IDIAP RESEARCH REPORT

A SPARSITY CONSTRAINT FOR TOPIC MODELS - APPLICATION TO TEMPORAL ACTIVITY MINING

Jagannadan Varadarajan[a] Remi Emonet
Jean-Marc Odobez

Idiap-RR-36-2010

OCTOBER 2010

[a]Idiap Research Institute

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

A Sparsity Constraint for Topic Models - Application to Temporal Activity Mining

Jagannadan Varadarajan[1,2], **Rémi Emonet**[1], **Jean-Marc Odobez**[1,2]
[1]Idiap Research Institute, CH-1920 Martigny, Switzerland.
[2]Ecole Polytechnique Federal de Lausanne, CH-1015 Lausanne, Switzerland.
{vjagann,remi.emonet,odobez}@idiap.ch

Abstract

We address the mining of sequential activity patterns from document logs given as word-time occurrences. We achieve this using topics that models both the co-occurrence and the temporal order in which words occur within a temporal window. Discovering such topics, which is particularly hard when multiple activities can occur simultaneously, is conducted through the joint inference of the temporal topics and of their starting times, allowing the implicit alignment of the same activity occurrences in the document. A current issue is that while we would like topic starting times to be represented by sparse distributions, this is not achieved in practice. Thus, in this paper, we propose a method that encourages sparsity, by adding regularization constraints on the searched distributions. The constraints can be used with most topic models (e.g. PLSA, LDA) and lead to a simple modified version of the EM standard optimization procedure. The effect of the sparsity constraint on our activity model and the robustness improvement in the presence of difference noises have been validated on synthetic data. Its effectiveness is also illustrated in video activity analysis, where the discovered topics capture frequent patterns that implicitly represent typical trajectories of scene objects.

1 Introduction

Topic models, which allow to extract dominant patterns in the data from simple un-ordered feature counts, have given encouraging results in several areas of computer vision. This is the case in automatic activity analysis from large volumes of video footages encountered for instance in the surveillance domain. There, one would like to automatically discover typical activity patterns, their start and end, or predict an object's behavior. Such information can be useful in better understanding the scene content and its dynamics, or to provide context for other tasks like object tracking. By considering quantized spatio-temporal visual features (e.g. optical flow) as words and short video clips as documents, topic models like pLSA [1] or LDA [2] have been used to discover scene level activity patterns and detect abnormal events [3, 4, 5].

However, activities occurring in scene are implicitly temporally ordered and using unordered word co-occurrence within a time window fails to represent their sequential nature. Recently, several topic models have been proposed to include sequential information, e.g. by modeling single word sequences [6, 7], or at the high level, i.e. by modeling the dynamics of topic distributions over time [8]. Many of these temporal models have been adapted for activity analysis. For instance, [9] introduced a Markov chain on scene level behaviors, but each behavior is still considered as a mixture of unordered (activity) words. And [10] exploits [11] to mine topical trends within a traffic signal cycle which requires explicit manual synchronization of clips to signal cycles.

Still, unlike in text analysis, a common situation in visual scenes is that multiple temporal activity patterns are happening simultaneously without being necessarily synchronized at the high level (e.g. motions of pedestrians and cars are independent, unless people want to cross the street). None of the above models actually were designed to handle this case. In [12], we introduced the Probabilistic

Latent Sequential Motif (pLSM) topic model to discover dominant activity patterns from sensor data logs represented by word×time count documents. Its main features are i) estimated patterns are not merely defined as static word distribution but also incorporate the temporal order in which words occur; ii) data with the temporal overlap between several activities can be dealt with; iii) automatic estimation of activity pattern starting times is done.

