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

Abstract

In this paper we address the problem of time based activity inference in unsupervised manner for an area under surveillance. We use a Latent Dirichlet Allocation based model that captures the activities and how they change over time. We use agglomerative clustering on optical flow vectors to code direction and spatial information. In this model each activity is associated with not only a mixture distribution over these cluster occurrences but also on the distribution over timestamps of their occurrences. Our method thus helps in determining the prominence and the correlation of activities over a period of time.

1 Introduction

Understanding and analyzing activities has been the central problem in the area of surveillance. Research over the past few decades has led to a variety of useful methods to model and detect activities. Most of these approaches rely on detecting and tracking objects throughout a scene and use the tracked motions to model activities [4]. 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.

The Probabilistic Latent Semantic Analysis (pLSA) [3] model was initially proposed for extracting semantically meaningful topics of linguistic words in text documents. Basic principle behind pLSA and other topic models is to discover co-occurrence patterns of words in a collection of documents. Several topic models, such as pLSA [3], Latent Dirichlet Allocation (LDA) [1] and Hierarchical Dirichlet Process (HDP) [8] have been applied to discover activities in videos for different scenarios. In these models, activities are distinguished from each other by their differing probability distribution over observed low level features.

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. For example in a surveillance system for banks it may be perfectly normal to discover *usual* activities near the safety vault during daytime, the discovery of same *usual* activities after office hours

BMVC 2009 doi:10.5244/C.23.70

is suspicious. An activity may also overlap or non-overlap with some other activities. For example, in a rail road crossing, the activity of train passing over the track and people crossing the track is mutually exclusive. Discovery of these two activities at the same instance is alarming. 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. In order to capture activity features we rely on local motion features, computed using optical flow. These features can be easily and reliably computed for crowded scenes. Since topic models work on documents and words, we divide a video into clips with a fixed number of frames. These clips are treated as documents. We use agglomerative clustering of optical flow with centroid coding to quantize the features in a word vocabulary. We also capture the timestamps associated with these clusters along with their occurrence count to construct document-word matrix. We use Beta distribution to discover the probability distribution associated with each activity over normalized timestamps. To justify the use of Beta distribution consider a surveillance video of a suburban railway station. This video will witness a myriad of activities associated with events such as people catching or alighting from trains, going to commercial establishments, buying tickets, looking up information and using public facilities. Capturing the time distribution of such activities over a period will lead to the discovery of the rise and fall of activity prominence and their co-occurrence patterns. These time distributions can have many different shapes. They can be narrow and recurring for localized activities happening at specific times like when people catch or alight from trains, broad and spread for activities that keep occurring, like people moving regularly near information booths or commercial establishments or bursty for activities associated due to sudden changes, such as change in station number of train arrival. Hence it is imperative to choose a distribution that can capture variety of shapes like Beta distribution which is a parameterized model that can take versatile shapes.

The generative process of activities can be thought of as follows: We first choose the possible number of activities that can occur. Then we choose the distribution of these activities in each clip by sampling a multinomial distribution from a Dirichlet. For each occurrence of optical flow cluster we associate an activity by sampling from this multinomial and draw the cluster representation (direction and location) from a per activity multinomial distribution conditioned on this activity. Finally, we choose the timestamp of this cluster by drawing from a Beta distribution conditioned on this activity [10]. To infer the values of parameters of these distributions, we present a new Variational Bayes algorithm.

The rest of paper is organized as follows: In section 2, we briefly comment on some related works. In section 3.1, we describe the low level features we use for generating the word-document matrix for a given video clip. In section 3, we use LDA model to discover activities from this matrix. In section 4, we describe the LDA model used to incorporate timestamps to determine time distribution. We also describe the variational Bayes inference algorithm and use it to infer the parameter values. We present the results on some real life data sets in section 5 and conclude in section 6.

2 Related Work

The approaches employed for activity discovery and analysis in visual surveillance can be broadly classified in two categories. In the first approach, features of interest are detected

and tracked. The tracks are then used to model activities. This modeling can be done in a supervised fashion. Among the supervised learning methods one method is to learn and classify basic events, such as “left-turn", “stop", “exit", and then model complicated activities using these basic units [4]. Other approaches use statistical models to learn from tracks. Coupled HMMs [2], Bayesian networks and ballistic dynamics [9] are some of the models that have been researched. These methods do not work well when they fail to detect and track the features of interest, for example in crowded scenes. The second approach directly uses low level features instead of tracks as the description of video. Methods using this approach avoid the pitfalls of detection and tracking, however, these methods do not scale when many activities simultaneously occur in the video. They can deal with only one activity occurring at a time and thus can detect only the whole video sequence as normal or abnormal.

