Vis Comput (2011) 27:1071–1082
DOI 10.1007/s00371-011-0652-1

ORIGINAL ARTICLE

Latent topic model-based group activity discovery

T.A. Faruquie · S. Banerjee · P. Kalra

Published online: 6 November 2011

Abstract Surveillance videos of public places often consist of group activities composed from multiple co-occurring individual activities. However, latent topic models, such as Latent Dirichlet Allocation (LDA), which have been successfully used to discover individual activities, do not discover group activities. In this paper we propose a method to discover group activities along with individual activities. We use a two layer latent structure where a latent variable is used to discover correlation of individual activities as a group activity using multinomial distribution. Each individual activity is in turn represented as a distribution over local visual features. We use a Gibbs sampling-based algorithm to jointly infer the individual and group 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 discovering activities and the salient correlation amongst them in real life videos of crowded public places.

Keywords Group activity · Surveillance · Topic models · Activity discovery

1 Introduction

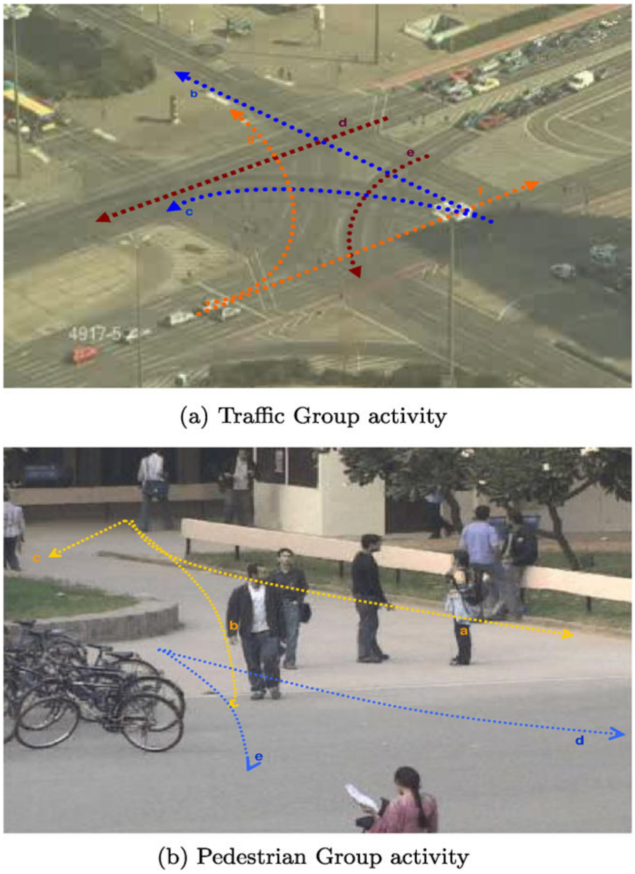
The proliferation of cameras in many public and private places is producing an ever increasing amount of surveillance videos. Though the objective of these camera installations is active scene monitoring, in practice the videos collected are often retroactively utilized on need basis because building an automatic video analysis system requires expensive and laborious manual analysis and labeling. Unsupervised scene understanding and behavior mining is thus important to analyze the increasing volume of video surveillance feeds. In this paper, we address the problem of unsupervised scene understanding and activity analysis. In particular, for a given video we want the system to automatically answer questions like: What are the individual activity patterns? Do these activities occur as a group? How is the scene composed from these individual and group activities?

We introduce the notion of group activity and define it as a set of individual activities that frequently occur together. Example of group activities include the set of individual traffic flows governed by the same traffic signal or the set of pedestrian motion as they disembark from train or emerge from a building. In Figs. 1(a) and (b) we show some of these sets of individual activities that constitute group activities. In Fig. 1(a) the group activities are traffic flows and in Fig. 1(b) they are pedestrian flows as people emerge from a building or a parking lot. We propose to model the occurrence of group activities as latent compositions of individual activities that co-occur with high likelihood. We use a multinomial distribution to measure these latent compositions. This results in a generative process that models each clip as a multinomial distribution of group activities, drawn from a Dirichlet distribution, and each group activity is composed as a multinomial distribution of individual activities drawn from group specific Dirichlet distribution. Individual activities are discovered as multinomial distributions over local features. We use a Gibbs sampling algorithm to jointly learn these distributions, thus discovering the individual activities and the composition of group activities.

T.A. Faruquie (✉) · S. Banerjee · P. Kalra
Dept. of Computer Science and Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
e-mail: tanveer@cse.iitd.ac.in

S. Banerjee
e-mail: suban@cse.iitd.ac.in

P. Kalra
e-mail: pkalra@cse.iitd.ac.in



(a) Traffic Group activity

(b) Pedestrian Group activity

Fig. 1 Example of group activities for traffic scene and a public place. The individual activities are labeled and represented with the same color

Past approaches based on detection and tracking have focused on detecting independent isolated events and do not extend to general settings of complicated multi-activity scenes. To address this shortcoming, topic models such as probabilistic Latent Semantic Analysis (pLSA) [7], Latent Dirichlet Allocation (LDA) [1] and Hierarchical Dirichlet Processes (HDP) [17] were applied to automatically learn activities by correlating local features. The approach taken is to divide the video into clips and then the model automatically correlates the local features present in each clip as latent structures, called topics. Each topic is considered as an activity which is a probability distribution over local features across video clips. Examples of local features that have been used with these models include epitome-based space-time shape features [3], foreground patches [12], spatio-temporal words [14] and optical flow vectors [6, 11]. These local features can be computed reliably even for complicated scenes 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. Though these methods can detect individual activities, they cannot detect the occurrence of activities as a group.

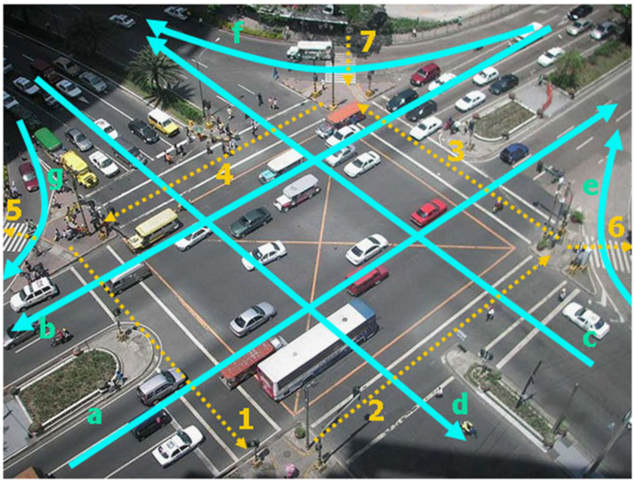


Fig. 2 Scene from a traffic surveillance camera consisting of vehicular and pedestrian activities. Some vehicular activities are marked from a to g and some pedestrian activities are marked from 1 to 7

The group activities can be understood using the following illustrative example. Consider a traffic surveillance system which is monitoring an intersection shown in Fig. 2. This scene depicts some individual vehicular activities (a–g) and pedestrian activities (1–7). 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. For example, the pairs of activities $(c, d, 1)$, $(b, a, 2)$ and (c, e) will almost always co-occur as they are synchronized and do not interfere with each other. Similarly, activities (c, f), (a, e) and (b, g) are less likely to occur as a group as each individual activity represent merging traffic which tend to stall each other for high traffic flows. The chances of co-occurrence of (b, c) and $(a, 3)$ is unlikely as they result in colliding traffic patterns. Pedestrian activities 5, 6 and 7 can occur independently.

