Unsupervised discovery of activity correlations using Latent Topic Models

Tanveer A Faruquie[*]
Indian Institute of Technology
Hauz Khas
New Delhi, India
tanveer@cse.iitd.ac.in

Subhashis Banerjee
Indian Institute of Technology
Hauz Khas
New Delhi, India
suban@cse.iitd.ac.in

Prem K Kalra
Indian Institute of Technology
Hauz Khas
New Delhi, India
pkalra@cse.iitd.ac.in

ABSTRACT

Topic models such as probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) have been successfully used to discover individual activities in a scene. However these methods do not discover group activities which are commonly observed in real life videos of public places. In this paper we address the problem of discovering activities and their associations as a group activity in an unsupervised manner. We propose a method that uses a two layer hierarchical latent structure to correlate individual activities in lower layer with group activity in higher layer. Our model considers each scene to be composed of a mixture of group activities. Each group activity is in turn composed as a mixture of individual activities represented as multinomial distributions. Each individual activity is represented as a distribution over local visual features. We use a Gibbs sampling based algorithm to infer these activities. Our method can summarize not only the individual activities but also the common group activities in a video. We demonstrate the strength of our method by mining activities and the salient correlation amongst them in real life videos of crowded public scenes.

1. INTRODUCTION

The proliferation of camera in many public and private places is producing an ever increasing amount of data. In practice this data is recorded for archival purposes and is often retroactively utilized on a need to know basis. For active utilization of this data it is imperative to avoid the expensive and laborious manual analysis and labeling of data. Automatic scene understanding and behavior mining is thus important to analyze the increasing volume of video surveillance feeds. In this work, we address the problem of automatic scene understanding and activity mining. Given a video containing scenes of outdoor public spaces we want the system to automatically answer questions like What are the activity patterns of individual? Do these activities occur as a group? If yes how? How is the global behavior composed from these individual and group activities?

Most of the past approaches based on detection and tracking have focused on independent isolated events and do not extend to general settings of complicated multi activity scenes. To address this shortcoming recent research has focused on using topic models, like probabilistic Latent Semantic Analysis (pLSA) [6], Latent Dirichlet Allocation (LDA) [1] and Hierarchical Dirichlet Processes (HDP) [16] to automatically learn activities by correlating local features. The approach taken is to divide the video into clips and then the model correlates the local features present in each clip as latent structures, called topics. Each topic is considered as an activity which are probability distributions over local features across video clips. Multiple topics are discovered by the model. Examples of local features that have been used with these models include optical flow vectors [11], foreground patches [12], spatio-temporal words [13] and space-time shape features extracted using epitomes [3]. Often these local features can be computed reliably and since the model inferences activities using global statistics across documents the model works well even if features are not detected in a few frames.

Although these methods detect multiple individual activities they do not model the occurrence of activities as a group. We propose to model the occurrence of group activities as association of individual activities that can co-occur with high likelihood. We model this latent association using a multinomial distribution. In our generative model each clip is composed as a multinomial of group activities drawn from a Dirichlet distribution and each group activity is composed as a multinomial of individual activities drawn from a group specific Dirichlet distribution. Individual activities are represented as multinomial distributions over local features. We use a Gibbs sampling algorithm to learn these distributions which gives us the individual activities and their associations as a group activity. The motivation of our method can be understood using an example presented in the following section.

Motivation

One of the most popular topic model to infer activities is LDA [1], which represents each activity as a multinomial distribution over local features and each video clip is represented as a mixture of these activities. While this model learns activities by capturing the co-occurrence of local features it does not capture the co-occurrence of activities within a scene. This is particularly needed in many realistic set-

[*] Corresponding author

ICVGIP '10, December 12-15, 2010, Chennai, India

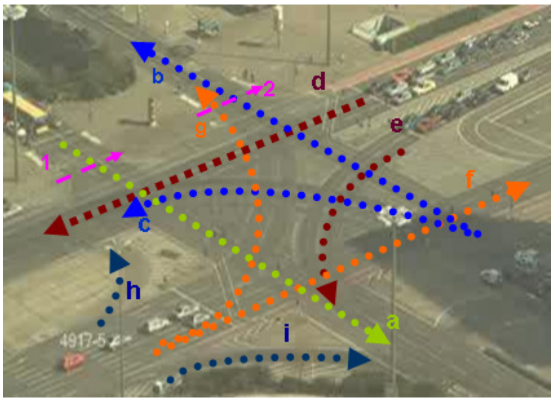

Figure 1: Scene from a traffic surveillance camera

ting of surveillance application where a scene often consists of multiple activities representing a group and individual activities. Consider for example a traffic surveillance system which is monitoring an intersection shown in figure 1. This scene consists of nine individual vehicular activities $a-i$ and two individual pedestrian activities $1-2$. Typically a scene will witness a combination of these individual activities. But not all co-occurrence of activities will have equal likelihood, some activities will co-occur more frequently than others. In particular the following pair of activities (b, c), (d, e) and (g, h) will almost always co-occur because each pair is governed by the same traffic signal. Similarly since activity h represents a free turn it can co-occur with any other activity but has less likelihood of co-occurring with c or d because of high chances of collision with on coming traffic on c and d than with, say g. The chances of co-occurrence of 1 or 2 with a, g or b is minimal as it is unlikely that pedestrians will cross a road when traffic is flowing on that road.