Recently topic models have been used to model behavior correlations within and across video clips represented as documents. Probabilistic semantic analysis (pLSA) [3] has been used for extracting object categories [5] and recognizing object actions in an unsupervised manner [7]. LDA and Hierarchical Bayesian models [10] have also been used to group low-level motion features into topics to discover activities. A Hierarchical pLSA [6] has been proposed to incorporate semantic scene representation to model global and local behavior patterns. However these models do not incorporate timestamps associated with the occurrence of activities. Our work is different from these models in which we treat time as an observed continuous variable like word occurrences. This also avoids the problem of time discretization. Our model essentially learns the co-occurrence patterns of activities over time, without labeling different activities at different points of time.

3 Latent Dirichlet Allocation

Before describing the incorporation of time in Latent Dirichlet Allocation model let us review the basic LDA model. The graphical model for LDA is shown in Figure 2. LDA models a document corpus consisting of M documents. Each document in this corpus is modeled as a mixture of K topics, where K is known a priori. In its generative process, for each document d having N_d words, a multinomial distribution θ_d is randomly sampled from a Dirichlet distribution with parameter α. To generate the i-th word in the document, first a topic z_{di} is chosen from this multinomial distribution and then a word w_{di} is generated by randomly sampling from a topic specific multinomial distribution β_z over the vocabulary V. In this model θ_d and z_{di} are the hidden variables and α and β are the hyper parameters to be optimized. For given α and β the joint probability distribution of clusters, $\mathbf{w_d}$, and activities, $\mathbf{z_d}$, for a clip is given by

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

Unfortunately since the marginal likelihood $p(\mathbf{w_d}|\alpha,\beta)$ is intractable, the posterior distribution $p(\theta_d, \mathbf{z_d}|\alpha,\beta)$ is also intractable. Therefore, a variational Bayes inference algorithm [1] is used to approximate the posterior distribution $p(\theta_d, \mathbf{z_d}|\alpha,\beta)$ using free variational parameters. This distribution is given by

$$q(\theta_d, \mathbf{z_d}|\gamma_d, \phi_d) = q(\theta_d|\gamma_d) \prod_{i=1}^{N_d} q(z_{di}|\phi_{di})$$

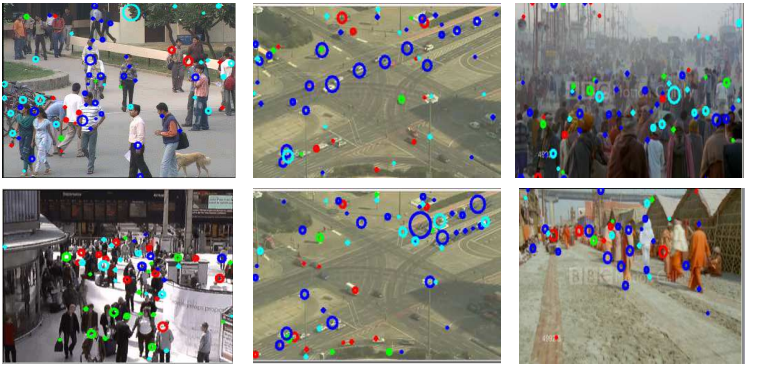

Figure 1: Example of visual words obtained for video sequences

Here, γ_d are variational Dirichlet parameters and ϕ_{di} are variational multinomial parameters. In the next section we present an approach to construct document word matrix for videos having crowded scenes.

3.1 Visual Words

We are working with surveillance video streams of public places recorded by a fixed camera. Processing this data is challenging because of object occlusions, lightening changes and different object types. We divide the complete video into M clips having fixed number of frames. These clips correspond to documents. Since tracking is unreliable for these situations we use optical flow features for computing visual words. In order to detect areas having cohesive motion we cluster these optical flow vectors. We first capture the directional cohesiveness by first clustering based on direction. 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 (320×280) into fixed size (10×10) cells. The direction (4) and location (centroid) of cluster (32×28) is thus coded with a word from this vocabulary ($32 \times 28 \times 4$).