Topic models like LDA and pLSA can learn individual activities a–g and 1–7 as multinomial probability distributions over local features and the composition of video clip as proportions of these activities. However, they ignore the fact that some activities co-occur more frequently as a group whereas the co-occurrence of some activities may not make any sense as they might preclude each other (like b and d). Our method addresses this shortcoming by jointly discovering individual and group activities as latent structures. We represent each video clip as a mixture of group activities drawn from a Dirichlet and then individual activities in turn are drawn from the chosen group activities [6]. We further show that our method can not only discover individual activities and complex interaction amongst them, but it is also able to explain novel videos better than other methods based solely on discovering individual activities. We demonstrate this generalization ability of our model by computing the perplexity of the model on previously unseen footage. Our model performs better for different parameter values and varying amount of data. Our method does not need any semantic input, camera calibration information or other prior knowledge. It is also independent of the algorithm used for computing local features.

The rest of the paper is organized as follows. We first present related work in Sect. 2. Subsequently we describe the model used to discover group activities in Sect. 3 and give the Gibbs sampling-based algorithm in Sect. 4. We demonstrate the effectiveness of our model in discovering activities from real life surveillance videos of outdoor scenes and its ability to explain novel video sequences in Sect. 5. Finally, Sect. 6 concludes with discussions and summarizes the findings.

2 Related work

Past work on automatic activity analysis can be divided in two main approaches. The first approach is object centric and often detects and tracks features of interest of an object. These tracks are subsequently used to automatically understand the activities. This approach includes supervised methods to learn specific semantic events and discover activities that are a composition of these events [8] and unsupervised methods that cluster trajectories into activities [10, 19]. This approach also includes other statistical models like Coupled HMMs [2], Bayesian networks [21] and ballistic dynamics [18] that incorporate spatial and temporal features of tracks. 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 is feature centric and consists of methods that directly use low level features instead of tracks as description of activities. Initial methods using this approach avoid the pitfalls of detection and tracking. However, they dealt with only one activity occurring at a given time. To overcome this shortcoming, topic models such as pLSA [4, 14, 16], LDA [22, 24] and HDP [23] have been used to model multiple activities across video clips using local features without detection and tracking. LDA has been used to discover behavior [9, 11], interactions across document clips [20] and time dependency of activities over a period of time [5]. All these methods can discover only individual isolated activities. They do not model group activities in the scene that exist in the form of correlations of individual activities. Our method not only discovers individual activities but also discovers group activities that may occur frequently in the video.

The method in [12] segments the scene into semantic regions and uses one pLSA per region to learn activities of that region. Global behavior pattern across semantic regions is discovered using hierarchical pLSA. This approach requires first to 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 method proposed in [20] assumes 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. Our objective is different where instead of clustering video clips based on activities we cluster activities across video clips using their co-occurrence statistics. Approaches in [9] and [11] model Markov dependency between activities 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 a scene in terms of group activities which are in turn composed of individual activities.

3 Group activity discovery

This section describes in detail the model to discover group activities and individual activities. We begin by briefly describing the local visual features on which the model is trained.

3.1 Local visual features

Automatic analysis of videos monitoring outdoor scenes is challenging because the scene may contain multiple objects in the presence of occlusions and lighting changes. The visual features chosen should be easily and reliably computed in the presence of these difficulties. We compute these features by finding the optical flow vectors across frames and also doing a background subtraction on the video. The complete video is then divided into M video clips containing a fixed number of frames. The camera view (320×280) is divided into fixed size (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. This results in a fixed size vocabulary, $V (= 32 \times 28 \times 4)$, of local visual features to represent a clip. The activities are learned as distributions over the local features described by this vocabulary V.

The above feature extractor is agnostic to the object which generates the local visual features. Hence, for example, both the individual or vehicular activities can be discovered by this method. In general, the temporal coherency of activities generate co-occurring local visual features in a clip. The shape of the object also generates spatially co-occurring local motion features in a clip. The spatial coherency of the features thus captures the approximate shape of the object. When the spatial coherency is observed consistently across multiple clips it is inherently captured statistically in the multinomial distribution for individual activities.

Other methods for computing the local features can also be substituted for or combined with this approach.

3.2 Group activity topic model

Latent Dirichlet Allocation model is used to learn activities present in a video in an unsupervised manner. 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 a priori. For each clip d consisting of N_d local visual features, a multinomial distribution θ_d having K components is sampled from a Dirichlet distribution with parameter α. The multinomial θ_d represent the mixing proportions of activities for that clip. For the tth visual feature slot in the clip, an activity z_{dt} is chosen by making a draw from the multinomial distribution θ_d and then a local visual feature representation $j = w_{dt}$, corresponding to the jth item in vocabulary, is generated by randomly sampling from an activity specific multinomial distribution $\phi_{z_{dt}}$. The distribution $\phi_{z_{dt}}$ is over the vocabulary V and is sampled from a Dirichlet with parameter β. This process uses a two layer latent structure: Activities and local features.

We model the scene as a three layer latent structure representing: group activities, individual activities and local features. We think of a video as composed of M clips consisting of multiple co-occurring individual activities. Individual activities that frequently co-occur as a group are deemed 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 number of group activities K_G and number of individual activities K_I are known a priori. The model utilizes a generative process that explains the observation of local visual features given the individual and group activities. The generative process is random and explains the observations using the joint probability distribution of observed and hidden variables. This generative process is described in detail below:

1. For each K_I individual activities sample a multinomial distribution ϕ having V components from a Dirichlet distribution with parameter β. The multinomial represent the individual activity as a distribution over visual features.
2. For each video clip d having N_d visual features
   (a) Sample a multinomial θ_{Gd} having K_G components, from a Dirichlet with parameter α_G, representing mixture of group activities for d.
   (b) Sample K_G multinomials, $\theta_{Id}^1, \ldots, \theta_{Id}^{K_G}$, having K_I components from each of K_G Dirichlets with parameters $\alpha_I^1, \ldots, \alpha_I^{K_G}$, respectively. These multinomials represent the composition of group activities as a mixture of individual activities. Each group activity is a distribution over individual activities and each individual activity can belong to multiple group activities.
   (c) For each of the tth local feature slot
      - Sample a group activity z_{dgt} from the multinomial θ_{Gd} corresponding to the gth component of θ_{Gd}.
      - Sample an individual activity z_{dit} from the multinomial θ_{Id}^g corresponding to the ith component of θ_{Id}^g.
      - Sample the visual word representation $j = w_{dt}$ from $\phi_{z_{dit}}$.