LDA can learn individual activities $a-i$ and $1-2$ as multinomial probability distributions over local features and the presence of these activities in a video clip is represented as a draw from a Dirichlet distribution representing the proportions of these activities. Thus any combination of activities is possible within a clip. This draw ignores the fact that some activities co-occur more frequently as a group whereas the co-occurrence of some activities may not make any sense (like a and d). To account for this drawback we propose a method to discover group activities using topic models.

Our Contribution

In this work we propose a method based on topic model to discover group and individual activities. In our method each video clip is represented as a mixture of group activities drawn from a Dirichlet and then individual activities in turn are drawn from the chosen group activities. The main contributions of our work can be summarized as follows.

1. Our method can extract group activities thus allowing us to discover complex interaction among different activities in the scene. It can also summarize the video into significant group activities as well as individual activities.
2. Our method can work with any local features and does not require expensive manual tuning and semantic input to discover activities and their groups within a scene.
3. We demonstrate the effectiveness of our model on real life surveillance videos of outdoor scenes and show its utility in activity discovery and interpretation.

The rest of the paper is organized as follows. In Section 2 we present related work. In Section 3 we describe our topic model for group activity discovery and in Section 4 we give an inference algorithm based on Gibbs sampling. We present the experimental results in Section 5.

2. RELATED WORK

Research on automatic activity understanding can be broadly classified in two categories. The first category uses an object centric view and often involves detecting and tracking features of interest of the object. These tracks are then used to automatically understand the activities. Methods within this category include supervised models which can learn specific semantic events and discover activities as a composition of these events [7] or unsupervised methods that cluster trajectories into activities [20] [9]. Other statistical models that incorporate spatial and temporal features of tracks include Coupled HMMs [2], Bayesian networks [19] and ballistic dynamics [17]. These methods do not work well when they fail to detect and track the features of interest, for example in crowded scenes.

The second category uses a feature centric approach and consists of methods that 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 can deal with only one activity occurring at a time and thus can detect only the whole video sequence as normal or abnormal. To overcome this shortcoming recently topic models such as pLSA [13] [4] [15], LDA [21] [23] and HDP [22] have been used to model multiple activities across video clips. LDA has been extended to model behavior across video clips [8] [11], interactions across document clips [18] and time dependency of activities over a period of time [5]. All these approaches model only individual isolated activities as multinomial distributions over low level features. They do not model group activities in the data that exist in the form of correlations of individual activities. We not only discover individual activities but also discover a group of activities that may occur frequently in data.

Work closest to our method is [18] and [12]. The approach in [12] segments the scene into semantic regions and uses a single pLSA for each region to learn activity within a semantic region. Global behavior pattern across semantic regions is modeled using hierarchical pLSA. This approach is not generative and requires to first segment scene into semantic regions. Moreover the assumption of activity correlation across semantic regions is restrictive as activities can be correlated within a region too. The approach in [18] overcomes the restrictive assumption of single Dirichlet prior over all video clips by assuming that video clips are coming from different clusters each governed by a separate Dirichlet prior. Thus common co-occurrence of activities in video clips governed by the same prior results in those video clips to be clustered together. We take a different approach where instead of clustering video clips based on activities we cluster activities across video clips using their co-occurrence statistics. Approaches in [8] and [11] model activity dependency across video clips using single and multiple Markov chains respectively. Instead of finding dependence of activities across video clips we find dependence of activities within a scene as co-occurrence patterns. Our work focuses on finding the composition of scene in terms of group activities which are in turn composed of individual activities.

3. GROUP ACTIVITY MINING

In this section we give a detailed description of our method that discovers group activities and individual activities. We begin with by describing the local visual features on which the model is trained.

3.1 Local Visual Features

Our goal is to construct a generative model capable of learning the individual and group activities in video data captured from single fixed view cameras monitoring public places. Processing this data is challenging because the scene may contain multiple group activities consisting of different objects in the presence of occlusions and lighting changes. As low level features we compute the optical flow across frames and also do a background subtraction. The complete video is then divided into M video clips containing fixed number of frames. The camera view (320 × 280) is divided into (10 × 10) pixel cells. When the proportion of the foreground pixel within a cell exceeds a threshold t_f and the magnitude of optical flow within a cell exceeds a threshold t_o we mark this as the presence of a local feature and code it using the position of the cell and the direction of optical flow quantized into one of the four directions. Thus a combination of $V = 3584$ $(32 \times 28 \times 4)$ local visual features are used to represent a clip and the activities are learnt as distributions over these local features.

