorized using a natural extension of the dynamic latent Dirichlet allocation model. Our extension involves use of multiple types of simultaneously observed features from multiple classes of objects within the video imagery. Within the prior for the model these features are treated as independent, and modeled as draws from a variety of appropriate distribution types. By including multiple features, the model generates a richer set of activities; we quantitatively show that this yields better predictions of physical attributes within the scene, relative to currently available models that use only the single best feature. We show this by comparing model prediction of traffic light states within a busy intersection, which we ground-truth manually within the video imagery.

***Index Terms*—** hierarchical models, Dirichlet process, activity modeling, dynamic latent Dirichlet allocation

1. INTRODUCTION

Much research effort is currently focused on the discovery and categorization of human activity within video data. The large body of existing work suggests both the difficulties and interest in the automated analysis of visual surveillance of objects and behavior (for a thorough survey see [1]). Of particular interest is the ability to automate the understanding of activity within complex scenes. There are predominantly two categories of approaches to this problem. The more traditional approach involves tracking objects through the scene and assessing activity based on these tracks [2], [3], [4], [5], [6]. Much previous research using a tracking approach relies heavily on expert knowledge to label a fixed set of activities. This supervised approach has limited robustness and is unable to expand beyond the fixed set of training actions. The other approach involves measuring motion within the video images directly to asses activity [7], [8], [9], [10]. In particular [11, 12] propose unsupervised approaches employing topic models operating on discrete spatio-temporal actions to learn meaningful co-occurrences of actions.

With the advent of more sophisticated detection and tracking algorithms that incorporate a probabilistic motion, shape, and color model for detection and tracking [13], it is possible to leverage measurements within the video images directly to create a more robust Bayesian tracking algorithm. Moreover, combining these robust tracks with hierarchical Bayesian models allows for a more sophisticated approach to activity modeling. While previous approaches that use hierarchical Bayesian models [11, 12] are powerful, they operate entirely on a single feature, extracted from quantized optical flow [14]. Further, they require the number of topics be set *a priori*, and rely on Gibbs sampling for inference. We propose a nonparametric model that can simultaneously consider multiple features, treats different classes of objects ($e.g.$, pedestrians vs. vehicles) in the scene independently (rather than using a common feature set across all classes), automatically infers the number of topics via a stick-breaking[15] representation of the Dirichlet Process [16], and employs efficient variational Bayesian (VB) inference [17]. Moreover, our proposed model also incorporates temporal dependence via the dynamic structure presented in [18].

2. REVIEW OF DYNAMIC TOPIC MODELING

Topic models [19], [20], [21], [22] refer to a family of models that attempt to learn meaningful co-occurrences of discrete tokens (words) from a corpus of documents. Currently, the most widely used topic model is *Latent Dirichlet Allocation* (LDA) mainly due to it being a fully generative model of documents while [19], [20] are not. In LDA, each document is characterized by a discrete distribution over a set of globally shared topics, where each topic is a discrete distribution over a fixed set of words. In LDA, observations are considered exchangeable and the inferred topic distributions are shared

*Author performed the work while at Duke University

across all documents. With observations that have a sequential time attribute, the dynamic nature of how topic mixtures evolve over time can be exploited by removing this global exchangeability assumption and constraining exchangeability to a varying time window [23]. This results in a dynamic latent Dirichlet allocation (dLDA) model [18] that generates a time varying mixture of topics. Pruteanu-Malinici *et al.* [18] considered a time-stamped sequence of documents, imposing that the topic proportions within a document at time t are constructed dependent on the topic proportions within documents of time $t-1$.

If we consider a collection of documents with known time stamps $t = 1,\ldots,T$, where the total number of independent documents at any given time is N_t, we can describe the full set of documents over time as $\{\boldsymbol{x}_{t,i}\}_{t=1,i=1}^{T,N_t}$ where $\boldsymbol{x}_{t,i}$ represents a vector of word counts in document i at time t. The dLDA generative model describes a process for a time-evolving mixture of topics and is written as

$$\begin{aligned}
\boldsymbol{x}_{t,i} &\sim F(\boldsymbol{\varphi}_{z_{t,i}}) \\
z_{t,i} &\sim Multi(\tau_t^1, ..., \tau_t^K), z_{t,i} \in \{1, K\} \\
z_{t=1,i} &\sim Multi(\boldsymbol{\pi}_1) \\
\boldsymbol{\tau}_t &= (1 - w_t)\boldsymbol{\tau}_{t-1} + (w_t)\boldsymbol{\pi}_t \\
\boldsymbol{\pi}_t &\sim Dir(\frac{\alpha}{K}, ..., \frac{\alpha}{K}) \\
w_t &\overset{i.i.d.}{\sim} Beta(c_0, d_0) \\
\boldsymbol{\varphi}_k &\overset{i.i.d.}{\sim} H
\end{aligned} \tag{1}$$