The local visual features w_d are the visible variables, $\theta_d = \{\theta_{Gd}, \theta_{Id}^1, \ldots, \theta_{Id}^{K_G}\}$ and ϕ are parameters to be learned and $z_d = \{z_{dgt}, z_{dit}\}$ are the hidden variables to be determined. The hyperparameters $\alpha = \{\alpha_G, \alpha_I^1, \ldots, \alpha_I^{K_G}\}$ and β are given. Input parameters K_I and K_G serve to parameterize the model that define the number of multinomials used in the model. Although it is difficult to know a priori the exact value of these parameters, the model is not very sensitive to the values of K_I and K_G as demonstrated by perplexity experiments shown in Figs. 7(a) and (b). If K_I and K_G are higher than the number of inherent activities present in the video, some of the discovered activities are similar or closely identical i.e. their respective multinomials have similar distributions. When K_I and K_G are less than the number of individual activities, some of the activities might not be discovered and some of the activities may be combined. While this might pose difficulties in the interpretation of activities it may still be able to explain the video probabilistically.

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

$$
\begin{aligned}
&p(w_d, z_d, \theta_d|\alpha, \phi) \\
&\quad = p(\theta_{Gd}|\alpha_G) \prod_{g=1}^{K_G} p\big(\theta_{Id}^g|\alpha_I^g\big) \\
&\qquad \times \prod_{t=1}^{N_d} p(z_{dgt}|\theta_{Gd}) p\big(z_{dit}|\theta_{Id}^g\big) p(w_{dt}|\phi_{z_{dit}})
\end{aligned}
$$

Integrating θ_d and z_d we get the marginal probability of d

$$
\begin{aligned}
&p(w_d|\alpha, \phi) \\
&\quad = \int p(\theta_{Gd}|\alpha_G) \prod_{g=1}^{K_G} p\big(\theta_{Id}^g|\alpha_I^g\big)
\end{aligned}
$$

$$\times \prod_{t=1}^{N_d} \sum_{z_{dgt}, z_{dit}} p(z_{dgt}|\theta_{Gd}) p\big(z_{dit}|\theta_{Id}^{g}\big) p(w_{dt}|\phi_{z_{dit}})\, d\theta_d \tag{1}$$

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 the Bayes rule we can write the posterior distribution as

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

Computing the exact marginal likelihood shown in (1) is intractable because of which computing the posterior distribution is also intractable. Hence we use approximate inferencing of the posterior distribution using Gibbs sampling.

4 Parameter estimation

The hidden parameters of our model are the multinomial distributions $\Theta = \{\theta_d, 1 \le d \le M\}$, the individual activity distributions ϕ and the assignment of individual and group activities z_{dit}, z_{dgt} to each observed local feature. Since the posterior distribution is difficult to compute, we use 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. Since the Dirichlet distribution and multinomial distribution belong to exponential family we can integrate out Θ and ϕ to find the conditional probability $p(z_{dgt}, z_{dit}|M, z_{-t}, \alpha, \beta)$ as shown in the Appendix. For the tth feature which has been assigned the visual feature $j \in V$ in document d the probability is given by

$$p(z_{dgt}, z_{dit}|M, z_{-t}, \alpha, \beta)$$
$$\propto \frac{n_{i,j}^{-t} + \beta}{n_i^{-t} + V \cdot \beta} \cdot \frac{n_{d,g,i}^{-t} + \alpha_{Ii}^{g}}{n_{d,g}^{-t} + \sum_{i'}^{K_I} \alpha_{Ii'}^{g}} \cdot \frac{n_{d,g}^{-t} + \alpha_G}{n_d^{-t} + K_G \cdot \alpha_G}$$

Here z_{dgt} and z_{dit} correspond to the global activity, g, and individual activity, i, assignment for a feature slot t assigned a visual feature j. Excluding the current local feature, $n_{d,g}^{-t}$ is the number of times that features in document d are assigned the global activity g and n_d^{-t} is the total number of global activity assignment in d. The number of times features are assigned the individual activity i when they are sampled from the global activity g is $n_{d,g,i}^{-t}$ and $n_{d,g}^{-t}$ is the total number of times global activity g is assigned to features in d. The total number of times that a visual feature j is assigned an individual activity i in complete video is $n_{i,j}^{-t}$ and n_i^{-t} is the number of times features are assigned the individual activity i in the complete video. The α_G, α_{Ii}^{g} and β are the Dirichlet parameters.

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} also have to be updated. The updates can be learned using the method of moments. This method estimates the Dirichlet parameters by finding that density which matches the moments of the data. The first two moments of Dirichlet parameters are given by

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

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

The above two equations can be solved to get

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

By multiplying (4) and (6) α_i^{g} can be estimated. However, since only one component α_i^{g} is used for the estimation we use the method suggested by [15] to use all components. Hence we have

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

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

Equations (4) and (8) can be multiplied to estimate α_i^{g}. This estimation is done after every iteration of Gibbs sampling.

Samples are taken after the burn in period to ensure that their autocorrelation is low. After drawing samples from the posterior distribution $p(\mathbf{z}|\mathbf{w}, \alpha_G, \alpha_I, \beta)$ the parameter values can be estimated. The parameter values from single sample are given by

$$\hat{\phi}_i^j = \frac{n_{i,j} + \beta}{\sum_{j'=1}^{V} n_{i,j'} + V \cdot \beta}$$

$$\hat{\theta}_{Gd}^g = \frac{n_{d,g} + \alpha_G}{\sum_{g'=1}^{K_G} n_{d,g'} + K_G \cdot \alpha_G}$$

$$\hat{\theta}_{Id}^{gi} = \frac{n_{d,g,i} + \alpha_{Ii}^g}{\sum_{i'}^{K_I} (n_{d,g,i'} + \alpha_{Ii'}^g)}$$

Integrating across the full set of samples gives parameter values that are independent of individual topic assignments.

5 Evaluations

This section presents the experimental evaluation of our model. We demonstrate the ability of our model to discover both important individual activities and prominent activity groups using real life videos of public places.