One common issue in non-parametric topic models is that distributions are often loosely constrained, resulting in non-sparse process representations which are often not desirable in practice. For instance, in PLSA, one would like each document d to be represented by few topics z with high weights $p(z|d)$, but nothing encourages this. The same applies to LDA models despite the presence of priors on the multinomial $p(z|d)$ [13]. This sparsity issue has been rarely dealt with in the literature. Very recently, [13] proposed a model that decouples the request for sparsity and the smoothing effect of dirichlet prior, by introducing explicit selector variables determining which terms appears in a topic. The Focused topic model of [14] addresses sparsity for Hierarchical Dirichlet Process by exploiting an Indian Buffet Process to impose sparse yet flexible document topic distributions.

In this paper, our contribution is to propose an alternative and simple approach to this problem. The main idea is to guide the learning process towards sparser (more peaky) distributions characterized by a smaller entropy. To simplify the learning, we achieve this indirectly by adding a regularization constraint in the EM optimization procedure that maximizes the Kullback-Leibler distance between the uniform distribution (maximum entropy) and the distribution to be learnt. This results in a simple procedure that can be applied to any distribution for which such a sparsity constraint is desirable. In this paper, we apply and demonstrate the usefulness of this approach for our model.

In the rest of the paper, we introduce our pLSM model, along with the proposed learning procedure that incorporates our sparsity constraints. The model and its properties are then validated on synthetic experiments and illustrated on real surveillance videos.

2 Probabilistic Latent Sequential Motif Model

In this section, we provide an overview of the generative model and then present our inference process, explaining how we enforce sparsity on some model distributions.

2.1 Model overview and generative process

Figure 1a illustrates how documents are generated. Let D be the number of documents d in the corpus, each spanning T_d discrete time steps. Let $V = \{w_i\}_{i=1}^{N_w}$ be the vocabulary of words that can occur at any given instant $t_a = 1, ..T_d$. A document is described by its count matrix $\mathrm{n}(w, t_a, d)$ indicating the number of times a word w occurs at the absolute time t_a. These documents are generated from a set of N_z topics $\{z_i\}_{i=1}^{N_z}$ assumed to be temporal patterns $p(w, t_r|z)$ with a maximal duration of T_z time steps (t_r denotes the relative time at which a word occurs within a topic) and that can start at any time instant t_s within the document. Qualitatively, documents triplets (w, t_a, d) are generated by sampling words in the topic temporal patterns and placing them in the document relatively to a sampled starting time according to (cf Fig.1a):

- draw a document d with probability $p(d)$;
- draw a latent topic $z \sim p(z|d)$;
- draw the starting time $t_s \sim p(t_s|z, d)$, where $p(t_s|z, d)$ denotes the probability that the topic z starts at time t_s within the document d;
- draw a word $w \sim p(w|z)$;
- draw the relative time $t_r \sim p(t_r|w, z)$, where $p(t_r|w, z)$ denotes the probability that the word w within the topic z occurs at time t_r;
- set $t_a = t_s + t_r$, which assumes that $p(t_a|t_s, t_r) = \delta(t_a - (t_s + t_r))$, that is, the probability density function $p(t_a|t_s, t_r)$ is a Dirac function.

The main assumption with the above model is that the occurrence of a word only depends on the topic, not on the time instant when a topic occurs. given the deterministic relation between the three time variables ($t_a = t_s + t_r$), the joint distribution of all variables can be written as:

$$p(w, t_a, d, z, t_s) = p(d)p(z|d)p(t_s|z, d)p(w|z)p(t_a - t_s|w, z) \quad (1)$$