3.2 Group Activity Topic Model

Standard LDA is a probabilistic graphical model which is used for learning activities present in a video in an unsupervised manner. This graphical model is depicted in figure 2(a). The video is divided into M clips and the scene within each clip is represented as a mixture of K activities, where K is known apriori. The generative process of scene is that for each clip d consisting of N_d local visual features, a multinomial distribution θ_d having K components, is randomly sampled from a Dirichlet distribution with parameter α. The probabilities associated with the K components of θ_d represent the mixing proportions of activities for that clip. To generate the j-th local visual feature in the clip, first an activity z_{dj} is chosen by making a draw from the multinomial distribution θ_d and then a local visual feature representation w_{dj} is generated by randomly sampling from an activity specific multinomial distribution ϕ_z. The distribution ϕ_z is over the vocabulary V consisting of all possible local visual features.

We model a scene in a video as consisting of three layer structure representing: group activities, individual activities and local features. This model is shown in figure 2(b) using the plate notation. The generative model can be thought of as follows. Suppose the scene is composed of M clips which consist of multiple co-occurring individual activities. Here, individual activities that frequently co-occur as a group is deemed as a group activity and there can be multiple group activities within a clip. Each individual activity can be part of more than one group activity. The possible number of group activities is given by K_G and the possible number of individual activities is given by K_I, where both K_G and K_I are known apriori. The process used to generate a scene is given by:

1. For each of the K_I individual activities sample a multinomial distribution ϕ_z having V components from a Dirichlet distribution with parameter β. The multinomial represents each individual activity as a distribtution over local visual features.

2. For each video clip d consisting of N_d local visual features

   (a) Sample a multinomial distribution θ_{Gd} having K_G components randomly from a Dirichlet distribution with parameter α_G. This multinomial represents the clip as a mixture of group activities.

   (b) Further K_G multinomials, $\theta^1_{Id}, \theta^2_{Id}, \ldots, \theta^{K_G}_{Id}$, having K_I components are randomly sampled from each of K_G Dirichlet distributions with parameters $\alpha^1_I, \alpha^2_I, \ldots, \alpha^{K_G}_I$ respectively. These multinomials represent the composition of group activities as a mixture of individual activities. Note each group activity is a distribution over individual activities and each individual activity can belong to multiple group activities.

   (c) For each of the j^{th} local visual feature of the N_d features

      - Sample a group activity z_{dgj} from the multinomial θ_{Gd} corresponding to the g^{th} component of θ_{Gd}.
      - Sample an individual activity z_{dij} from the multinomial θ^g_{Id} corresponding to the i^{th} component of θ^g_{Id}.
      - Sample the visual word representation w_{dj} from the multinomial distribution ϕ_{z_i} over the vocabulary V. Here V consists of all possible local visual features.

As is clear from the above generative process the local visual features w_d are the visible variables whereas $\theta_d = \{\theta_{Gd}, \theta^1_{Id}, \theta^2_{Id}, \ldots, \theta^{K_G}_{Id}\}$ and $z_d = \{z_{dgj}, z_{dij}\}$ are the hidden variables. The parameters are $\alpha = \{\alpha_G, \alpha^1_I, \alpha^2_I, \ldots, \alpha^{K_G}_I\}$ and ϕ which have to be learnt and the hyperparameter β is given.

Given the hyperparameter, the joint probability distribution of group activities, individual activities, topic assignments z_d and the local visual words w_d for clip d is given by

$$p(w_d, z_d, \theta_d|\alpha, \phi) = p(\theta_{Gd}|\alpha_G) \prod_{g=1}^{K_G} p(\theta^g_{Id}|\alpha^g_I)$$

$$\times \prod_{j=1}^{N_d} p(z_{dgj}|\theta_{Gd}) p(z_{dij}|\theta^g_{Id}) p(w_{dj}|\phi_{z_{dij}})$$

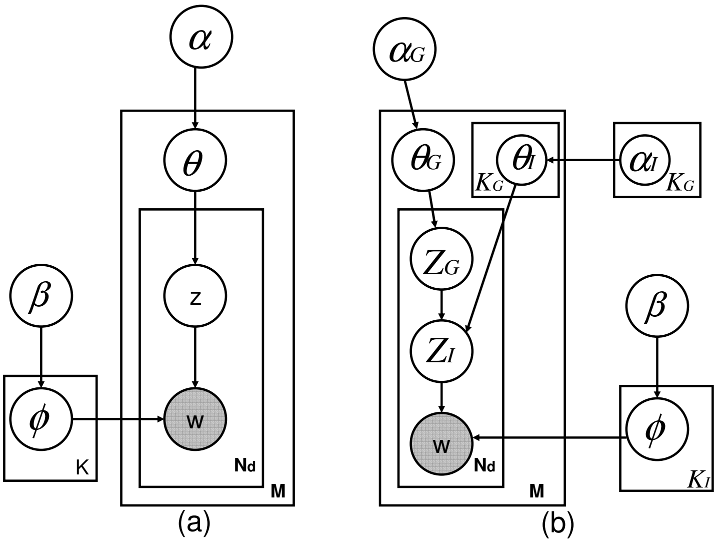

Figure 2: Plate notation for (a) LDA model (b) Group Activity Model

On integrating over the hidden variables θ_d, z_d and ϕ we get the marginal probability of clip d as

