Unsupervised Activity Perception by Hierarchical Bayesian Models

Xiaogang Wang Xiaoxu Ma Eric Grimson
Computer Science and Artificial Intelligence Lab
Massachusetts Tnstitute of Technology, Cambridge, MA, 02139, USA
xgwang@csail.mit.edu, xiaoxuma@mit.edu, welg@csail.mit.edu

Abstract

We propose a novel unsupervised learning framework for activity perception. To understand activities in complicated scenes from visual data, we propose a hierarchical Bayesian model to connect three elements: low-level visual features, simple "atomic" activities, and multi-agent interactions. Atomic activities are modeled as distributions over low-level visual features, and interactions are modeled as distributions over atomic activities. Our models improve existing language models such as Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP) by modeling interactions without supervision. Our data sets are challenging video sequences from crowded traffic scenes with many kinds of activities co-occurring. Our approach provides a summary of typical atomic activities and interactions in the scene. Unusual activities and interactions are found, with natural probabilistic explanations. Our method supports flexible high-level queries on activities and interactions using atomic activities as components.

1. Introduction

The goal of this work is to understand activities and interactions in a complicated scene, e.g. a crowded traffic scene (see Figure 1), a busy train station or a shopping mall. In such scenes individual objects are often not easily tracked because of frequent occlusions among objects, and many different types of activities often occur simultaneously. Nonetheless, we expect a visual surveillance system to: (1) find typical single-agent activities (e.g. car makes a U-turn) and multi-agent interactions (e.g. vehicles stop waiting for pedestrians to cross the street) in this scene, and provide a summary; (2) label short video clips in a long sequence by interaction, and localize different activities involved in an interaction; (3) show abnormal activities, e.g. pedestrians cross the road outside the crosswalk; and abnormal interactions, e.g. jay-walking (pedestrians cross the road while vehicles pass by); (4) support queries about an interaction that has not yet been discovered by the system. Ideally, a system would learn models of the scene to answer such questions in an unsupervised way.

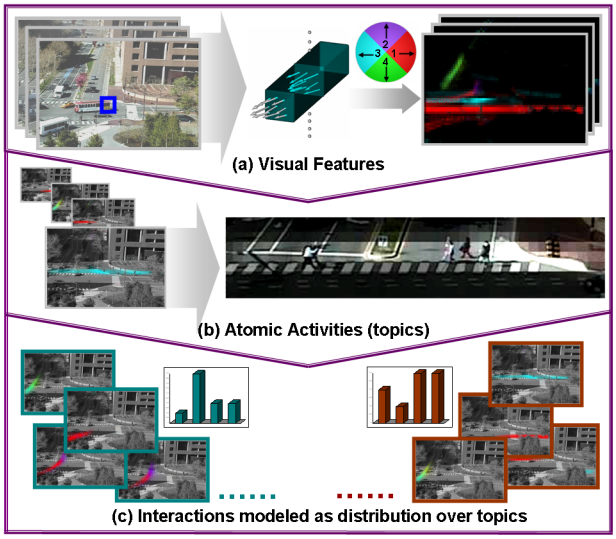


Figure 1. Our approach connects: low-level visual features, atomic activities and interactions. (a) The video sequence is divided into short clips as documents. In each clip, local motions are quantized into visual words based on location and motion direction. The four quantized directions are represented by colors. Each video clip has a distribution over visual words. (b) Atomic activities (e.g. pedestrians cross the road) are discovered and modeled as distributions over visual words. (c) Each video clip is labeled by type of interaction, modeled as a distribution over atomic activities.

