Time based Activity Inference using Latent Dirichlet Allocation

Tanveer A Faruquie
tanveer@cse.iitd.ac.in
Prem K Kalra
pkalra@cse.iitd.ac.in
Subhashis Banerjee
suban@cse.iitd.ac.in

Dept. of Computer Science & Engg.
Indian Institute of Technology
New Delhi, India

Understanding and analyzing activities has traditionally relied on detecting and tracking objects throughout the scene. The tracked motions are then used to model activities [3] [2]. The advantage of tracking is that it inherently separates activities of one object from another. In crowded videos or poor quality videos it is difficult to track objects; usually only low level features can be computed. Latent topic models have recently been applied to alleviate this problem by modeling activities as hidden variables termed as topics. Basic principle behind these topic models is to discover co-occurrence patterns of low level features in a scene. The activities are distinguished from each other by their differing probability distribution over these low level features. Several topic models, such as Probabilistic Latent Semantic Analysis (pLSA) [5] [4], Latent Dirichlet Allocation (LDA) [1] and Hierarchical Dirichlet Process (HDP) [6] have been applied to discover activities in videos for different scenarios.

However, activities do not just have static co-occurrence patterns among features. Activities are localized in time and may have relevance only for a certain period. An activity may also overlap or non-overlap with some other activities at the same time. Topic models employed so far do not incorporate the dynamic time dependent nature of activities. In this paper, we focus on discovering the time-dependent behavior of activities using a Latent Dirichlet Allocation (LDA) Model.

We consider videos of crowded scenes where a host of problems like occlusions, object view changes, and different object shapes make tracking difficult. We divide the complete video into M clips having fixed number of frames. These clips correspond to documents. We compute optical flow between frames within the clip and cluster these optical flow vectors using agglomerative clustering. The number of clusters is determined using a threshold on the distance metric. We first cluster based on direction and all the vectors belonging to a particular direction cluster are then further clustered based on their location. Finally we prune clusters which do not have enough number of features. The clusters which survive pruning are treated as words w_i. These words come from a fixed size vocabulary V. The vocabulary is constructed by dividing the frame into fixed size cells and encoding the direction and location of these cells as vocabulary.

The graphical model for incorporating time in LDA is shown in Figure 1. This model takes cluster co-occurrences along with the temporal information of these clusters. We associate a continuous distribution over time for each of the K activities [7]. The parameterized distribution chosen for time is beta distribution, which defines a probability distribution over a normalized time range from 0 to 1.

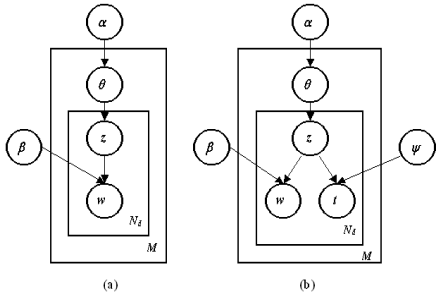

Figure 1: Graphical Representation of Topic Models (a) LDA (b) Time LDA

The generative model for timestamps and clusters using this model is as follows:

1. For each clip d, draw a multinomial distribution θ_d from a Dirichlet prior α.
2. For the i-th cluster, w_{di} in clip d (where $i = 1,\ldots,N_d$):
   (a) Draw an activity z_{di} from the multinomial θ_d.
   (b) Draw a cluster w_{di} from $p(w_{di}|z_{di},\beta)$, a multinomial distribution conditioned on activity z_{di}.
   (c) Draw a timestamp t_{di} from a Beta distribution $\psi_{z_{di}}$.

As can be seen from the above generative process the posterior distribution of activities depends on both, the clusters that represent the activities and the time when they occur. In this model both the clusters and the timestamps are the visible variables whereas θ_d and z_{di} are the hidden variables. The hyperparameters are α, β and ψ. Given these hyperparameters the joint probability distribution of clusters, activities and timestamp occurrences for clip d is given by

$$p(\mathbf{w_d},\mathbf{t_d},\mathbf{z_d},\theta_d|\alpha,\beta,\psi) = p(\theta_d|\alpha)\prod_{i=1}^{N_d} p(z_{di}|\theta_d)p(t_{di}|\psi_{z_{di}})p(w_{di}|z_{di},\beta)$$

Like LDA, here too exact inference cannot be done because computing the marginal likelihood $p(\mathbf{w_d},\mathbf{t_d}|\alpha,\beta,\psi)$ is intractable. Therefore, we propose a variational Bayes inference algorithm to approximate the posterior distribution $p(\theta_d,\mathbf{z_d}|\alpha,\beta,\psi)$. The algorithm is

1. For every clip d, find the optimal values for the variational parameters, γ_d^* and ψ_d^* for fixed hyperparameters. This is the E-step. The update equations for variational parameters for each clip given n-th cluster and i-th activity are

$$\gamma_i = \alpha_i + \sum_{n=1}^{N}\phi_{ni}$$

$$\phi_{ni} \propto \beta_{iw_n}\exp\{\Psi(\gamma_i) - \Psi\Big(\sum_{j=1}^{k}\gamma_j\Big) + \psi_{i1}\log t_n + \psi_{i2}\log(1-t_n)\}$$

   Here Ψ is the digamma function and ψ_{i1}, ψ_{i2} are the Beta distribution parameters.

2. Maximize the bound on log likelihood with respect to α, β and ψ for fixed variational parameters. This is the M-step. The bound is given by

$$l(\alpha,\beta,\psi) = \sum_{d=1}^{M}\log p(\mathbf{w_d},\mathbf{t_d}|\alpha,\beta,\psi).$$

   We use variational inference algorithm to bound the log likelihood. Here β can be computed analytically [1], whereas for α and ψ we use an efficient Newton-Raphson method.

The two steps are repeated alternatively until the log likelihood converges.

[1] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *CVPR*, 1997.

[3] S. Hongeng and R. Nevatia. Multi-agent event recognition. In *ICCV*, pages 84–93, 2001.

[4] J. Li, S. Gong, and T. Xiang. Global behaviour inference using probabilistic latent semantic analysis. In *BMVC*, 2008.

[5] J. C. Niebles, H. Wang, and F. Li. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[6] Y.W. Teh, M.I. Jordon, M.J. Beal, and D.M. Blei. Hierarchical dirichlet process. *Journal of the American Statistical Association*, pages 1566–1581, 2006.

[7] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. pages 424–433. KDD, 2006.