5.1 Setup and dataset

We used two videos from a single view camera monitoring crowded public scenes to evaluate our model. One is from a University Campus and the other from a Traffic Junction.

University campus: This contains 45 mins of video at 24 fps with a frame size of 384 × 288. It monitors a university area having a bicycle stand, book shop, coffee shop and a department in close proximity. Different activities are performed by people in the scene depending on the interaction amongst each other and their use of these services. People enter or exit a department, go toward 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 segmentation and correlation of individual activities difficult.

Traffic junction: This video is a far field video of a busy traffic junction 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 known a priori because the sequence of signal activations and the corresponding traffic motion governed by the signal is known.

The frame for both videos is divided into 10 × 10 cells. After using background subtraction and quantizing the 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 a Gibbs sampler for a total of 4000 iterations and trained 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 trained the model using the visual features extracted from the University Campus video to learn the individual and group activities. The number of individual activities K_I is set to 12 and the number of group activities K_G 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.

After training the model on a training sequence we obtain activities as multinomial distributions over local visual features. These local motion features are ranked according to their probability score for each multinomial distribution. The top local motion features are plotted on the scene which represent the activity modeled by that distribution. We further provide an interpretation of these distributions as annotations over the scene.

Individual activities

The individual activities discovered by our method are shown in Fig. 3. As described earlier the activities in this area are 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 these, people walk across this area and may meet and discuss in the vicinity. The activities discovered by our model are shown in Figs. 3(b)–(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 toward their residences either when coming from department or moving from right to left. Figures 3(c) and (e) are the activities when people leave the department and move toward residence and office, respectively. Figure 3(g) represents the activity when people move toward the office either coming from location 1

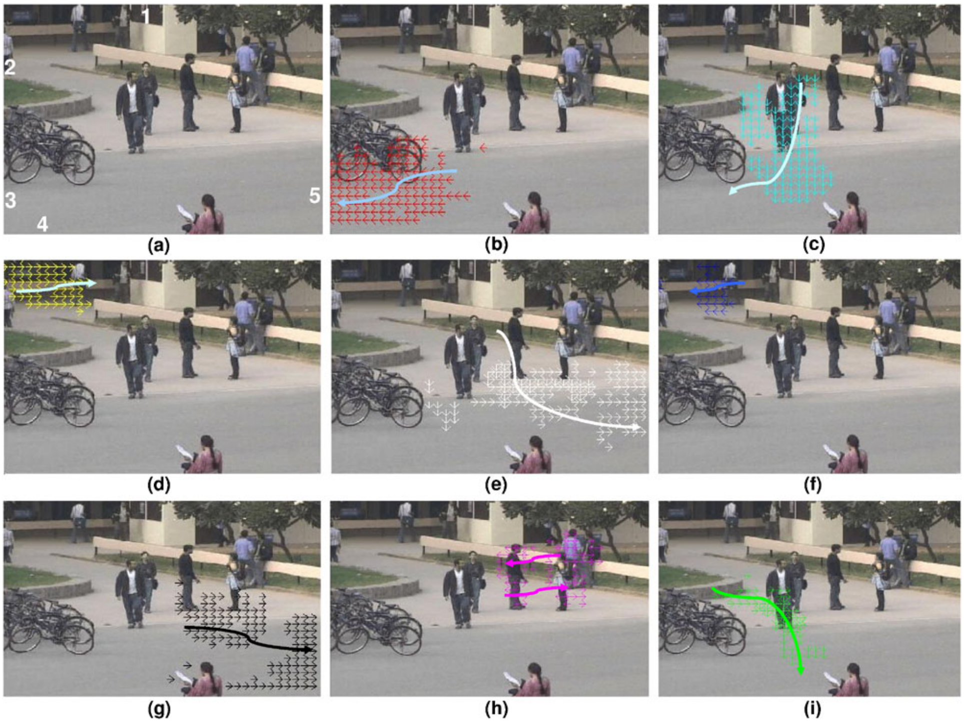


Fig. 3 The activities discovered by our model. (**a**) Represent the location of different services which gives rise to activities in this area. (**b**)–(**i**) Describe the different activities ϕ_z which are discovered by our model. Local visual features which have high $p(w|\phi_z)$ are also plotted

or move straight from left to right. Figures 3(d) and (f) represent the activities when people move toward the bookshop or leave 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. Another interesting activity is shown in Fig. 3(h), 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 Fig. 4. The prominent group activities discovered by our model is shown in Fig. 4. As described in the previous section the group activity is a correlation among individual activities. The correlation among activities is given by the parameter α_I^g which represents the prior on mixing weights of individual activities to be selected to compose a group activity g. In Fig. 4(a) the group activity represents the two different directions which people might take when they exit the department. This is intuitive because usually people left in a group and then split and moved toward residence or office. This phenomenon is also observed in group activity Fig. 4(d), which is the activity which is observed when some people are walking left to right simultaneously while some people are walking right to left. Although each of these individual activities 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 directions when they come from location 5. Group activity Fig. 4(c) describes the scene when people leave the department, walk down toward 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 Fig. 4(b) captures the activity when people are discussing and other people in the group are moving from right to left. In summary, since our method is able to combine individual topics we not only 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 trained 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.