To answer these challenges for visual surveillance systems, we must determine: how to compute low-level visual features, and how to model activities and interactions. Our approach is shown in Figure 1. We compute local motion as low-level visual feature, avoiding difficult tracking problems in crowded scenes. We do not adopt global motion features ([15, 14]), because in our case multiple activities occur simultaneously and we want to separate single-agent activities from interactions. Word-document analysis is then performed by quantizing local motion into visual words and dividing the long video sequence into short clips as documents. We assume that visual words caused by the same

1-4244-1180-7/07/$25.00 ©2007 IEEE

kind of atomic activities often co-exist in video clips (documents) and that interaction is a combination of atomic activities occuring in the same clip. Given this problem structure, we employ a hierarchical Bayesian approach, in which atomic activities are modeled as distributions over low-level visual features, and interactions are modeled as distributions over atomic activities. Under this model, surveillance tasks like clustering video clips and abnormality detection have a nice probabilistic explanation. Because our data is hierarchical, a hierarchical model can have enough parameters to fit the data well while avoiding overfitting problems, since it is able to use a population distribution to structure some dependence into the parameters [2].

Hierarchical Bayesian models for word-document analysis include *LDA* [1] and *HDP* [12]. In these models, words often co-existing in the same documents are clustered into the same topic. *HDP* is nonparametric and can automically decide the number of topics while *LDA* requires knowing that in advance. Directly applying these models to our problem, atomic activities (corresponding to topics) can be discovered and modeled, however modeling interactions is not straightforward. Although *LDA* and *HDP* allow inclusion of more hierarchical levels, they require manually labeling documents into groups. For example, [12] modeled multiple corpora but required knowing to which corpus each document belonged; [5] used *LDA* for scene categorization, but had to label each image in the training set into different categories. We improve *LDA* and *HDP* approaches by simultaneously grouping words into topics and documents into clusters in an unsupervised way. In the case of visual surveillance, this means we can learn atomic activities as well as interactions among activities.

1.1. Related Work

Most existing approaches to activity analysis fall into two categories. In the first, objects of interest are detected, tracked, and classified into different classes. The object tracks are used to model activities [8, 11, 13]. These approaches fail when object detection, tracking, and/or recognition do not work well, especially in crowded scenes. Some systems model primitive events, such as "move, stop, enter-area, turn-left", and use these primitives as components to model complicated activities and interactions [4, 3]. These primitive events are learned from labeled training examples, or their parameters are manually specified. When switching to a new scene, new training samples must be labeled and parameters must be tuned or re-learned.

The second approach [15, 10] directly uses motion feature vectors instead of tracks to describe video clips. Without object detection and tracking, a particular activity can not be separated from other activities simultaneously occuring in the same clip, as is common in crowded scenes. These approaches treat a video clip as an integral entity and flag the whole clip as normal or abnormal. They are often applied to simple data sets where there is only one kind of activity in a video clip. It is difficult for them to model both single-agent activities and multi-agent interactions.

In computer vision, hierarchical Bayesian models have been applied to scene categorization [5], object recognition [9], and human action recognition [7]. Both [5] and [7] are supervised learning frameworks in the sense that they need to manually label the documents. The video clip in [7] usually contains a single activity and [7] did not model interactions among multiple objects. [9], which directly applied an LDA model, was an unspervised framework assuming a document contains only one major topic. This does not work for our problem where each document typically contains several topics. It did not model interactions either.

Our approach avoids tracking in crowded scenes, using only local motion as features. It can separate co-occuring activities in the video clip by modeling activities and interactions. The whole learning procedure is unsupervised without manual labeling of video clips or local motions.

2. Low-Level Visual Features

Our data set is a challenging far-field traffic scene (Figure 1) video sequence lasting 1.5 hours, recorded by a fixed camera. There are myriads of activities and interactions in the video data. It also involves many challenging problems, such as lighting changes, occlusions, a variety of object types, object view changes and environmental effects.