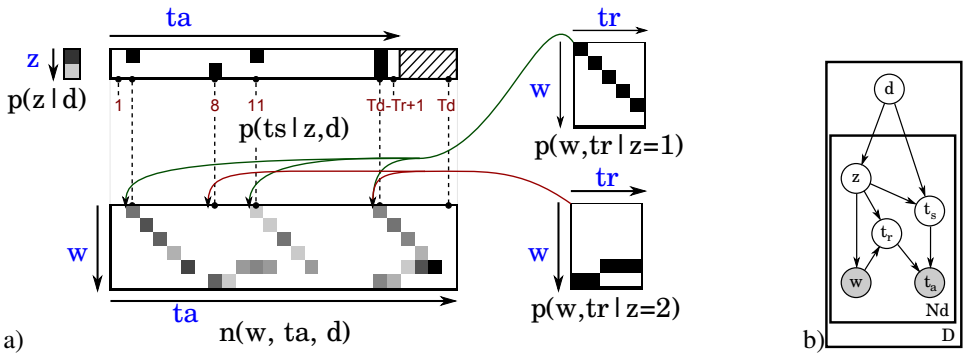


Figure 1: a) document $\mathrm{n}(w, t_a, d)$ generation. Words $(w, t_a = t_s + t_r)$ are obtained by first sampling the topics and their starting times from the $p(z|d)$ and $p(t_s|z, d)$ distributions, and then sampling the word and its temporal occurrence within the topic from $p(w, t_r|z)$. b) graphical model.

2.2 Model inference and sparsity

Our goal is to discover the topics and their starting times given the set of documents $\mathrm{n}(w, t_a, d)$. The model parameters Θ can be estimated by maximizing the log-likelihood of the observed data $\mathcal{D}$, which is obtained through marginalization over the hidden variables $Y = \{t_s, z\}$:

$$\mathcal{L}(\mathcal{D}|\Theta) = \sum_{d=1}^{D} \sum_{w=1}^{N_w} \sum_{t_a=1}^{T_d} n(w, t_a, d) \log \sum_{z=1}^{N_z} \sum_{t_s=1}^{T_{ds}} p(w, t_a, d, z, t_s) \tag{2}$$

Such an optimization can be performed using an Expectation-Maximization (EM) approach, maximizing the expectation of the complete log-likelihood. However, as motivated in the introduction, the estimated distributions may exhibit a non-sparse structure that is not desirable in practice. In our model this is the case of $p(t_s|z, d)$: one would expect this distribution to be peaky, exhibiting high values for only a limited number of time instants t_s. To encourage this, we propose to guide the learning process towards sparser distributions characterized by smaller entropy, and achieve this indirectly by adding to the data likelihood a regularization constraint to maximize the Kullback-Leibler distance $D_{KL}(U||p(t_s|z, d))$ between the uniform distribution U (maximum entropy) and the distribution of interest. After development and removing the constant term, our constrained objective function is now given by:

$$\mathcal{L}_c(\mathcal{D}|\Theta) = \mathcal{L}(\mathcal{D}|\Theta) - \sum_{z,d} \frac{\lambda_{z,d}}{T_{ds}} * \log(p(t_s|z, d)) \tag{3}$$

The EM algorithm can be easily applied to the modified objective function. In the E-step, the posterior distribution of hidden variables is calculated as (the joint probability is given by Eq. 1):

$$p(z, t_s|w, t_a, d) = \frac{p(w, t_a d, z, t_s)}{p(w, t_a, d)} \text{ with } p(w, t_a, d) = \sum_{z=1}^{N_z} \sum_{t_s=1}^{T_{ds}} p(w, t_a, d, z, t_s) \tag{4}$$

In the M-step, the model parameters (the probability tables) are updated according to:

$$p(z|d) \propto \sum_{t_s=1}^{T_{ds}} \sum_{t_r=0}^{T_z-1} \sum_{w=1}^{N_w} \mathrm{n}(w, t_s + t_r, d) p(z, t_s|w, t_s + t_r, d) \tag{5}$$

$$p(t_s|z, d) \propto \max\left(0, \sum_{w=1}^{N_w} \sum_{t_r=0}^{T_z-1} \mathrm{n}(w, t_s + t_r, d) p(z, t_s|w, t_s + t_r, d) - \frac{\lambda_{z,d}}{T_{ds}}\right) \tag{6}$$

$$p_w(w|z) \propto \sum_{d=1}^{D} \sum_{t_s=1}^{T_{ds}} \sum_{t_r=0}^{T_z-1} \mathrm{n}(w, t_s + t_r, d) p(z, t_s|w, t_s + t_r, d) \tag{7}$$

$$p_{t_r}(t_r|w, z) \propto \sum_{d=1}^{D} \sum_{t_s=1}^{T_{ds}} \mathrm{n}(w, t_s + t_r, d) p(z, t_s|w, t_s + t_r, d) \tag{8}$$