$$p(w_d|\alpha,\phi) = \int p(\theta_{Gd}|\alpha_G) \prod_{g=1}^{K_G} p(\theta^g_{Id}|\alpha^g_I) \tag{1}$$

$$\times \prod_{j=1}^{N_d} \sum_{z_{dgj}} \sum_{z_{dij}} p(z_{dgj}|\theta_{Gd}) p(z_{dij}|\theta^g_{Id}) p(w_{dj}|\phi_{z_{dij}}) d\theta$$

The probability of generating the whole scene $\mathbf{M}$ consisting of M video clips is the product of probabilities for each individual clip and is given by

$$p(\mathbf{M}|\alpha,\phi) = \prod_{d=1}^{M} p(w_d|\alpha,\phi)$$

On integrating out the multinomial distributions over individual activities we get

$$p(\mathbf{M}|\alpha,\beta) = \int \prod_{i}^{K_I} p(\phi_{z_i}|\beta) \prod_{d=1}^{M} p(w_d|\alpha,\phi) d\phi \tag{2}$$

Using bayes rule we can write the posterior distribution as

$$p(\theta_d, z_d, \phi|\alpha,\beta,w_d) = \frac{p(\theta_d, z_d, w_d, \phi|\alpha,\beta)}{p(w_d|\alpha,\beta)} \tag{3}$$

Finding the posterior distributions results in the discovery of latent variables. Computing the exact marginal likelihood shown in equation 1 is intractable because of which computing the posterior distribution is also intractable. Hence we have to use inferencing method which can approximate the posterior distribution. We present a Gibbs sampling algorithm to achieve this task.

4. PARAMETER ESTIMATION

As described in the previous section the hidden parameters of our model are the multinomial distributions Θ, the assignment of individual and group activities z to each local feature, the individual activity distributions ϕ and the Dirichlet parameters α_I of group activities. Since the posterior distribution in equation 3 is difficult to compute we use a Gibbs sampling based approximate inference. In Gibbs sampling we sample the individual activity and group activity assignment for each local feature based on the conditional probability of this assignment given the observation and assignments to other features. Note that the Dirichlet distribution and multinomial distribution belong to exponential family with the Dirichlet distribution having parameters α_I being the conjugate prior of multinomials Θ and β being the conjugate prior for ϕ. Hence we can integrate out Θ and ϕ to find the conditional probability $p(z_{dgj}, z_{dij}|M, z_{-j}, \alpha, \beta)$. For a local feature j in document d this is given by

$$p(z_{dgj}, z_{dij}|M, z_{-j}, \alpha, \beta) \propto$$
$$\frac{n_{ij}+\beta}{n_i + V.\beta} \cdot \frac{n_{d,gi} + \alpha^g_{Ii}}{n_{d,g} + \sum_{i'}^{K_G} \alpha^g_{Ii'}} \cdot \frac{n_{d,g} + \alpha_G}{n_d + K_G.\alpha_G}$$

Here z_{dgj} and z_{dij} correspond to the global activity, g, and individual activity, i, assignment for a local feature. Excluding the current local feature, $n_{d,g}$ is the number of times features in document d is assigned the global activity g and n_d is the total number of global activity assignment in d. $n_{d,gi}$ is the number of times features are assigned the individual activity i when they are sampled from the global activity g and $n_{d,g}$ is the total number of times global activity g is assigned to features in d. n_{ij} is the total number of times a feature j is assigned an individual activity i in complete video and n_i is the number of times features are assigned the individual activity i in the complete video. The α_G, α^g_{Ii} and β are the Dirichlet parameters.

The above equation is intuitive. Here the first ratio expresses the probability that the local feature j from V will belong to an individual activity. The second ratio expresses the probability of an individual activity i participating in composing a global activity g and the third ratio expresses the probability of global activity g being part of clip d. Since the association of individual activities to form a global activity has to be determined by the data the Dirichlet parameters α_I^g has to be updated. The updates can be learned using a maximum likelihood estimation (MLE). There is no closed form solution for MLE estimation of a Dirichlet and one can use iterative methods such as Gradient Ascent or Newton Raphson [10]. However to gain efficiency we approximate MLE using the method of moments. This method estimates the Dirichlet parameters by finding that density which matches the moments of data. The first two moments of Dirichlet parameters is given by

$$E[\bar{\alpha}_i^g] = \frac{1}{M}\sum_{d=1}^{M}(\frac{n_{d,gi}}{n_{d,g}}) = \frac{\alpha_i^g}{\sum_{i'}\alpha_{i'}^g} \tag{4}$$

$$E[\bar{\alpha}_i^{g2}] = \frac{1}{M}\sum_{d=1}^{M}(\frac{n_{d,gi}}{n_{d,g}})^2 = E[\bar{\alpha}_i^g]\frac{1+\alpha_i^g}{1+\sum_{i'}\alpha_{i'}^g} \tag{5}$$