We use agglomerative clustering because the number of clusters is not required *a priori*. In agglomerative clustering clusters are generated by using a threshold on the distance metric. The threshold in our case translates to the maximum permissible distance between motion clusters thus allowing us to ensure that local motion features of different close by objects are not merged together. Figure 1 shows examples of clusters obtained for our video sequences. The distance threshold used is the same in all the clips which results in variable number of clusters. The clusters obtained for different directions are shown in different colors. In next section we describe how LDA model over time is used for discovering activities in video sequences.

4 Time LDA

The LDA model described in the previous section takes cluster co-occurrences into account but do not account for temporal information associated with these clusters. By incorporating normalized timestamps of these clusters one can model the long range dependencies of activ-

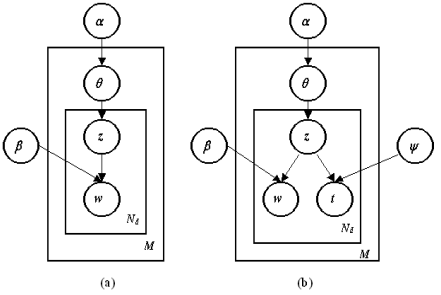

Figure 2: Graphical Representation of Topic Models (a) LDA (b) Time LDA

ities over time. This is useful to predict activity probability for a given time. One possibility is to discretize time and find the activities within these discrete ranges. Activities discovered in the time range depicts where they are dominant. Unfortunately it is very difficult to chose granularity of time slice which is suitable for all activities. The time slice chosen may be too large for some activities or too small for others. In order to avoid this problem we associate a continuous distribution over time for each activity [10]. The parameterized distribution chosen for time is Beta distribution, which defines a probability distribution over a normalized time range from 0 to 1. The generative model for timestamps and clusters of clips 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 a 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 $\psi_{z_{di}}$.

The graphical model representation for this model is shown in figure 2. 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

$$p(\mathbf{w_d},\mathbf{t_d},\mathbf{z_d},\theta_d|\alpha,\beta,\psi) = p(\theta_d|\alpha)p(\mathbf{z_d}|\theta_d)p(\mathbf{t_d}|\mathbf{z_d},\psi)p(\mathbf{w_d}|\mathbf{z_d},\beta)$$

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

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 $p(\theta_d,\mathbf{z_d}|\alpha,\beta,\psi)$. In section 4.1 we present the algorithm to compute the variational parameters and in section 4.2 we present the algorithm to compute the hyperparameter for time.

4.1 Variational Inference

If we drop the explicit notation showing the dependence on the clips, the marginal log likelihood for a clip is given by

$$\log p(\mathbf{w},\mathbf{t}|\alpha,\beta,\psi) = \log \int_{\theta} \sum_{z} p(\mathbf{w},\mathbf{t},\theta,\mathbf{z}|\alpha,\beta,\psi).$$

This probability can be bounded using Jensen's inequality of concave functions for some variational probability distribution $q(\cdot)$. The free variational parameters are the Dirichlet parameter γ and the multinomial parameter ϕ.

$$\log p(\mathbf{w},\mathbf{t}|\alpha,\beta,\psi) = \log \int_{\theta} \sum_{z} \frac{p(\mathbf{w},\mathbf{t},\theta,\mathbf{z}|\alpha,\beta,\psi)}{q(\theta,\mathbf{z}|\gamma,\phi)} \cdot q(\theta,\mathbf{z}|\gamma,\phi)$$

$$\geq \int_{\theta} \sum_{z} q(\theta,\mathbf{z}|\gamma,\phi) \log \frac{p(\mathbf{w},\mathbf{t},\theta,\mathbf{z}|\alpha,\beta,\psi)}{q(\theta,\mathbf{z}|\gamma,\phi)}$$

We show in Appendix 7.1 that using Variational Bayes, $\log p(\mathbf{w},\mathbf{t}|\alpha,\beta,\psi)$ can be approximated by maximizing the lower bound on this probability with respect to γ and ϕ. Maximizing the lower bound minimizes the KL divergence between the variational posterior probability and true posterior probability. Hence the set of variational parameters γ^* and ϕ^* that minimize the KL divergence

$$(\gamma^*,\phi^*) = argmin_{\gamma,\phi} D(q(\theta,\mathbf{z}|\gamma,\phi)||p(\mathbf{w},\mathbf{t},\theta,\mathbf{z}|\alpha,\beta,\psi)),$$

is obtained by computing the derivative of lower bound and setting them to zero. Doing so we obtain the following update equations for variational parameters for each clip given n-th cluster and i-th activity