We compute local motion as our low-level feature. Moving pixels are detected in each frame as follows. We compute the intensity difference between two successive frames, on a pixel basis. If the difference at a pixel is above a threshold, the pixel is detected as a moving pixel. The motion direction at each moving pixel is obtained by computing optical flow [6]. The moving pixels are quantized according to a codebook, as follows. Each moving pixel has two features: position and direction of motion. To quantize position, the scene(480×720) is divided into cells of size 10 by 10. The motion of a moving pixel is quantized into four directions as shown in Figure 1(a). Hence the size of the codebook is $48 \times 72 \times 4$, and thus each detected moving pixel is assigned a word based on rough position and motion direction. The whole video sequence is divided into 540 clips, each 10 seconds in length. In our framework, video clips are treated as documents and moving pixels are treated as words for word-document analysis as described in Section 3.

3. Hierarchical Bayesian Models

LDA and *HDP* are hierarchical Bayesian models for language processing. In these models, words that often co-exist in the same documents are clustered into the same

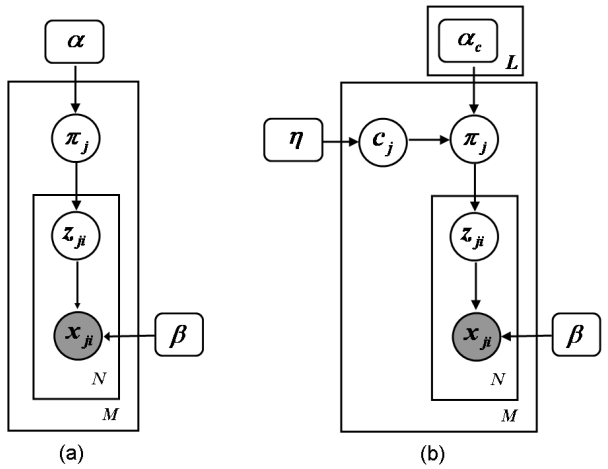


Figure 2. (a) *LDA* model proposed in [1]; (b) our *LDA* model.

topic. We extend these models by enabling clustering of both documents and words, thus finding co-occuring words (topics) and co-occuring topics (interactions). For far-field surveillance videos, words are quantized local motion of pixels; moving pixels that tend to co-occur in clips (or documents) are modeled as topics. Our goal is to infer the set of activities (or topics) from video by learning the distributions of features that co-occur, and to learn distributions of activities that co-occur, thus finding interactions.

3.1. LDA

Figure 2(a) shows the *LDA* model of [1]. Suppose the corpus has M documents. Each document is modeled as a mixture of K topics, where K is assumed known. Each topic k is modeled as a multinomial distribution over a word vocabulary given by $\beta = \{\beta_k\}$. α is a Dirichlet prior on the corpus. For each document j, a parameter π_j of the multinomial distribution is drawn from Dirichlet distribution $Dir(\pi_j|\alpha)$. For each word i in document j, a topic z_{ji} is drawn with probability π_{jk}, and word x_{ji} is drawn from a multinomial distribution given by $\beta_{z_{ji}}$. π_j and z_{ji} are hidden variables. α and β are hyperparameters to be optimized. Given α and β, the joint distribution of topic mixture π_j, topics $\mathbf{z}_j = \{z_{ji}\}$, and words $\mathbf{x}_j = \{x_{ji}\}$ is:

$$p(\mathbf{x}_j, \mathbf{z}_j, \pi_j|\alpha, \beta) = p(\pi_j|\alpha)\prod_{i=1}^{N} p(z_{ji}|\pi_j)p(x_{ji}|z_{ji}, \beta)$$
$$= \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\pi_{j1}^{\alpha_1-1}\cdots\pi_{jK}^{\alpha_K-1}\prod_{i=1}^{N}\pi_{jz_{ji}}\beta_{z_{ji}x_{ji}} \quad (1)$$

where N is the number of words in document j. Unfortunately, the marginal likelihood $p(\mathbf{x}_j|\alpha, \beta)$ and thus the posterior distribution $p(\pi_j, \mathbf{z}_j|\alpha, \beta)$ are intractable for exact inference. Thus in [1], a Variational Bayes (VB) inference algorithm used a family of variational distributions:

$$q(\pi_j, \mathbf{z}_j|\gamma_j, \phi_j) = q(\pi_j|\gamma_j)\prod_{i=1}^{N} q(z_{ji}|\phi_{ji}) \quad (2)$$

to approximate $p(\pi_j, \mathbf{z}_j|\alpha, \beta)$, where the Dirichlet parameter γ_j and multinomial parameters $\{\phi_{ji}\}$ are free variational parameters. The optimal (γ_j, ϕ_j) is computed by finding a tight lower bound on $\log p(\mathbf{x}_j|\alpha, \beta)$.