The above two equations can be solved to get

$$\sum_{i'}\alpha_{i'}^g = \frac{E[\bar{\alpha}_i^g] - E[\bar{\alpha}_i^{g2}]}{E[\bar{\alpha}_i^{g2}] - E[\bar{\alpha}_i^g]^2} \tag{6}$$

By multiplying equation 4 and 6 α_i^g can be estimated. However since only one α_i^g is used for the estimation we use the method suggested by [14] to use all α_i^g. Hence we get

$$var(\bar{\alpha}_i^g) = \frac{E[\bar{\alpha}_i^g](1-E[\bar{\alpha}_i^g])}{1+\sum_{i'}\alpha_{i'}^g} = \frac{1}{M}\sum_{d=1}^{M}(\frac{n_{d,gi}}{n_{d,g}} - E[\bar{\alpha}_i^g])^2 \tag{7}$$

$$\sum_{i'}\alpha_{i'}^g = \exp\left[\frac{1}{K_I-1}\sum_{i'=1}^{K_I}\log\left(\frac{E[\bar{\alpha}_i^g](1-E[\bar{\alpha}_i^g])}{var(\bar{\alpha}_i^g)} - 1\right)\right] \tag{8}$$

equation 4 and 8 can be multiplied to estimate α_i^g. This estimation is done after every iteration of the Gibbs sampling.

5. EVALUATIONS

In this section we present the experimental evaluation of our model. We demonstrate the strength of our model on both the aspects, its ability to discover important individual activities in a scene and its ability to discover prominent activity groups in the video. We demonstrate our model on real life videos of public places.

5.1 Setup and Dataset

We evaluated the performance of our model using data from a single view camera meant to monitor crowded public scenes. We used two data sets to evaluate our method with video footage from a University Campus and a Traffic Junction.

University Campus: This data contained 45 mins of video at 24 fps with a frame size of 384 × 288. The camera monitored an area in a university surrounded by various services including a bicycle stand, book shop, coffee shop and a department. Different activities are performed by people in the scene depending on the interaction among each other and their use of these services. People enter or exit a department, go towards office, meet and discuss and fetch their bicycles. Apart from these people can walk across the area. Many of these activities can happen simultaneously in a scene. Apart from discovering the possible activities we are also interested in detecting the co-occurrence of these activities as a group. The camera was mounted for a near view scene at a lower level and hence results in severe occlusions which makes it even more difficult to segment and correlate individual activities.

Traffic Junction: This data is a video from a busy traffic junction. The scene is a far field video captured at 20 frames per second with a frame size of 320×210. The total length of the video is 20 minutes. Here the activities are paths taken by vehicles and the pedestrians and the co-occurrence of this activities is governed by the sequence of the four traffic signals that regulate the traffic flow. Hence the co-occurrence of different activity motions is know a prior because the sequence of signal activations and the corresponding traffic motion governed by the signal is known.

For both these videos the frame is divided into cells of size 10 × 10. After using background subtraction and quantizing optical flow in four directions the total number of local visual features obtained is 4408 (38 × 29 × 4) for the university campus data and 2688 (32 × 21 × 4) for the traffic junction data. The threshold for foreground t_f is set to 0.35 and the threshold for average flow magnitude t_o is set to 20. We ran the Gibbs sampler for a total of 4000 iterations and learnt the model by taking 100 samples at an interval of 10 iterations after a burn in period of 3000 iterations. The hyperparameter α_G is set to 0.1 and the hyperparameter β is set to 0.05.

5.2 Discovering Activities

We learnt the individual and group activities using our model by training it on the visual features extracted from the video of University Campus data. The number of individual activities K_I to be discovered was set to 12 and the number of group activities K_G to be discovered is set to 6. We ran the Gibbs sampling algorithm which took around 90 minutes on a 2.6 GHz machine with 2 GB RAM.