Here, each word is drawn from a distribution $F(\varphi)$ parameterized by a topic φ, and the topic assignment $z_{t,i}$ for document i at time t is drawn from a multinomial distribution. In standard LDA, this multinomial is parameterized by a draw $\boldsymbol{\pi}$ from a symmetric Dirichlet distribution, providing a constant set of mixing weights. In dLDA, a time-evolving aspect is introduced by parameterizing the draw of $z_{t,i}$ by τ_t. This parameter is a sum of the mixing weights in the previous time step and current time step, proportioned by an innovation weight w_t. When w_t is small, the mixing weights from the previous time step have more influence in the model than the newly drawn mixing weights. Whereas, when w_t is large, the new mixing weights have more influence and the model is more likely to transition to a new set of topic mixtures. The kth topic assigned to the jth word is drawn *i.i.d.* from a distribution H^j. By selecting the distributions F and H as conjugate pairs, efficient variational inference is possible. Observations in the form of words are well modeled as a draw from a multinomial distribution, $F(\boldsymbol{\varphi}^j_{z_{t,i}}) = Multi(\boldsymbol{\varphi}^j_{z_{t,i}})$, and conjugacy dictates that the topics are drawn from a Dirichlet distribution, $H^j = Dir(\frac{\beta}{J}, ..., \frac{\beta}{J})$. To ensure that a sufficient number of topics are available within the model, the total number of possible topics K is set large enough that not all will have an observation associated with them. Any unpopulated clusters are pruned away after parameter inference.

3. FEATURES OBSERVED

The type of distribution used to model observables and topics within the dynamic topic modeling framework is flexible, and should consist of whatever is most appropriate for the observation itself. When considering a dynamic topic model framework applied to systems that do not involve text, such as video data of a busy intersection of roads, a more sophisticated notion of observation needs to be considered. For instance, observations could consist of multiple simultaneous measures, such as the distribution of vehicle speeds and headings, as well as the spatial density of vehicles within an intersection. In such a setting, instead of modeling topics in the traditional sense, we model topics of activity.

Along with multiple features, we also incorporate multiple categories of observed objects. For instance, by using existing tracking software [13] that creates tracks for vehicles in the scene, and a different track set for pedestrians in the scene, we have two categories of tracked objects to consider. Each of these object categories has its own set of features. Vehicle tracks contain a rich set of information to consider. In particular, the features we measure from the collection of vehicle tracks in the scene are vehicle spatial density, heading change, acceleration, heading of decelerating vehicles, and heading of accelerating vehicles. Each measurement category is treated as an observational feature of vehicles in the scene. The features we find useful from pedestrian tracks are their spatial density and distribution of headings. The spatial density of vehicles or pedestrians is modeled as a draw from a product of Poisson distributions, one Poisson per spatial block j and indexed by object category i, $F(\varphi^j_{z_{t,i}}) = Poisson(\varphi^j_{z_{t,i}})$. The parameter of each Poisson is modeled as a draw from a gamma distribution $H^j = Gamma(a, b)$. Each of the other features mentioned are modeled as draws from a J_n-dimensional multinomial, a different length for each mode n, with parameters drawn from a Dirichlet distribution, $F(\boldsymbol{\varphi}^j_{z_{t,i}}) = Multi(\boldsymbol{\varphi}^j_{z_{t,i}})$ and $H^j = Dir(\frac{\beta}{J_n}, ..., \frac{\beta}{J_n})$.

Rather than treating each feature as a single concatenated feature, we propose a more robust model that allows these features to be drawn independently from differing distribution types, allows different sets of features for different categories of observed object, and allows for missing features or object types. This also differs from LDA where the features are dependent on one another since the topics normalize to one. The more flexible model we propose is the multi-feature dLDA.