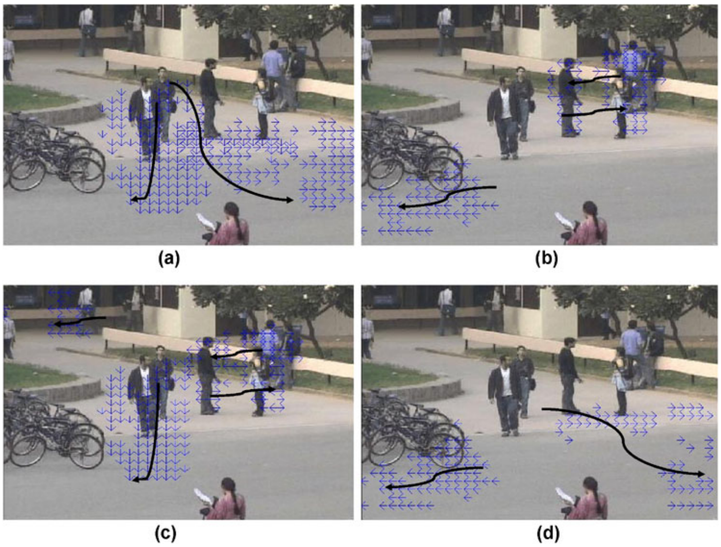


Fig. 4 The prominent group activities discovered by our method

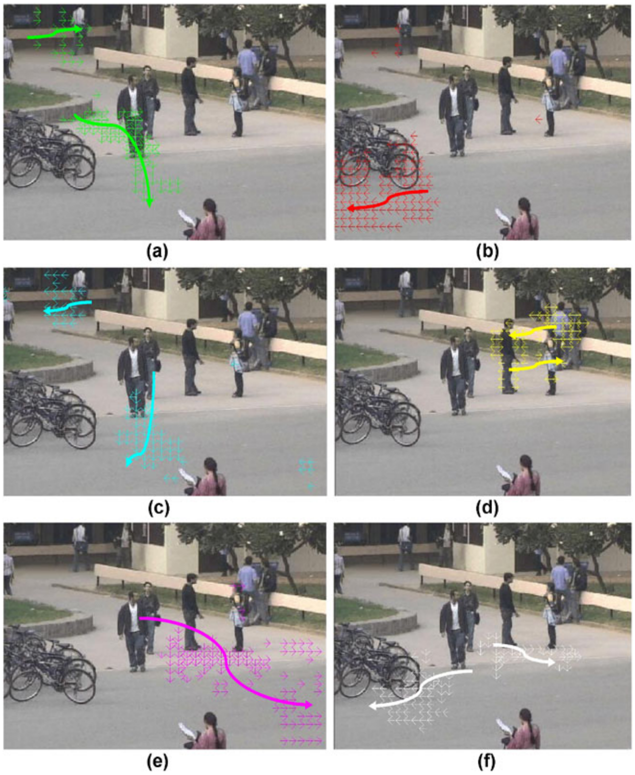


Fig. 5 The activities discovered by the LDA which does not find the correlations among individual activities

Since LDA is not able to distinguish between individual and group activities it produces activities which consist of individual and group activities without labeling which is which. The activities discovered by LDA model are shown in Fig. 5. Some of the activities discovered by the LDA model correspond to the individual activities discovered by our model. For example, the activities discovered in Figs. 5(b), (d) and (e) correspond to the individual activities shown in Figs. 3(b), (h), and (e). The activity in Fig. 5(f) can correspond to activity in Fig. 3(g). However, this activity does not preserve the coherency of local features that was observed in our model. Other activities like those in Figs. 5(c) and (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 training a model on these 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 are shown in Figs. 6(a) to (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 Fig. 6(e) shows that it is composed of individual activities Figs. 6(a) and (c). This validates our method because this 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 Fig. 6(f) shows that it is composed of two individual activities Figs. 6(b) and (d), which is also in confirmation with our ground truth as these two activities co-occur, because the traffic motions in these two directions are governed by the same traffic signal.

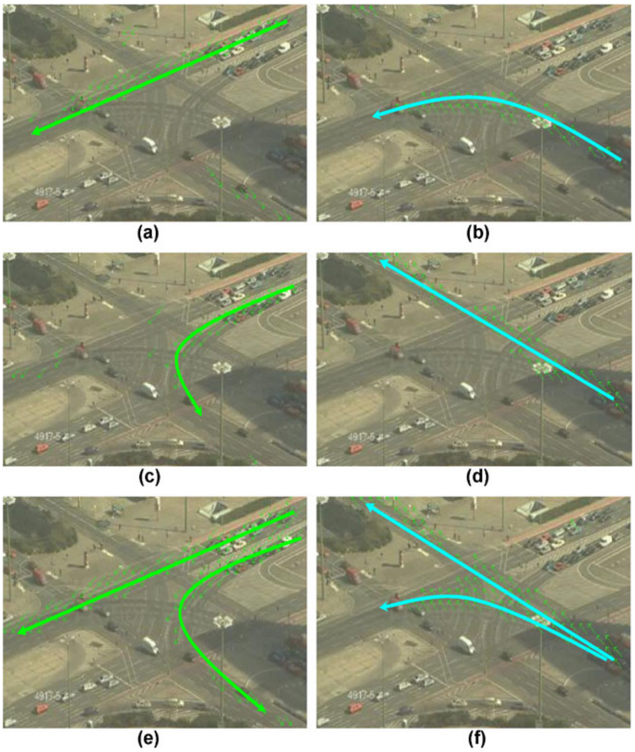


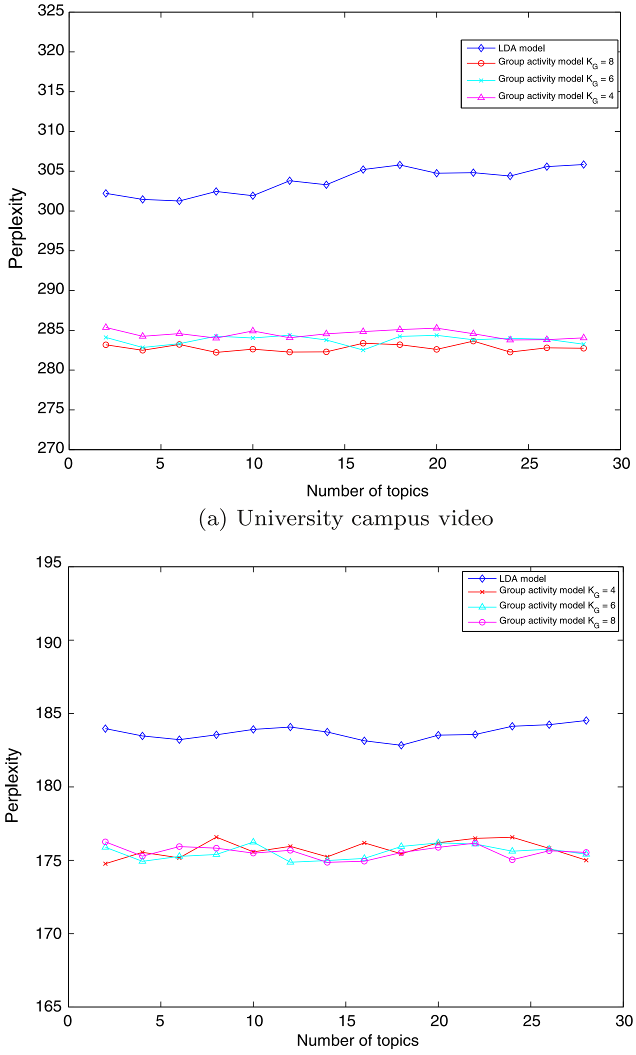
Fig. 6 Activities discovered by our method on a video feed of a traffic intersection. Individual discovered activities are (**a**)–(**d**), while (**e**) is a group activity with components (**a**) and (**b**), and (**f**) with components (**c**) and (**d**)