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

$$\phi_{ni} \propto \beta_{iw_n} \exp\{\Psi(\gamma_i) - \Psi\Big(\sum_{j=1}^{k} \gamma_j\Big) + \psi_{i1}\log t_n + \psi_{i2}\log(1-t_n)\} \tag{2}$$

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

4.2 Hyperparameter estimation

Given a surveillance video V divided into M clips with each clip d represented by $\mathbf{w_d}$ clusters observed at normalized times $\mathbf{t_d}$, we wish to find the parameters α, β and ψ that maximize the log likelihood of observing the clusters and their timings in this video.

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

As stated earlier $p(\mathbf{w_d},\mathbf{t_d}|\alpha,\beta,\psi)$ cannot be computed tractably and hence we use variational inference algorithm to bound the log likelihood. For this we use an alternating variational EM algorithm that maximizes the lower bound for fixed variational parameters γ and ϕ. The maximization for α and β is identical to LDA [1]. The derivation of ψ is given in Appendix 7.2. This yields the following iterative algorithm

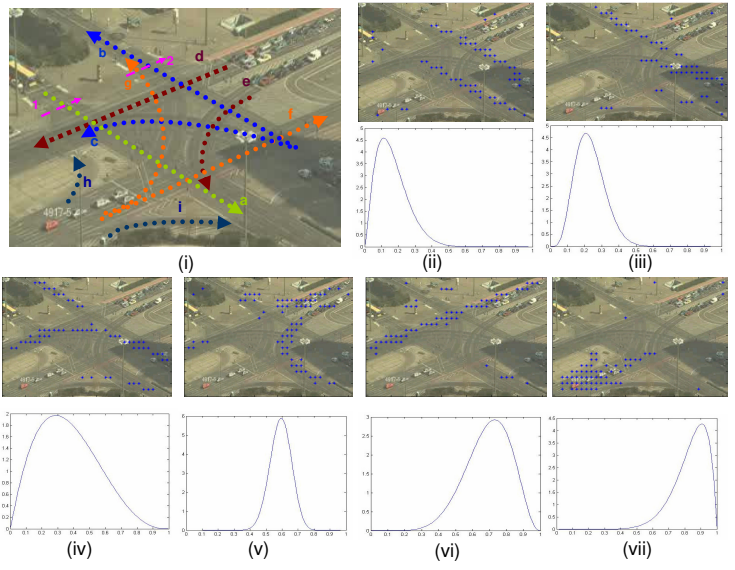

Figure 3: Activities discovered along with their time distribution. (i) The actual pedestrian and traffic flow paths (ii)-(vii) The discovered activities, arranged in the order they are observed in time.

1. For every clip d, find the optimal values for the variational parameters, γ_d^* and ψ_d^* for fixed hyperparameters, as described in section 4.1. This is the E-step.

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

The two steps are repeated alternatively until the log likelihood converges. The M-step for β can be computed analytically [1], whereas the M-step for α and ψ uses an efficient Newton-Raphson method.

5 Experimental Results

We evaluated the performance of this model using video data from a busy traffic intersection which is controlled by four traffic lights. The traffic data is a far field video at 10 frames per second with a frame size of 320 x 210. The total number of frames used is 43000 (i.e. 4300 non-overlapping clips where each clip is of 10 frames). This results in a vocabulary size of 2688 words. The number of unique clusters discovered were 367. The time taken for a complete traffic signal cycle is considered to be an epoch. All the timestamps of the clusters are drawn from the time range of this epoch. The actual timestamps are offset and normalized between 0 and 1 with respect to the beginning of the epoch. Clusters and their normalized timestamps are considered for several cycles. Figure 3 (i) shows some of the paths which the vehicles can take either freely (*h, i*) or guided by traffic signals (*a* to *g*). Besides vehicle movement there are pedestrian movements (1, 2). We expect that paths that

are taken freely (*h, i*,1 ,2) will contribute to broad time distribution whereas those paths which depend on traffic signalling (*a* to *g*) will tend to have narrow time distribution since they will be more localized in time. We show the result of training our model in figure 3 (ii - vii) where we have set $K = 11$. We plotted the cluster co-occurrences for some of the activity clusters. The centroid of the cluster discovered are marked in blue. We have also plotted the activity-dependent time distribution obtained for these activities. The activities have been arranged in the order in which they occur in one cycle of time in the video.