4. MULTI-FEATURE DLDA MODEL

In the multi-feature dLDA model, simultaneous measures, or multiple observational features, are modeled collectively where each feature is modeled with an independent generative distribution. Since each feature is independent, observations can consist of differing distribution types. The multi-feature

form of dLDA we propose is as follows.

$$
\begin{aligned}
\{\boldsymbol{x}_{n,t,i}\}_{n=1}^{N} &\sim \{F_n(\boldsymbol{\Phi}_{n,z_{t,i}})\}_{n=1}^{N} \\
z_{t,i} &\sim Multi(\tau_t^1, ..., \tau_t^K), z_{t,i} \in \{1, K\} \\
z_{t=1,i} &\sim Multi(\boldsymbol{\pi}_1) \\
\boldsymbol{\tau}_t &= (1 - w_t)\boldsymbol{\tau}_{t-1} + (w_t)\boldsymbol{\pi}_t \\
\boldsymbol{\pi}_t &\sim Dir(\frac{\alpha}{K}, ..., \frac{\alpha}{K}) \\
w_t &\overset{i.i.d.}{\sim} Beta(c_0, d_0) \\
\{\boldsymbol{\Phi}_{n,k}\}_{n=1}^{N} &\overset{i.i.d.}{\sim} \{\boldsymbol{H}_n\}_{n=1}^{N}
\end{aligned}
\tag{2}
$$

Here the observational set of N features $\{\boldsymbol{x}_{n,t,i}\}_{n=1}^{N}$ are collectively generated from the same topic or activity groups $\{\boldsymbol{\Phi}_{n,k}\}_{n=1}^{N}$ that are shared globally. The activities themselves $\boldsymbol{\Phi}_{n,k}$ are multi-featured, matching the observational set. Since this model is not specifically designed for document analysis, the document index i can be thought of as an observed object category, such as vehicles or pedestrians in the scene. Not all object categories need to be present in each time step, and their total number available in each time step is tracked as I_t. The observations are written as vectors allowing the generative process modeling them to be written more generally. In considering the observation $\boldsymbol{x}_{n,t,i} \sim F_n(\boldsymbol{\Phi}_{n,z_{t,i}})$ of feature n and treating the observation as a vector of length J, the activities of an individual feature take the form $\boldsymbol{\Phi}_k = \left\{\varphi_k^j\right\}_{j=1}^{J}$ for a collection of J multinomials, as in (1). Other forms include $\boldsymbol{\Phi}_k = \boldsymbol{\varphi}_k$, where the observations consists of multiple draws from the same multinomial, as well as $\boldsymbol{\Phi}_k = \left\{\varphi_k^j\right\}_{j=1}^{J}$, where the observations are represented by a collection of J Poisson distributions. Representing features this way is convenient since any missing features can be marginalized out. A graphical representation for this multi-feature dLDA model is shown in Figure 4. In the work presented here, we empirically find the model is robust to various settings of the hyperparameters and we consistently observe empty clusters ($K = 50$) suggesting that the model consistently learns the appropriate degree of complexity in the data.

5. EXPERIMENTAL RESULTS

To evaluate the ability of the multi-feature dLDA model to learn types of activity, we used publicly available video data from the Next Generation Simulation Community (NGSIM) (http://ngsim-community.org) of vehicle and pedestrian traffic centered on an intersection in a U.S. city. The video consists of a fixed view of an intersection on Peachtree Street in Atlanta, GA, sampled at 20Hz and taken from the vantage point of the top of a nearby building. To match the activity categories the model finds with easily identifiable attributes of each intersection, we chose to match the model outputs to

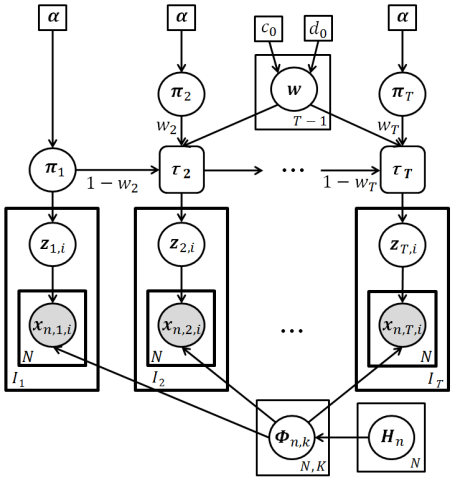
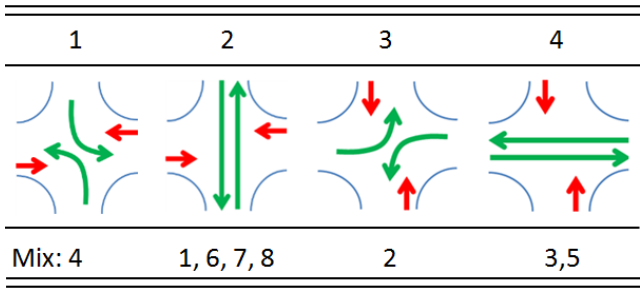