5.3 Perplexity comparison

We also provide a quantitative comparison of the group activity model with the LDA model. The objective is to evaluate the generalization achieved when group activities are learned compared to learning individual activities using the LDA model. Since the videos are unlabeled the task is to achieve high likelihood on a held out test set. In particular, we compute perplexity on held out test data for both videos to compare the model. Perplexity is the number of bits required to encode the data and is inversely proportional to the log likelihood of the data. The lower the perplexity score, the better is the generalization performance. For a held out test set $\mathbf{M}_{\text{test}}$ containing M documents the perplexity is given by

$$perplexity(\mathbf{M}_{\text{test}}) = e^{-\frac{\sum_{d=1}^{M}\log(w_d|\phi)}{\sum_{d=1}^{M}N_d}}$$

Here, w_d and N_d are the visual features and the number of words, respectively, in video clip d of the test set $\mathbf{M}_{\text{test}}$. Given the complete video sequence, we randomly sample video clips to keep aside 20% of data for test purpose. The remaining 80% of the data are used for training purposes. Hence, the test data come from the same underlying mechanism used to produce the training data, including the observed scene, camera view and the calibration used to generate the complete video. The activities ϕ are first learned on the training data given the hyperparameters. Once ϕ are learned they are fixed and perplexity is computed on the video documents in the test data.

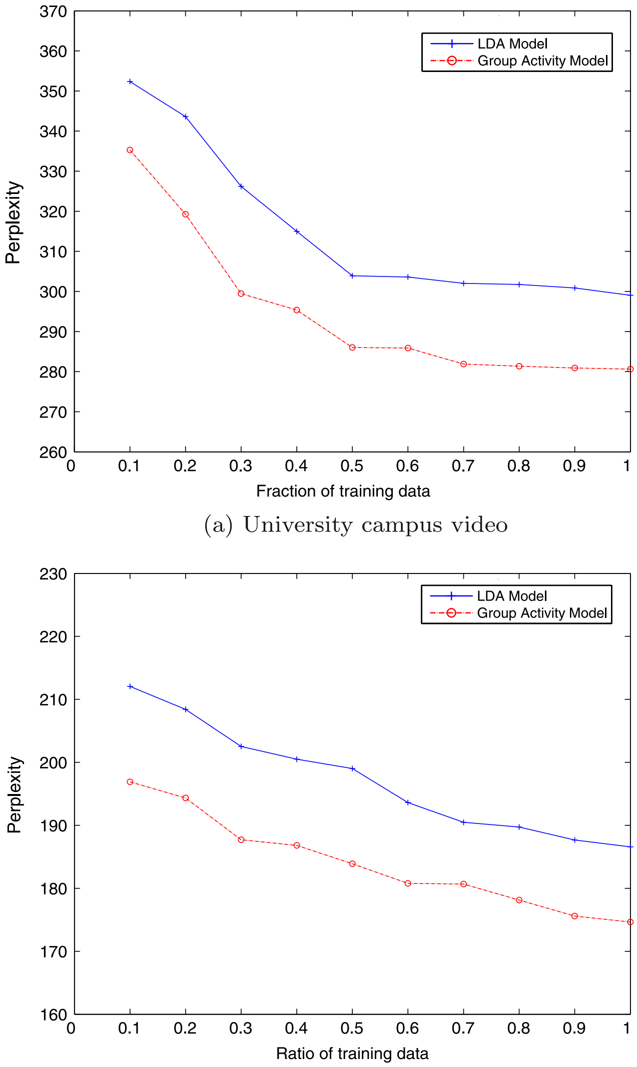


(a) University campus video

(b) Traffic junction video

Fig. 7 Perplexity comparison with LDA for different numbers of individual activities and different numbers of group activities

We plot the perplexity of the group activity model and LDA model by varying K_I and K, respectively, the number of individual activities learned by the model. The perplexity for the university campus video and traffic junction video is shown in Figs. 7(a) and (b), respectively. The perplexity of the group activity model is consistently lower than the perplexity of LDA for different numbers of individual activities. These figures also show that the performance is maintained even when the number of group activities discovered are varied because the perplexity of group activity model for different numbers of group activities is lower than the LDA model. These graphs show that group activities, along



(a) University campus video

(b) Traffic junction video

Fig. 8 Perplexity comparison with LDA for different amounts of training data

with individual activities are able to model the activities in a scene better than just using the individual activities.

In Figs. 8(a) and (b), we present the perplexities for the university campus video and the traffic junction video, respectively, for different amount of video used for training the model. In these plots the number of individual activities was set to 12 and 10 for the university campus video and the traffic junction video, respectively, including the LDA model. The number of group activities K_G was set to 6 for both the videos. From both these graphs we see that learning group activities along with individual activities give better generalization performance, which is maintained even when the amount of available training data is varied.

For a novel video the likelihood of the video given our model is given by $\sum_{d=1}^{M} \log(w_d|\phi)$. This likelihood can be used to find the presence of abnormal activities when the model is trained on video containing usual activities. Any unusual activity present in the video will result in a lower likelihood for that video when evaluated using the model. This unusualness can be caused by abnormal individual activity or abnormal group activity or both.

6 Conclusions and future work

Most of real life surveillance installations monitor public places like train stations, airports, university campus 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. We also demonstrate that discovering group activities results in better explaining previously unseen footage using perplexity measures. Discovering the 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 our model to utilize non-parametric approaches, such as a Dirichlet process mixture model [13], to automatically determine the number of activities and group activities in a data driven fashion. In order to distinguish between activities by different objects we plan to construct a feature vocabulary that can code position and motion as well as shape and appearance information. We also plan to extend this work by finding the dynamic group behavior and discovering how these individual groups behave over time. We plan to discover the time dependency of individual activities and the group activities.

Appendix

Let $\mathbf{w}$ be the local visual features in the whole scene $\mathbf{M}$ consisting of M video clips. The individual activity and global activity assignments $\mathbf{z} = \{\mathbf{z_I}, \mathbf{z_G}\}$ are drawn from multinomial $\theta_I = \{\theta_I^1, \theta_I^2, \ldots, \theta_I^{K_G}\}$ and θ_G, which in turn are sampled from Dirichlets with parameters $\alpha_I = \{\alpha_I^1, \alpha_I^2, \ldots, \alpha_I^{K_G}\}$ and α_G, respectively. The topics are drawn from multinomials ϕ which are sampled from Dirichlet with parameter β. The joint probability of the observed variables, the hidden variables and the model parameters, given the hyperparameters, are described by

$$p(\mathbf{w}, \mathbf{z_G}, \mathbf{z_I}, \theta_G, \theta_I, \phi|\alpha_G, \alpha_I, \beta) =$$

$$p(\phi|\beta)p(\theta_I|\alpha_I)p(\theta_G|\alpha_G)p(\mathbf{w}|\phi, \mathbf{z_I})p(z_I|\theta_I, \mathbf{z_G})p(\mathbf{z_G}|\theta_G) \quad (9)$$