Qualitatively, in the E-step, the responsibilities of the topic occurrences in explaining the word pairs (w,t_a) are computed (high responsibilities are obtained for informative words, i.e. words appearing

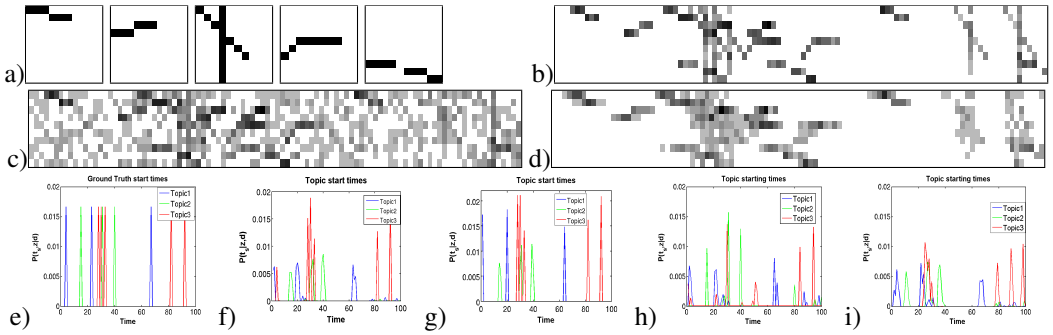


Figure 2: Synthetic experiments. (a) the five topics, (b) a segment of a generated document, (c,d) the same segment perturbed with: (c) uniform noise ($\sigma_{snr} = 1$), (d) Gaussian noise ($\sigma = 1$) added to each word time occurrence t_a. (e) the true topic occurrences (only 3 of them are shown for clarity) in the document segment shown in (b). (f-i) the recovered topic occurrences $p(t_s|z, d)$; (f) the clean document (cf b) and no sparsity constraint $\lambda = 0$ (g) and $\lambda = 0.5$; (h) the noisy document (c) and $\lambda = 0.5$ (i) the noisy document (d) and $\lambda = 0.5$.

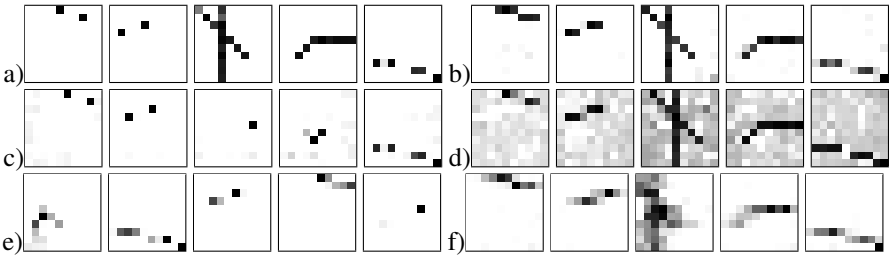


Figure 3: Recovered topics without (a,c,e) and with (b,d,f) sparsity constraints $\lambda = 0.5$ (a,b) from clean data; (c,d) from documents perturbed with random noise words, $\sigma_{snr} = 1$, cf Fig.2c; (e,f) from documents perturbed with Gaussian noise on location $\sigma = 1$, cf Fig.2d.

in only one topic and at a specific time), whereas the M-steps aggregates these responsibilities to infer the topic patterns and occurrences. Importantly, thanks to the E-steps, the multiple occurrences of an activity in documents are implicitly aligned in order to learn its pattern.
Finally, when looking at Equation 6, we see that the effect of the introduced constraint is to set to 0 the probability of terms which are lower than $\frac{\lambda_{z,d}}{T_{ds}}$ thus increasing the sparsity as desired.