As can be seen, the time distributions have correctly discovered the ordering of activities on the time axis. Activities which tend to happen earlier in the cycle have a higher probability of occurring earlier, whereas, activities that happen later have a right sided distribution. Any deviation from the discovered order can be flagged as unusual. The activity shown in figure 3 (ii) occurs when the traffic starts flowing in direction *a* and *b*. The time distribution for this activity states that the probability of seeing motion in direction *a* and *b* is highest at the beginning (time starts when the first traffic light goes green). The motion in *b* is divided into two separate activities (ii and iii) because it co-occurs with *a* and *c*. Therefore, the time distribution for these activities have overlapping time range.

Apart from the controlled traffic pattern there are other pedestrian motions and free turns which are also discovered in these activity patterns. Surprisingly, these uncontrolled motions also have time dependent behavior. The probability of pedestrian motion 1 and 2 is highest in the normalized time range of 0.5 to 0.7 and co-occurs with motion *e* and some portion of motion *d*. This is the reason why activity in (vi) have a broader time distribution whereas activity (v) have a narrower distribution because (vi) is trying to accommodate the time dependence of *d* and to some extent of 2. The free turns also exhibit temporal dependence. The free turn shown by *i* occurs only in activities (iii) and (iv) which comprises of flows *b* and *c*. This can be explained by the fact that at all other times the road carries traffic of flows *a, e* which inhibits people from using *i*. The time distribution of activity (iv) is broad because it captures the time dependence of flows *c* and *i*. The free turn *h* and *i* is observed to be used only at the start of flow *f*. This is because the bulk of traffic waits to take direction *f* which prevents the flow in direction *h* and *i*. Therefore time distribution of activity (vii) captures the time dependence of *f, h* which tend to co-occur. The presence of the start of flow *i* contributes to the long tail in left direction of the distribution.

6 Conclusion

In several surveillance applications like train stations, banks, traffic etc, there can be specific time distributions associated with some activities. Our motivation in this paper was to develop a framework that mines these activities from poor quality crowded videos, determines their time distributions, and extracts any hidden time dependencies about the activities. To overcome the problems associated with crowded scenes we use an agglomerative clustering approach that works reliably on low level motion vectors. We present an efficient variational inference algorithm to model the time distribution of these activities with LDA. We validated our method on real life traffic video. Our method can determine the prominence and the correlation of activities over a period of time. It can also adapt with the evolution of activities over time. Discovering this information can be used in a variety of ways like Egress planning, security management, facility management and discovering hotspots prominence over time.

7 Appendix

In this section we present additional derivations that are required for capturing activity dependent time distributions.

7.1 Variational Multinomial

Finding a tight bound on posterior probability distribution will require maximizing $L(\gamma,\phi;\alpha,\beta,\psi)$ [1] which is given by

$$L \equiv E_q[\log\, p(\theta|\alpha)] + E_q[\log\, p(\mathbf{z}|\theta)] + E_q[\log\, p(\mathbf{t}|\mathbf{z},\psi)] + E_q[\log\, p(\mathbf{w}|\mathbf{z},\beta)] \\ -E_q[\log\, q(\theta)] - E_q[\log\, q(\mathbf{z})].$$

Specifically, the expectation of timestamps for document d is given by,

$$E_q[\log\, p(\mathbf{t}|z,\psi)] = \sum_{n=1}^{N_d} E_q[\log\, p(t_{dn}|z_{dn},\psi)]$$

$$= \sum_{n=1}^{N_d}\sum_{i=1}^{K}\Big(\phi_{ni}\log\,\Gamma(\psi_{i1}+\psi_{i2}) \;+\; \phi_{ni}\log\,\Gamma(\psi_{i1}) \;+$$

$$\phi_{ni}\log\,\Gamma(\psi_{i2}) \;+\; \phi_{ni}\psi_{i1}\,\log\,t_n \;+\; \phi_{ni}\psi_{i1}\,\log\,(1-t_n)\Big)$$