This can be expanded as

$$= \prod_{i=1}^{K_I} p(\phi_i|\beta) \prod_{d=1}^{M} p(\theta_{Gd}|\alpha_G) \prod_{g=1}^{K_G} p\big(\theta_{Id}^g|\alpha_I^g\big)$$

$$\times \prod_{j=1}^{N_d} p(z_{dgj}|\theta_{Gd}) p\big(z_{dij}|\theta_{Id}^g\big) p(w_{dj}|\phi_{z_{dij}}) \qquad (10)$$

We have used the parametric representation for Dirichlet and multinomial distributions

$$= \left[\prod_{i=1}^{K_I} \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \prod_{j=1}^{V} \phi_{i,j}^{\beta-1}\right]\left[\prod_{d=1}^{M} \frac{\Gamma(K_G\alpha_G)}{\Gamma(\alpha_G)^{K_G}} \prod_{g=1}^{K_G} (\theta_G)_{d,g}^{\alpha_G-1}\right]$$

$$\times \left[\prod_{i=1}^{K_I}\prod_{j=1}^{V} \phi_{i,j}^{n_{i,j}}\right]\left[\prod_{d=1}^{M}\prod_{g=1}^{K_G} \frac{\Gamma(\prod_{i'=1}^{K_I}\alpha_{Ii'}^g)}{\prod_{i'=1}^{K_I}\Gamma(\alpha_{Ii'}^g)} \prod_{i=1}^{K_I} \big(\theta_I^g\big)_{d,g,i}^{\alpha_{Ii}^g-1}\right]$$

$$\times \left[\prod_{d=1}^{M}\prod_{g=1}^{K_G} (\theta_G)_{d,g}^{n_{d,g}}\right]\left[\prod_{d=1}^{M}\prod_{g=1}^{K_G}\prod_{i=1}^{K_I} \big(\theta_I^g\big)_{d,i}^{n_{d,i}}\right] \qquad (11)$$

$$= C \cdot \left[\prod_{i=1}^{K_I}\prod_{j=1}^{V} \phi_{i,j}^{n_{i,j}+\beta-1}\right]\left[\prod_{d=1}^{M}\prod_{g=1}^{K_G} (\theta_G)_{d,g}^{n_{d,g}+\alpha_G-1}\right]$$

$$\times \left[\prod_{d=1}^{M}\prod_{g=1}^{K_G}\prod_{i=1}^{K_I} \big(\theta_I^g\big)_{d,i}^{n_{d,g,i}+\alpha_{Ii}^g-1}\right] \qquad (12)$$

where C is a constant and we have used conjugacy between Dirichlet and multinomial distribution. Integrating θ_G, θ_I and ϕ we get

$$p(\mathbf{w}, \mathbf{z}_G, \mathbf{z}_I|\alpha_G, \alpha_I, \beta)$$

$$\propto \iiint \left[\prod_{j=1}^{V}\prod_{i=1}^{K_I} \phi_{i,j}^{n_{i,j}+\beta-1}\right]\left[\prod_{d=1}^{M}\prod_{g=1}^{K_G} (\theta_G)_{d,g}^{n_{d,g}+\alpha_G-1}\right]$$

$$\times \left[\prod_{d=1}^{M}\prod_{g=1}^{K_G}\prod_{i=1}^{K_I} \big(\theta_I^g\big)_{d,i}^{n_{d,g,i}+\alpha_{Ii}^g-1}\right] d\phi\, d\theta_I\, d\theta_G \qquad (13)$$

Since integration of a product of independent variables can be integrated separately we can write the above equation as

$$= \prod_{i=1}^{K_I} \int \prod_{j=1}^{V} \phi_{i,j}^{n_{i,j}+\beta-1}\, d\phi \prod_{d=1}^{M} \int \prod_{g=1}^{K_G} (\theta_G)_{d,g}^{n_{d,g}+\alpha_G-1}\, d\theta_G$$

$$\times \prod_{d=1}^{M}\prod_{g=1}^{K_G} \int \prod_{i=1}^{K_I} \big(\theta_I^g\big)_{d,i}^{n_{d,g,i}+\alpha_{Ii}^g-1}\, d\theta_I^g \qquad (14)$$

Using the fact that integration over Dirichlet distribution equals to 1 we have

$$= \prod_{i=1}^{K_I} \frac{\prod_{j=1}^{V}\Gamma(n_{i,j}+\beta)}{\Gamma(n_i+V\beta)} \cdot \prod_{d=1}^{M} \frac{\prod_{g=1}^{K_G}\Gamma(n_{d,g}+\alpha_G)}{\Gamma(n_d+K_G\alpha_G)}$$

$$\times \prod_{d=1}^{M}\prod_{g=1}^{K_G} \frac{\prod_{i=1}^{K_I}\Gamma(n_{d,g,i}+\alpha_{Ii}^g)}{\Gamma(n_{d,g}+\sum_{i'}^{K_G}\alpha_{Ii'}^g)} \qquad (15)$$

Let z_{dgt} and z_{dit} be the global and local activity assignments to the current local feature t belonging to document d with local feature assignment $j \in V$. Let $\mathbf{z}_{-t}$ be the assignment to all other local features excluding this local feature.

According to the Bayes rule and using the fact that z_t depends only on w_t we have

$$p(z_{dgt}, z_{dit}|\mathbf{z}_{-t}, \mathbf{w}, \alpha_G, \alpha_I, \beta)$$

$$\propto \frac{p(\mathbf{w}, \mathbf{z}|\alpha_G, \alpha_I, \beta)}{p(\mathbf{w}, \mathbf{z}_{-t}|\alpha_G, \alpha_I, \beta)}$$

$$= \frac{p(\mathbf{w}, \mathbf{z}|\alpha_G, \alpha_I, \beta)}{p(\mathbf{w}_{-t}, \mathbf{z}_{-t}|\alpha_G, \alpha_I, \beta)}$$

The numerator in the above equation is the same as (15). The denominator is also almost the same but differs from the numerator because it does not include the count for the assigned global and individual activities to the tth feature. In fact the count in the numerator for the gth global and ith individual activity assignments are related to the counts for the global and individual activities in denominator by a difference of one.