Individual Activities

The individual activities which are discovered by our method is shown in figure 3. As described earlier the activities in this area is driven by the presence of services. Figure 3(a) locates the presence of these services and it includes the department entrance and the book shop (1), entry to bicycle stand (2), way to residences and coffee shop (3), library (4) and the way to office building (5). Besides this people walk across this area and may meet and discuss in vicinity. The activities discovered by our model is shown in figures 3(b)-3(i). Since this video is taken at a time when most of the people leave the department a majority of activities follow the pattern from location 1 to locations 2-5. Figure 3(b) is an activity when people move towards their residences either when coming from department or moving right to left. Figures 3(c) and 3(e) are the activities when people leave the department and move towards the residence and towards the office respectively. Figure 3(g) represent the activity when people move towards the office either coming from location 1 or moving straight from left to right. Figures 3(d) and 3(f) represent the activities when people are moving towards the bookshop or leaving the department. Figure 3(h) is an interesting activity discovered by our model which is observed because of the presence of a couple of people standing and talking. While talking these people move which results in bidirectional optical flow vectors. Our model is able to detect this meeting as a separate activity without confusing it with other movements in the neighborhood. Similarly an-

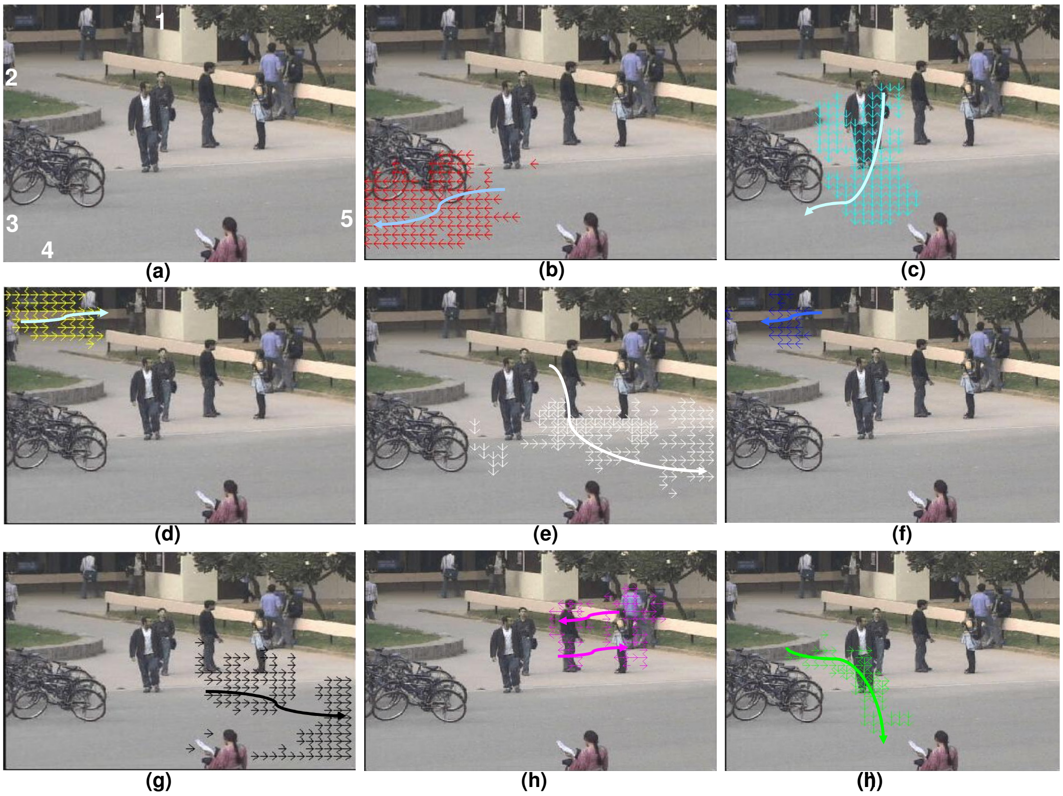

Figure 3: The activities discovered by our model (a) Represent the location of different services which gives rise to activities in this area. (b) - (i) describes the different activities ϕ_z which are discovered by our model. The local visual features which has high $p(w|\phi_z)$ is also plotted.

other interesting activity is shown in figure 3(h) which is the activity when people take their bicycle and leave the parking area. This activity is also clearly detected and separated from other activities.

Group Activities

Our model also discovers the group activities by finding the salient correlation among individual activities shown in figure 3. The prominent group activities discovered by our model is shown in figure 4. As described in the previous section the group activity is a correlation among individual activities. This correlation among activities is given by the parameter α_I^g which represent the prior on mixing weights of individual activities to be selected to compose a group activity g. In figure 4(a) the group activity represent the two different direction which people might take when they exit the department. This is intuitive because most often people leave in a group and then they decide to split and move towards residence or towards the office. This phenomenon is also observed in group activity 4(d) which is the activity which is observed when some people are walking left to right simultaneously when some people are walking right to left. Although each of this individual activity may be observed with people either coming from location 5 or walking on the straight road the correlation is observed mainly because people split and walk in both direction when they come from location 5. Group activity 4(c) describes the scene when people leave the department, walk down towards the road and at the same time a set of people are standing and discussing in the area. This is interesting because our model is able to discover the group activity of people standing and meeting while other people are going about walking on the side after leaving the department. Similarly figure 4(b) captures the activity when people are discussing and other people in the group are moving in the right to left direction. In summary, since our method is able to combine individual topics we are not only able to discover group activities as a combination of individual activities but even the individual activities are discovered as fine grained coherent structures of local features which are separated from each other even if present simultaneously in the scene.