3 Experiments on synthetic data

Synthetic data is used to demonstrate the strength of our model and the effect of the sparsity constraint. Using a vocabulary of 10 words, we created five topics with duration ranging between 6 and 10 time steps (see Fig. 2a). Then, we created 10 documents of 2000 time steps assuming equiprobable topics and 60 random occurrences per topic. In the rest of the article, average results from the 10 documents and corresponding error-bars are reported. One hundred time steps of one document are shown in Fig. 2b, where the intensities represents the word count (larger counts are darker), and Fig. 2e shows the corresponding starting times of three out of the five topics. As can be noticed, there is a large amount of overlap between topics. Finally, in Eq. 6 we defined $\lambda_{z,d} = \lambda \frac{\mathrm{n}_d}{N_z}$, where n_d denotes the total number of words in the document, and use λ to denote the sparsity level. Note that when $\lambda = 1$, the correction term $\frac{\lambda_{z,d}}{T_{ds}}$ is, on average, of the same order of magnitude than the first part of the right hand side in Eq. 6.

Results on clean data. Figures 3a and 3b illustrate the recovered topics with and without the sparsity constraint. As can be seen, without sparsity, two of the obtained topics are not well recovered. This can be explained as follows. Consider the first of the five topics. Samples of this topic motif starting at a given instant t_s in the document can be equivalently obtained by sampling words from the learnt topic 3a and sampling the starting time from three consecutive t_s values with lower probabilities instead of one. This can be visualized in Fig.2f, where the peaks in the blue curve $p(t_s|z = 1, d)$ are

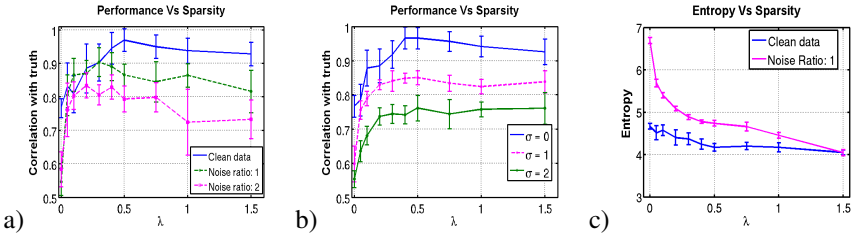


Figure 4: Average topic correlation between the estimated and the ground truth topics for different sparsity weight λ and for different levels of (a) the uniform noise, (b) the Gaussian noise on a word time occurrence t_a. (c) Average entropy of $p(ts|z,d)$ in function of the sparsity λ.

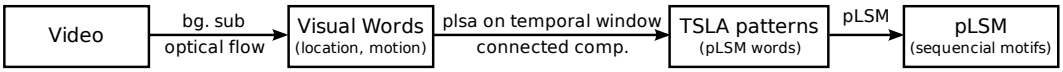


Figure 5: Flowchart for discovering sequential activity motifs in videos.

three times wider and lower than in the ground truth. When using the sparsity constraint, the topics are well recovered, and the starting time occurrences better estimated.