Accounting for this difference in (15), most of the factors cancel out except the factors associated with the tth feature. Hence

$$p(z_{dgt}, z_{dit}|\mathbf{z}_{-t}, \mathbf{w}, \alpha_G, \alpha_I, \beta)$$

$$\propto \frac{\frac{\prod_{j=1}^{V}\Gamma(n_{i,j}^{-t}+1+\beta)}{\Gamma(n_i^{-t}+1+V\beta)}}{\frac{\prod_{j=1}^{V}\Gamma(n_{i,j}^{-t}+\beta)}{\Gamma(n_i^{-t}+V\beta)}} \cdot \frac{\frac{\prod_{g=1}^{K_G}\Gamma(n_{d,g}^{-t}+1+\alpha_G)}{\Gamma(n_d^{-t}+1+K_G\alpha_G)}}{\frac{\prod_{g=1}^{K_G}\Gamma(n_{d,g}^{-t}+\alpha_G)}{\Gamma(n_d^{-t}+K_G\alpha_G)}}$$

$$\cdot \frac{\frac{\prod_{i=1}^{K_I}\Gamma(n_{d,g,i}^{-t}+1+\alpha_{Ii}^g)}{\Gamma(n_{d,g}^{-t}+1+\sum_{i'}^{K_I}\alpha_{Ii'}^g)}}{\frac{\prod_{i=1}^{K_I}\Gamma(n_{d,g,i}^{-t}+\alpha_{Ii}^g)}{\Gamma(n_{d,g}^{-t}+\sum_{i'}^{K_I}\alpha_{Ii'}^g)}} \qquad (16)$$

Using the factorial representation of the Gamma function, $\Gamma(y) = (y-1)\Gamma(y-1)$, we get

$$p(z_{dgt}, z_{dit}|\mathbf{M}, \mathbf{z}_{-t}, \alpha, \beta)$$

$$\propto \frac{n_{i,j}^{-t}+\beta}{n_i^{-t}+V\cdot\beta} \cdot \frac{n_{d,g}^{-t}+\alpha_G}{n_d^{-t}+K_G\cdot\alpha_G} \cdot \frac{n_{d,g,i}^{-t}+\alpha_{Ii}^g}{n_{d,g}^{-t}+\sum_{i'}^{K_I}\alpha_{Ii'}^g}$$

This gives the update rule for drawing samples for global and local activity assignments.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I., Lafferty, J.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Brand, M., Oliver, N., Pentland, A.: Coupled hidden Markov models for complex action recognition. In: CVPR (1997)
3. Choudhary, A., Pal, M., Banerjee, S., Chaudhury, S.: Unusual activity analysis using video epitomes and PLSA. In: ICVGIP, pp. 390–397 (2008)
4. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS, pp. 65–72 (2005)
5. Faruquie, T.A., Kalra, P.K., Banerjee, S.: Time based activity inference using latent Dirichlet allocation. In: BMVC (2009)
6. Faruquie, T.A., Banerjee, S., Kalra, P.K.: Unsupervised discovery of activity correlations using latent topic models. In: Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing, pp. 25–32 (2010)
7. Hoffmann, T.: Probabilistic latent semantic analysis. In: SIGIR, pp. 50–57 (1999)
8. Hongeng, S., Nevatia, R.: Multi-agent event recognition. In: ICCV, pp. 84–93 (2001)
9. Hospedales, T., Gong, S., Xiang, T.: A Markov clustering topic model for mining behaviour in video. In: ICCV, pp. 1165–1172 (2009)
10. Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., Maybank, S.: A system for learning statistical motion patterns. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1450–1464 (2006)
11. Kuettel, D., Breitenstein, M.D., Gool, L.V., Ferrari, V.: What's going on? discovering spatio-temporal dependencies in dynamic scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, June (2010)
12. Li, J., Gong, S., Xiang, T.: Global behaviour inference using probabilistic latent semantic analysis. In: BMVC (2008)
13. Neal, R.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. 249–265 (2000)
14. Niebles, J., Wang, H., Li, F.: Unsupervised learning of human action categories using spatial-temporal words. Int. J. Comput. Vis. **79**(3) (2008)
15. Ronning, G.: Maximum likelihood estimation of Dirichlet distributions. J. Stat. Comput. Simul. **32**(4), 215–221 (1989)
16. Savarese, S., Pozo, A.D., Niebles, J.C., Li, F.F.: Spatial temporal correlations for unsupervised action classification. In: IEEE Workshop on Motion Video Compute, pp. 1–8 (2008)
17. Teh, Y., Jordon, M., Beal, M., Blei, D.: Hierarchical Dirichlet process. J. Am. Stat. Assoc. 1566–1581 (2006)
18. Vitaladevuni, S., Kellokumpu, V., Davis, L.: Action recognition using ballistic dynamics. In: CVPR (2008)
19. Wang, X., Tieu, K., Grimson, E.: Learning semantic scene models by trajectory analysis. In: ECCV, pp. 110–123 (2006)
20. Wang, X., Ma, X., Grimson, E.: Unsupervised activity perception by hierarchical Bayesian models. In: Proc. CVPR (2007)
21. Wang, X., Ma, X., Grimson, W.: Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 539–555 (2009)
22. Wang, X., Tieu, K., Grimson, W.E.L.: Correspondence-free activity analysis and scene modeling in multiple camera views. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 56–71 (2010)
23. Wang, Y., Mori, G.: Human action recognition by semilatent topic models. IEEE Trans. Pattern Anal. Mach. Intell. 1762–1774 (2009)
24. Zhang, J., Gong, S.: Action categorization by structural probabilistic latent semantic analysis. Comput. Vis. Image Understan. (2010)

T.A. Faruquie received the B.E. degree in electronics and telecommunications from Rani Durgawati University, India and an M. Tech. degree in electrical engineering from the Indian Institute of Technology, Mumbai, India. He is pursuing his Ph.D. in department of computer science and engineering, Indian Institute of Technology, New Delhi, India. His research interests include computer vision, statistical inference, unstructured information processing and speech processing.

S. Banerjee obtained his Ph.D. from the Indian Institute of Science in 1989 and is a Professor at the department of Computer Science and Engineering, Indian Institute of Technology, Delhi, India. He currently holds the Microsoft Chair Professorship. His broad areas of research include Computer Vision, Robotics, Real-time systems, Image processing and Pattern recognition.

P. Kalra is a professor in the Department of Computer Science and Engineering at the Indian Institute of Technology (IIT), Delhi. He is with IIT Delhi since December 1997. Before joining IIT Delhi, he was at University of Geneva (Switzerland). He did B.Sc. Mechanical Engineering from (Dayalbagh) Agra University, India, M.Tech. in Industrial Engineering from IIT Delhi, M.S. in Computer Science from University of New Brunswick, Canada and Ph.D. in Computer Science from Swiss Federal Institute of Technology- Lausanne (EPFL), Switzerland. His research interests include computer vision-based modeling and rendering, 3D visualization and animation, and image/video super-resolution. He has published more than 50 papers in reputed international journals and conferences.