This LDA model in [1] does not model clusters of documents. All the documents share the same Dirichlet prior α. In activity analysis, we assume that video clips (documents) of the same type of interaction would include a set of atomic activities (topics) in a similar way, so they could be grouped into the same cluster and share the same prior over topics. Our *LDA* model is shown in Figure 2(b). The M documents in the corpus will be grouped into L clusters. Each cluster c has its own Dirichlet prior α_c. For a document j, the cluster label c_j is first drawn from multinomial distribution η and π_j is drawn from $Dir(\pi_j|\alpha_{c_j})$. Given $\{\alpha_c\}$, β, and η, the joint distribution of hidden variables c_j, π_j, $\mathbf{z}_j$ and observed words $\mathbf{x}_j$ is $p(\mathbf{x}_j, \mathbf{z}_j, \pi_j, c_j|\{\alpha_c\}, \beta, \eta)$, given by:

$$p(c_j|\eta)p(\pi_j|\alpha_{c_j})\prod_{i=1}^{N} p(z_{ji}|\pi_j)p(x_{ji}|z_{ji}, \beta) \quad (3)$$

The marginal log likelihood of document j is:

$$\log p(\mathbf{x}_j|\{\alpha_c\}, \beta, \eta) = \log\sum_{c_j=1}^{L} p(c_j|\eta)p(\mathbf{x}_j|\alpha_{c_j}, \beta) \quad (4)$$

Using VB [1], $\log p(\mathbf{x}_j|\alpha_{c_j}, \beta)$ can be approximated by a tight lower bound $L_1(\gamma_{jc_j}, \phi_{jc_j}; \alpha_{c_j}, \beta)$. However because of the marginalization over c_j, hyperparameters are coupled. So we use both EM and VB to estimate hyperparameters. After using VB to compute the lower bound of $\log p(\mathbf{x}_j|\alpha_{c_j}, \beta)$, an averaging distribution $q(c_j|\gamma_{jc_j}, \phi_{jc_j})$ can provide a further lower bound on the log likelihood,

$$\log p(\mathbf{x}_j|\{\alpha_c\}, \beta, \eta) \geq \log\sum_{c_j=1}^{L} p(c_j|\eta)e^{L_1(\gamma_{jc_j}, \phi_{jc_j}; \alpha_{c_j}, \beta)}$$
$$= \log\sum_{c_j=1}^{L} q(c_j|\gamma_{jc_j}, \phi_{jc_j})\frac{p(c_j|\eta)e^{L_1(\gamma_{jc_j}, \phi_{jc_j}; \alpha_{c_j}, \beta)}}{q(c_j|\gamma_{jc_j}, \alpha_{jc_j})}$$
$$\geq \sum_{c_j=1}^{L} q(c_j|\gamma_{jc_j}, \phi_{jc_j})\left[\log p(c_j|\eta) + L_1(\gamma_{jc_j}, \phi_{jc_j}; \alpha_{c_j}, \beta)\right]$$
$$- \sum_{c_j=1}^{L} q(c_j|\gamma_{jc_j}, \phi_{jc_j})\log q(c_j|\gamma_{jc_j}, \phi_{jc_j})$$
$$= L_2(q(c_j|\gamma_{jc_j}, \phi_{jc_j}), \{\alpha_c\}, \beta, \eta) \quad (5)$$