Robustness to Noise and sparsity effect. Two types of noise were used to test the method's robustness. In the first case, words were added to the clean documents by randomly sampling the time instant t_a and the word w from a uniform distribution, as illustrated in Fig. 2c. The amount of noise is quantified by the ratio $\sigma_{snr} = N_w^{noise}/N_w^{true}$ where, N_w^{noise} denotes the number of noise words added and N_w^{true} the number of words in the clean document. The learning performance is evaluated by measuring, the average correlation between the learned topics $\hat{p}(t_r, w|z)$ and the true topics $p(t_r, w|z)$ (i.e. $\frac{1}{N_z}\sum_{t_r,w}\hat{p}(t_r, w|z).p(t_r, w|z)$) (See Fig. 4). Noise can also be due to variability in the temporal execution of the activity. This 'location noise' was simulated by adding random shifts (sampled from Gaussian noise with $\sigma \in [0, 2]$) to the time occurrence t_a of each word, resulting in blurry documents (see Fig. 2d). Fig. 2c-f illustrates the recovered topics. Without sparsity constraint, the topic patterns are not well recovered (even the vertical topic). With the sparsity constraint, topics are well recovered, but reflect the effects of the generated noise, i.e. uniform noise in the first case, temporal blurring in the second case. Fig. 4 shows that the model is able to handle quite a large amount of noise in both cases, and that the sparsity approach provide significantly better results. Finally, we validate that, as desired, there is an inverse relation between the sparsity constraint and the entropy of $p(t_s|z,d)$ which is clearly seen in Fig. 4c.

Number of Topics and Topic Length. In [12], it was shown that requesting longer topics than in the ground truth or more topics than necessary does not affect learning, but that performance are usually significantly worse with no sparsity constraint.

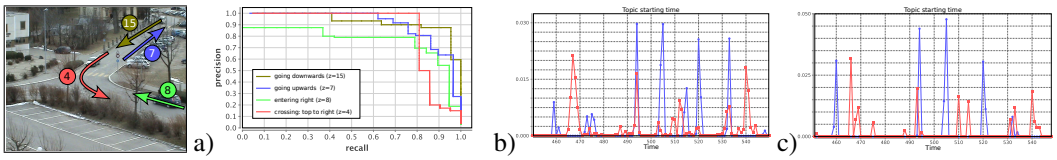
4 Scene activity patterns

4.1 Activity words

We also applied our pLSM model to discover temporal activity patterns from real life scenes. This work flow is summarized in Fig. 5. To apply the pLSM model on videos, we need to define the words w forming its vocabulary. Instead of using low-level features directly, we perform a dimensionality reduction step on the low level features as done in [12] by applying pLSA on location, and optical flow velocity features obtained from the video. Thus, we obtain temporally and spatially localized activity (TSLA) patterns from the low-level features and use the occurrences of these as our words to discover sequential activity motifs (SM) in pLSM model.Thus, N_A dominant TSLA patterns obtained from pLSA define our words for PLSM i.e. $N_w = N_A$. The word counts defining the PLSM documents d are then built from the amount of presence of these TSLA patterns in the sequence of d_{t_a} documents.

4.2 Results

Experiments were carried out on two complex scenes. The **Far Field** video contains 108 minutes of a three-road junction captured from a distance, where typical activities are moving vehicles. As the scene is not controlled by a traffic signal, activities have large temporal variations. The **Traffic**



a) b) c)

Figure 6: a)Interpolated Precision/Recall curves for the detection of 4 types of events mapped onto 4 topics evaluated on a 10 minute test video.b)$p(ts|z,d)$ for two activities without sparsity constraint, c)$p(ts|z,d)$ with sparsity constraints.

Junction video is 45 minutes long and captures a portion of a busy traffic-light-controlled road junction. Activities include people walking on the pavement or waiting before crossing over the zebras, and vehicles moving in and out of the scene.

In the interest of space and better illustration we have provided sample clips and comprehensive results at `http://www.idiap.ch/paper/1930/sup.html` [15]. Given the scene complexity and the expected number of typical activities, we arbitrarily set the number N_z of sequential motifs (SM) to 15 and the motif duration T_z to 10 time steps (10 seconds). Top ranking SMs from the datasets, (provided in the web-page [15]) exactly correspond to the dominant patterns in the scene namely, vehicle moving along the main road in both directions in the far field data. In the Traffic Junction scene, despite the low amount of data, the motifs represent well vehicular movements, pedestrian activities, and complex temporal interactions between vehicles and pedestrians.