Fig. 1. (a) Graphical representation for the multi-feature dLDA model.

traffic light states within each intersection since they control vehicle and pedestrian flow through the intersection. The ground-truth of these light states can effectively be gathered by watching the video and compared directly with activity topics (traffic flow) learned by the model, thereby allowing a quantitative analysis of model performance. Through visual inspection of the video, the traffic light states of go (green), stop (red), and protected turn (green arrow) are evident, whereas the warning of transition from go to stop (yellow) is not, and is not considered here as a separate state. In the video data there are four distinct traffic light states that are cycled through, depicted in Table 1.

Table 1. Association of multi-feature dLDA mixture components with traffic light states.



In this video data, the multi-feature dLDA learns eight topic mixture components of traffic activity within this intersection (vehicles shown in Fig. 2). The topic mixtures capture well the traffic flow corresponding to particular traffic light states. The model allows an increasing sophistication through time, and additional mixture components (5-8) are found in the second half of the data. The vehicle tracks in activity one show a mixture of right of way traffic flow and left turns. However, in the video, it is clear that many left turns occur at

times when the traffic going straight has the right of way. In this scenario, the left turns occur when there is a break in the oncoming traffic and are not protected left turns. This is accurately captured by the model. The protected left turns in this direction are captured by activity four. In activity two, there is a similar mixing of left turns and oncoming straight traffic. However, here the model captures mostly protected turns, but there is a little overlap with the oncoming straight "right of way" traffic light state.

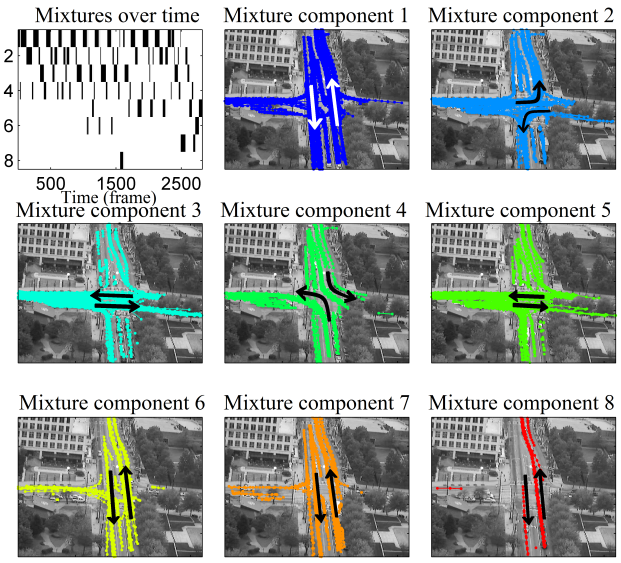


Fig. 2. Multi-feature dLDA mixture components through time (upper left) and vehicle tracks segmented by most likely mixture component. Mixtures are ranked by their abundance.

To generate quantitative results of model prediction, we used the first 75% (2233 frames) of the video for model training and the last 25% (758 frames) for testing. In our testing, we ran the multi-feature dLDA model with both vehicle and pedestrian tracks as well as with vehicle tracks alone. By removing the dynamic timing portion of the model, we effectively create a multi-feature LDA model that we are able to generate comparative results with using the vehicle tracks. Additionally, we compare these results with both the dLDA and the LDA model using only the single best feature (heading of accelerating vehicles). The comparative ROC curves for each model's ability to predict traffic light state, based on the manually extracted ground-truth is shown in Fig. 5, and the area under the curve (AUC) calculations for each curve are provided in Table 2. In these results, the multi-feature dLDA model outperforms each of the other models we tested for traffic light state prediction. The multi-feature dLDA model with vehicles alone has a marginally better AUC than the multi-feature dLDA model with both vehicles and pedestrian tracks. Due to the nature of the more erratic pedestrian traffic, it adds complexity and degrades performance. This suggests the need for future work in feature refinement. Both LDA models learned only the two most prominent states and lead to the least accurate predictions. Overall, it is clear that allowing the activity mixtures to vary in time allows better traffic light state prediction, and that predictive performance improves by having multiple observable features.