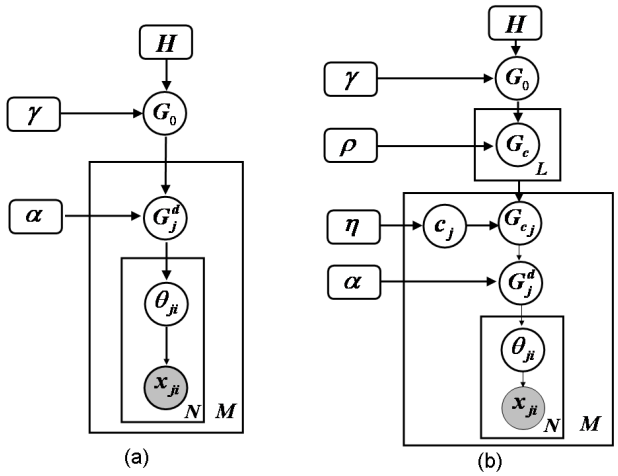


Figure 3. (a) *HDP* model proposed in [12]; (b) our *HDP* model.

L_2 is maximized when choosing

$$q(c_j|\gamma_{jc_j},\phi_{jc_j}) = \frac{p(c_j|\eta)e^{L_1(\gamma_{jc_j},\phi_{jc_j};\alpha_{c_j},\beta)}}{\sum_{c_j} p(c_j|\eta)e^{L_1(\gamma_{jc_j},\phi_{jc_j};\alpha_{c_j},\beta)}}. \quad (6)$$

Our EM algorithm for hyperparameters estimation is:

1. For each document j and cluster c_j, find the optimal values of the variational parameters $\{\gamma^*_{j,c_j},\phi^*_{j,c_j} : j = 1,\ldots,M, c_j = 1,\ldots,L\}$ to maximize L_1 (using VB [1]).
2. Compute $q(c_j|\gamma^*_{jc_j},\phi^*_{jc_j})$ using (6) to maximize L_2.
3. Maximize L_2 with respect to $\{\alpha_c\}$, β, and η. β and η are optimized by setting the first derivative to zero,

$$\eta_c \propto \sum_{j=1}^{M} q(c_j = c|\gamma^*_{jc},\phi^*_{jc}) \quad (7)$$

$$\beta_{kw} \propto \sum_{j=1}^{M}\sum_{c_j=1}^{L} q(c_j|\gamma^*_{jc_j},\phi^*_{jc_j})\left[\sum_{i=1}^{N}\phi^*_{jc_jik}x^w_{ji}\right] \quad (8)$$

where $x^w_{ji} = 1$ if $x_{ji} = w$, otherwise it is 0. The $\{\alpha_c\}$ are optimized using a Newton-Raphson algorithm.

L_2 monotically increases after each iteration.

3.2. HDP

HDP is a nonparametric hierarchical Bayesian model and automatically decides the number of topics. The *HDP* model proposed in [12] is shown in Figure 3 (a). A global random measure G_0 is distributed as a Dirichlet Process with concentration parameter λ and base probability measure H (H is a Dirichlet prior in our case):

$$G_0|\gamma, H \sim DP(\gamma, H).$$

G_0 can be expressed using a stick-breaking representation, $G_0 = \sum_{k=1}^{\infty}\pi_{0k}\delta_{\phi_k}$, where $\{\phi_k\}$ are parameters of multinomial distributions, $\phi_k \sim H$, $\pi_{0k} = \pi'_{0k}\prod_{l=1}^{k-1}(1-\pi'_{0l})$, $\pi'_{0k} \sim Beta(1,\lambda)$. G_0 is a prior distribution over the whole corpus. For each document j, a random measure G^d_j is drawn from a Dirichlet process with concentration parameter α and base probability measure G_0:

$$G^d_j|\alpha, G_0 \sim DP(\alpha, G_0).$$

Each G^d_j has support at the same points $\{\phi_k\}_{k=1}^{\infty}$ as G_0, and can be written as $G^d_j = \sum_{k=1}^{\infty}\pi_{jk}\delta_{\phi_k}$. G^d_j is a prior distribution of all the words in document j. For each word i in document j, a factor θ_{ji} is drawn from G^d_j (θ_{ji} is sampled as one of the ϕ_k's). Word x_{ji} is drawn from multinomial distribution $F(x_{ji};\theta_{ji})$. In [12], Gibbs sampling schemes were used to do inference under an *HDP* model.