Event detection. We also did a quantitative evaluation of how well pLSM can be used to detect particular events. We can create an event detector by considering the most probable occurrences $p(t_s, z|d)$ of a topic z in a test document d. By setting and varying a threshold on $p(t_s, z|d)$ we can control the trade-off between precision and completeness. For this event detection task, we labelled a 10 minute video clip from the far field scene, distinct from the training set, and considered 4 events depicted in Fig. 6. To each event type, we manually associated a topic, built an event detector and varied the decision threshold to obtain precision/recall curves. Fig. 6 shows the obtained results.

4.3 Sparsity effect

The Sparsity constraint employed on $p(t_s, z|d)$ distribution resulted in clear peaks for the motif start times (see Fig. 6c) as opposed to smoother distributions obtained without the sparsity constraint (Fig. 6b). This was useful in removing some of the false alarms and improving the quantitative results in the event detection task. However, looking at the motifs qualitatively revealed that a sparse $p(t_s, z|d)$ distribution results in smoother motifs: the uncertainty in start times is transfered to the time axis of the motifs. This effect can be clearly observed on synthetic data (in Fig. 3f vs Fig. 3b) and the examples in the web page [15].

5 Conclusion

In this article, we extended a topic-based method for temporal activity mining. The underlying model used here extracts temporal patterns from documents where multiple activities occur simultaneously. Our contribution is to encourage sparsity in the model demonstrated specifically on the motif start times.

We introduced sparsity by adding a regularization constraint on learnt distributions. The effect of sparsity on both synthetic data under variety of noise and real life data was studied. Results in both settings show that sparsity constraint improves the quality of recovered activity patterns and increases the model's robustness to noise. The formulation of the regularization constraint as an entropy minimization makes it straightforward to introduce in the EM optimization, and can be similarly introduced in most topic models like pLSA and LDA.

References

[1] T. Hofmann. Unsupervised learning by probability latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[2] D. M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Machine Learning Research*, (3):993–1022, 2003.

[3] J. Varadarajan and J.M. Odobez. Topic models for scene analysis and abnormality detection. In *ICCV-12th International Workshop on Visual Surveillance*, Kyoto, Japan, 2009.

[4] Jian Li, S. Gong, and T. Xiang. Global behaviour inference using probabilistic latent semantic analysis. In *British Machine Vision Conference*, 2008.

[5] Xiaogang Wang, Xiaoxu Ma, and Eric L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Trans. on PAMI*, 31(3):539–555, 2009.

[6] Hanna M. Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977 – 984, Pittsburgh, Pennsylvania, 2006.

[7] Amit Gruber, Michal Rosen-Zvi, and Yair Weiss. Hidden topic markov model. *Intelligence and Statistics (AISTATS)*, March 2007.

[8] D. Blei and J. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[9] Timothy Hospedales, S. Gong, and Tao Xiang. A markov clustering topic model for mining behavior in video. In *International Conference in Computer Vision*, Kyoto, Japan, 2009.

[10] Tanveer A Faruquie, Prem K Kalra, and Subhashis Banerjee. Time based activity inference using latent dirichlet allocation. In *British Machine Vision Conference*, London, UK, 2009.

[11] Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *In Conference on Knowledge Discovery and Data Mining (KDD) 2006*, Philadelphia, USA, 2006.

[12] J. Varadarajan, R. Emonet, and J.-M. Odobez. Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes. In *British Machine Vision Conference (BMVC), Aberystwyth*, 2010.

[13] C. Wang and D.M. Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *NIPS*, page 19821989, 2009.

[14] Sinead Williamson, Chong Wang, Katherine Heller, and David Blei. Focused topic models. In *NIPS workshop on Applications for Topic Models: Text and Beyond*, Whistler, Canada., 2009.

[15] http://www.idiap.ch/paper/1930/sup.html.