Here ϕ_{ni} is the probability that n-th word is generated by topic $i(=z_{dn})$. By taking all the other terms containing ϕ_{ni} [1], adding Lagrange multipliers, and then taking derivative with respect to ϕ_{ni} we get

$$L_{\phi_{ni}} = \phi_{ni}(\Psi(\gamma_i) - \Psi(\textstyle\sum_{j=1}^{k}\gamma_j)) + \phi_{ni}\log\beta_{iv} - \phi_{ni}\log\phi_{ni} + \phi_{ni}\psi_{i1}\log t_n + \\ \phi_{ni}\psi_{i2}\log(1-t_n) + \phi_{ni}\log\Gamma(\psi_{i1}+\psi_{i2}) - \phi_{ni}\log\Gamma(\psi_{i1}) - \phi_{ni}\log\Gamma(\psi_{i2}) + \lambda_n(\textstyle\sum_{j=1}^{k}\phi_{ni}-1)$$

$$\frac{\partial L}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi(\textstyle\sum_{j=1}^{k}\gamma_j) + \log\beta_{iv} - \log\phi_{ni} - 1 + \psi_{i1}\log t_n + \psi_{i2}\log(1-t_n) + \\ \log\Gamma(\psi_{i1}+\psi_{i2}) - \log\Gamma(\psi_{i1}) - \log\Gamma(\psi_{i2}) + \lambda_n$$

Setting this derivative to zero will give the value of variational parameter ϕ_{ni} that maximizes L, and is given by

$$\phi_{ni} \propto \beta_{iv}\exp\Big\{\Psi(\gamma_i) - \Psi\Big(\sum_{j=1}^{k}\gamma_j\Big) + \psi_{i1}\log t_n + \psi_{i2}\log(1-t_n)\Big\} \tag{3}$$

7.2 Conditional Beta distribution

The model parameters that maximize the intractable log likelihood over all documents can be determined by considering variational lower bound in place of marginal log likelihood. Therefore, collecting the terms containing $\psi_{ij}(j=1,2)$ in L and taking the derivative with respect to ψ_{ij}. we get

$$L_{\psi_{ij}} = \sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{i=1}^{K}\phi_{dni}\left(\log\Gamma(\sum_{l=1}^{2}\psi_{il}) - \sum_{j=1}^{2}\log\Gamma(\psi_{ij}) + \sum_{j=1}^{2}\psi_{ij}\log\, t_{nj}\right)$$

$$\frac{\partial L}{\partial \psi_{ij}} = \sum_{d=1}^{M}\sum_{n=1}^{N_d} \phi_{dni}\left(\Psi(\sum_{l=1}^{2}\psi_{il}) - \Psi(\psi_{ij})\right) + \sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}\log t_{nj}$$

Because of the coupling between ψ's the equation has to be solved in an iterative manner using Newton Raphson method. The gradient is given by

$$\frac{\partial L}{\partial \psi_{ij}\partial \psi_{il}} = \begin{cases} \sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}\left(\Psi'\left(\sum_{r=1}^{2}\psi_{ir}\right) - \Psi'(\psi_{ij})\right) & \text{for } j = l \\ \sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}\Psi'(\sum_{r=1}^{2}\psi_{ir}) & \text{for } j \neq l \end{cases}$$

i.e., for $\forall i$,

$$\frac{\partial L}{\partial \psi_{ij}\partial \psi_{il}} = -\delta(j,l)\Psi'(\psi_{il})K^i + K^i\Psi'(\sum_{j=1}^{2}\psi_{il}), \tag{4}$$

where $K^i = \sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}$. This is in a Hessian form $H^i = K^i(diag(h) + 1z1^T)$ where the j-th diagonal element is $h^i_j = -\Psi'(\psi_{ij}).K^i$ and $z = \Psi'(\sum_{j=1}^{2}\psi_{ij}).K^i$. This avoids the costly matrix inversion operation and can be solved efficiently in an iterative manner.

References

[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*, volume 29, pages 213–244, 1997.

[3] T. Hoffmann. Probabilistic latent semantic analysis. In *SIGIR*, pages 50–57, 1999.

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

[5] F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

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

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

[8] 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.

[9] S.N.P. Vitaladevuni, V. Kellokumpu, and L.S. Davis. Action recognition using ballistic dynamics. In *CVPR*, 2008.

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

[11] X. Wang, X. Ma, and W.E.L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE PAMI*, 31(3):539–555, 2009. ISSN 0162-8828.