In our *HDP* model, as shown in Figure 3 (b), clusters of documents are modeled and each cluster c has a random probability measure G_c. G_c is drawn from Dirichlet process $DP(\rho, G_0)$. For each document j, a cluster label c_j is first drawn from multinomial distribution $p(c_j|\eta)$. Document j chooses G_{c_j} as the base probabilty and draws its own prior from Dirichlet process $DP(\alpha, G_{c_j})$. We also use Gibbs sampling for inference. In our *HDP* model, there are two kinds of hidden variables to be sampled: (1) variables $\mathbf{k} = \{k_{ij}\}$ assigning words to topics, base distributions G_0, $\{G_k\}$; and (2) cluster label c_j. The key issue to be solved in this paper is how to sample c_j. Given c_j, the first kind of variables can be sampled using the schemes in [12].

At some sampling iteration, suppose that there have been K topics, $\{\phi_k\}_{k=1}^{K}$, generated and assigned to the words in the corpus (K is variable during the sampling procedure). G_0, G_c, and G^d_j can be expressed as, $G_0 = \sum_{k=1}^{K}\pi_{0k}\delta_{\phi_k} + \pi_{0u}G_{0u}$, $G_c = \sum_{k=1}^{K}\pi_{ck}\delta_{\phi_k} + \pi_{cu}G_{cu}$, $G^d_j = \sum_{k=1}^{K}\omega_{jk}\delta_{\omega_k} + \omega_{ju}G^d_{ju}$, where G_{0u}, G_{cu}, and G^d_{ju} are distributed as Dirichlet process $DP(\gamma, H)$.

Using the sampling schemes in [12], topic mixtures $\pi_0 = \{\pi_{01},\ldots,\pi_{0K},\pi_{0u}\}$, $\pi_c = \{\pi_{c1},\ldots,\pi_{cK},\pi_{cu}\}$ are sampled, while $\{\phi_k\}$, G_{0u}, G_{cu}, G^d_{ju}, and $\omega^d_j = \{\omega_{j1},\ldots,\omega_{jK},\omega_{ju}\}$ can be integrated out without sampling. In order to sample the cluster label c_j of document j, the posterior $p(c_j = c|(m_{j1},\ldots,m_{jK}),\pi_0,\{\pi_c\})$ has to be computed where m_{jk} is the number of words assigned to topic k in document j and is computable from $\mathbf{k}$.

$$\begin{aligned} &p(c_j = c|(m_{j1},\ldots,m_{jK}),\pi_0,\{\pi_c\}) \\ \propto &p(m_{j1},\ldots,m_{jK}|\pi_c)p(c_j = c) \\ = &\eta_c\int p(m_{j1},\ldots,m_{jK}|\omega^d_j)p(\omega^d_j|\pi_c)d\omega^d_j \end{aligned}$$

$p(m_{j1},\ldots,m_{jK}|\omega^d_j)$ is a multinomial distribution. Since G^d_j is drawn from $DP(\alpha, G_c)$, $p(\omega^d_j|\pi_c)$ is Dirichlet distri-