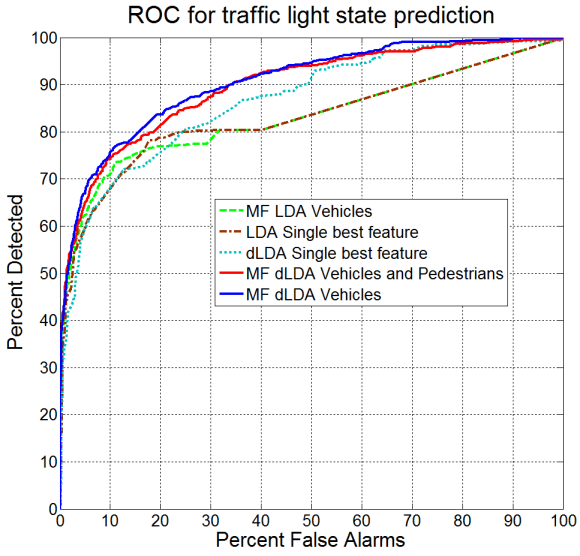


Fig. 3. Performance results for traffic light state prediction using various models. The proposed model, MF dLDA, performs best. Neither LDA method is able to find either of the protected turn traffic light states.

Table 2. Area Under the Curve Measures for Traffic Light State Prediction

Training model	Data	AUC
Multi-feature dLDA	vehicle tracks	0.904
Multi-feature dLDA	vehicle tracks and pedestrian tracks	0.896
dLDA	single best feature from vehicle tracks	0.861
Multi-feature LDA	vehicle tracks	0.773
LDA	single best feature from vehicle tracks	0.768

6. CONCLUSION

In this paper we present a natural extension of dLDA [18] that accommodates multiple independent observational features and simultaneous object classes, applied to complex video data of intersections in a U.S. city. The use of dLDA to activity modeling is new, and we quantitatively show in our example that the proposed multi-feature dLDA model outperforms a variety of established topic models. Additionally, the multi-feature nature of the model inherently lends itself to a wide range of time-evolving systems, making it adaptable to any number of scenarios where unsupervised time-evolving clustering is useful.

7. REFERENCES

[1] W. Hu, T. Tan, L. Wang, and S Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 3, pp. 334–352, 2004.

[2] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: Real-time surveillance of people and their activities," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 22, pp. 809–830, 2000.

[3] N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 22, pp. 831–843, 2000.

[4] C. Stauffer and E. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 22, pp. 747–757, 2000.

[5] G. Medioni, I. Cohen, F. BreAmond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 23, pp. 873–889, 2001.

[6] T.T. Truyen, D.Q. Phung, H.H. Bui, and S. Venkatesh, "Adaboost.mrf: Boosted markov random forests and application to multilevel activity recognition," in *Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition*, 2006.

[7] L. Zelnik-Manor and M. Iran, "Event-based analysis of video," in *Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition*, 2001.

[8] H. Zhong, J. Shi, and M. Visontaii, "Detecting unusual activity in video," in *Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition*, 2004.

[9] T. Xiang and S. Gong, "Video behaviour profiling and abnormality detection without manual labelling," in *Proc. Intl Conf. Computer Vision*, 2005.

[10] Y. Wang, T. Jiang, M.S. Drew, Z. Li, and G. Mori, "Unsupervised discovery of action classes," in *Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition*, 2006.

[11] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 31, no. 3, pp. 539–555, 2009.

[12] T. Hospedales, S. Gong, and T. Xiang, "A markov clustering topic model for mining behaviour in video," in *IEEE 12th International Conference on Computer Vision*, Kyoto, 2009, pp. 1165–1172.

[13] J. Woodworth, A. Eliazar, C. Lunsford, L. Kennedy, M. Groenert, S. Jellish, and J. Hilger, "Automated exploitation of wide area persistent surveillance imagery.," in *Parallel Meeting of the Military Sensing Symposium, Passive Sensors*, February 2009.

[14] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. International l Joint Conference Articial Intelligence*, 1981, pp. 674–680.

[15] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[16] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[17] M.J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[18] I. Pruteanu-Malinici, L. Ren, J. Paisley, E. Wang, and L. Carin, "Hierarchical bayesian modeling of topics in time-stamped documents," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 32, pp. 996–1011, June 2010.

[19] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R.Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[20] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50–57.

[21] D. Blei, A. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–10224, March 2003.

[22] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 55–56, November 2010.

[23] L. Ren, D.B. Dunson, and L. Carin, "The dynamic hierarchical dirichlet process," in *Proceedings of the International Conference on Machine Learning*, 2008.