In order to measure the performance of our model we compare the activities discovered by our model with the activities discovered by LDA. We trained an LDA model on our video sequence using a Gibbs sampling algorithm. We ran the Gibbs sampler for a total of 3000 iterations and learnt the model by taking 100 samples at an interval of 10 iterations after a burn in period of 2000 iterations. The hyperparameter α was set to 0.1 and the hyperparameter β is set to 0.05. The number of (individual) topics K was set to 12.

Since LDA is not able to distinguish between individual and group activities it produces activities which consists of individual and group activities without labeling which is which. The activities discovered by LDA model is shown in figure 5. Some of the activities discovered by the LDA model correspond to the individual activities discovered by

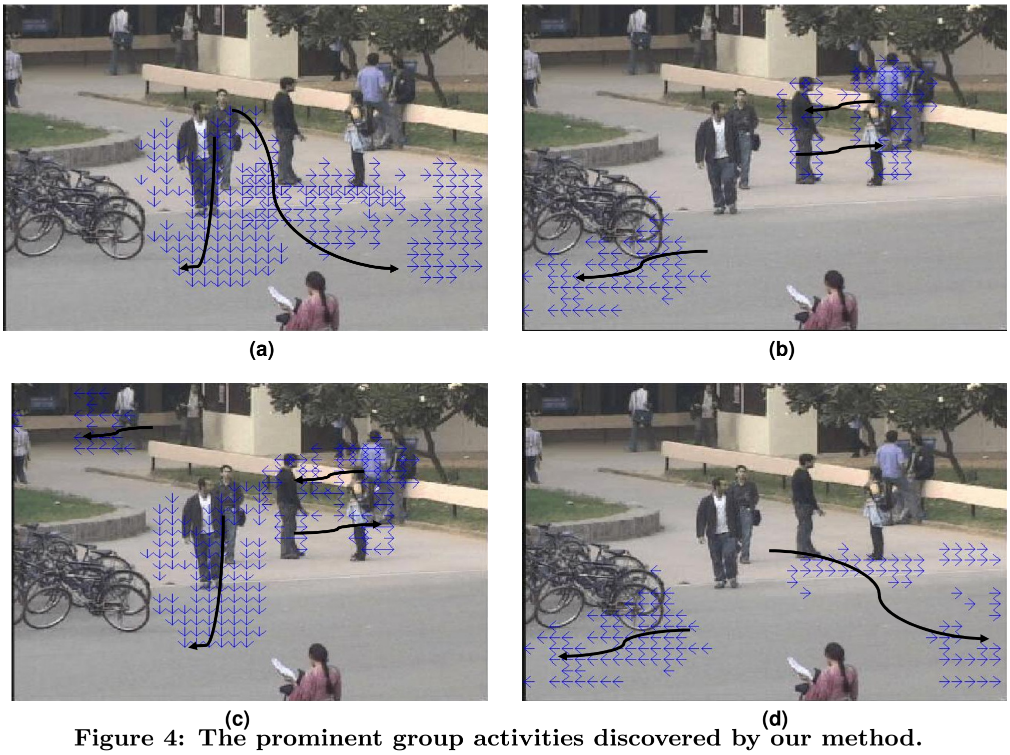

Figure 4: The prominent group activities discovered by our method.

our model. For example the activities discovered in figure 5(b), 5(d) and 5(e) correspond to the individual activities shown in figure 3(b) 3(h) and 3(e). The activity in figure 5(f) can correspond to activity in figure 3(g). However this activity does not preserve the coherency of local features that was observed in our model. Other activities like in figure 5(c) and 5(a) are clearly a combination of different activities. Instead of explaining the observations of these local features by fitting two or more mixture components of individual activities corresponding to a group activity, LDA fits a single component to explain the complete observation. This demonstrates the advantage of our method over LDA and presents the strength of our method for discovering activities in real life video feeds.

We further experiment with the traffic junction data by learning a model on this data to discover individual and group activities. The number of individual activities K_I to be discovered was set to 10 and the number of group activities K_G to be discovered is set to 6. We ran the Gibbs sampling algorithm which took around 60 minutes on a 2.6 GHz machine with 2 GB RAM. Note that our model can be directly applied to this method without requiring any configuration or tuning.

Some of the individual activities discovered by our model is shown in figure 6(a) to 6(d). As can be seen these accurately capture the individual traffic flows along specific paths. We also show a couple of group activities discovered by our model. Group activity in figure 6(e) shows that it is composed of individual activities 6(a) and 6(c). This validates our method because is confirms with the ground truth as both these traffic motions are expected to co-occur because they are governed by the same traffic signal. Similarly group activity 6(f) shows that it is composed of two individual activities 6(b) and 6(d) which is also in confirmation with our ground truth as these two activities co-occur because the traffic motions in these two directions is governed by the same traffic signal.

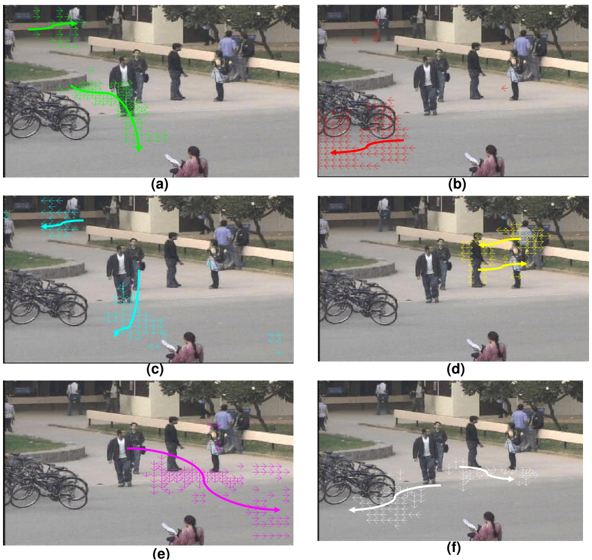

Figure 5: The activities discovered by the LDA without finding the correlations among activities.

6. CONCLUSIONS AND FUTURE WORK

Most of real life surveillance installations monitor public places like train stations, airports, university campus

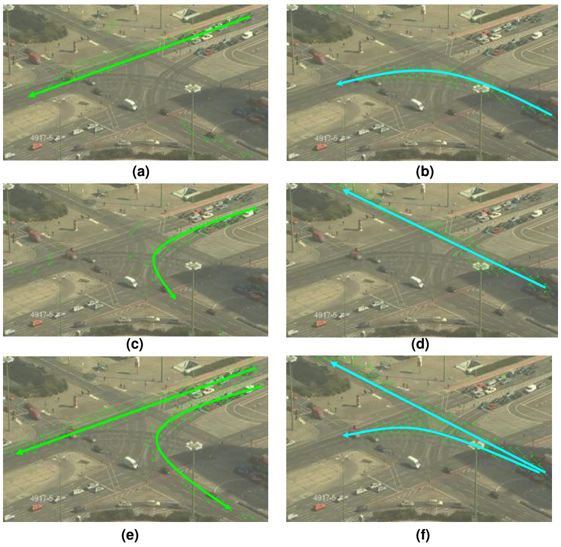

Figure 6: Activities discovered by our method on a video feed of a traffic intersection. a - d are the individual activities discovered. A group activity discovered with component a and b is shown in e. Another group activity with components c and d is shown in f.

etc. Most of the installation monitor crowded scenes with multiple objects behaving not only as individuals but also as a group. We present an unsupervised method that not only discovers the usual activities present in a scene but can also extract the hidden association of these activities among themselves. Discovering this group activities can help in various applications like crowd management, Egress planning, facility management and floor management. Our method does not require any semantic input and can be used in different scenarios with minimal tuning and configuration. In future we plan to extend this work by finding the dynamic group behavior and discovering how these individual groups behave over time. We also plan to discover the time dependency of individual activities and the group activities.

7. 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*, 1997.

[3] A. Choudhary, M. Pal, S. Banerjee, and S. Chaudhury. Unusual activity analysis using video epitomes and plsa. In *ICVGIP*, pages 390–397, 2008.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, pages 65–72, 2005.

[5] T. A. Faruquie, P. K. Kalra, and S. Banerjee. Time based activity inference using latent dirichlet allocation. In *BMVC*, 2009.

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

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

[8] T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In *ICCV*, pages 1165–1172, 2009.

[9] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1450–1464, 2006.

[10] J. Huang. Maximum Likelihood Estimation of Dirichlet Distribution Parameters. *CMU Technical Report*, 2005.

[11] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari. WhatŠs going on? discovering spatio-temporal dependencies in dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2010.

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

[13] J. Niebles, H. Wang, and F. Li. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3), 2008.

[14] G. Ronning. Maximum likelihood estimation of Dirichlet distributions. *Journal of statistical computation and simulation*, 32(4):215–221, 1989.

[15] S. Savarese, A. D. Pozo, J. C. Niebles, and F. F. Li. Spatial temporal correlatons for unsupervised action classification. In *IEEE Workshop on Motion Video Compute*, pages 1–8, 2008.

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

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

[18] X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception by hierarchical bayesian models. In *Proc. CVPR*, 2007.

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

[20] X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *ECCV*, pages 110–123, 2006.

[21] X. Wang, K. Tieu, and W. E. L. Grimson. Correspondence-free activity analysis and scene modeling in multiple camera views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:56–71, 2010.

[22] Y. Wang and G. Mori. Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1762–1774, 2009.

[23] J. Zhang and S. Gong. Action categorization by structural probabilistic latent semantic analysis. *Computer Vision and Image Understanding*, 2010.