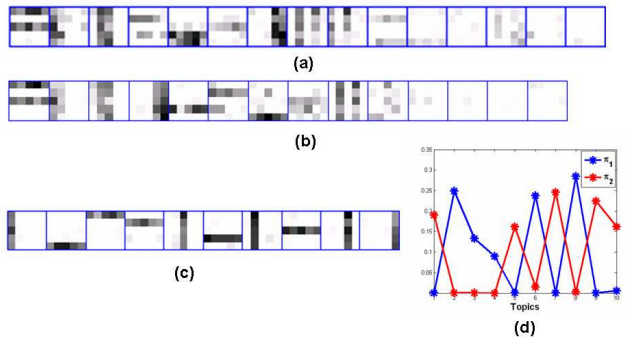


Figure 4. Bars example. (a) Topic distributions learnt by the *HDP* model in [12]. (b) Topics distributions learnt by the *HDP* model in Figure 3(b), however, the cluster labels of documents are randomly assigned and not updated by sampling. (c) Topic distributions learnt by our *HDP* model in Figure 3(b) where the cluster labels are updated using our Gibbs sampling algorithms. (d) Topic mixtures of two clusters π_1 and π_2.

bution $Dir(\omega_j^d|\alpha \cdot \pi_c)$. Thus we have

$$
\begin{aligned}
& p(c_j = c|(m_{j1}, \ldots, m_{jK}), \pi_0, \{\pi_c\}) \\
\propto & \eta_c \int \frac{\Gamma(\alpha\pi_{cu} + \alpha\sum_{k=1}^K \pi_{ck})}{\Gamma(\alpha\pi_{cu})\prod_{k=1}^K \Gamma(\alpha\pi_{ck})} \omega_{ju}^{\alpha\pi_{cu}-1} \prod_{k=1}^K \omega_{jk}^{\alpha\pi_{ck}+m_{jk}-1} d\omega_j^d \\
\propto & \frac{\Gamma(\alpha\pi_{cu} + \alpha\sum_{k=1}^K \pi_{ck})}{\Gamma(\alpha\pi_{cu})\prod_{k=1}^K \Gamma(\alpha\pi_{ck})} \frac{\Gamma(\alpha\pi_{cu})\prod_{k=1}^K \Gamma(\alpha\pi_{ck} + m_{jk})}{\Gamma(\alpha\pi_{cu} + \sum_{k=1}^K (\alpha\pi_{ck} + m_{jk}))} \\
\propto & \eta_c \frac{\prod_{k=1}^K \Gamma(\alpha \cdot \pi_{ck} + m_{jk})}{\prod_{k=1}^K \Gamma(\alpha \cdot \pi_{ck})}. \qquad (9)
\end{aligned}
$$

So the Gibbs sampling procedure repeats the following two steps alternatively at every iteration:

1. given $\{c_j\}$, sample $\mathbf{k}$, π_0, and $\{\pi_c\}$ using the schemes in [12];

2. given $\mathbf{k}$, π_0, and $\{\pi_c\}$, sample cluster labels $\{c_j\}$ using posterior Eq 9.

3.3. Simple explanatory example

We use an example of synthetic data to demonstrate the strength of our model. The word vocabulary is 5×5 cells. There are 10 topics with distributions over horizontal bars and vertical bars, i.e., cells tend to co-occur along the same row or column, but not arbitrarily. If we generate documents by randomly combining several bars and adding noise, there are only two levels of structures (topics and words) in the data and the HDP model in [12] usually can perfectly discover the 10 topics. However, in our experiments in Figure 4, we add one more cluster level to the data. Documents are from two clusters: a vertical-bars cluster and a horizontal-bars cluster. If a document is from the vertical-bars cluster, it randomly combines several vertical-bars, otherwise, it randomly combines horizontal bars. As seen in Figure 4 (a), *HDP* in [12] has much worse performance on this data. There are two kinds of correlation among words: if words are on the same bar, they often co-exist in the same documents; if words are all on horizontal bars or vertical bars, they are also likely to be in the same documents. It is improper to use a two-level *HDP* to model data with three-level structure. 15 topics are discovered and many of the topics include more than one bar. Even using a three-level *HDP*, if the cluster labels are not updated properly, the performance is still poor as shown Figure 4(b). Using our *HDP* model and simultaneously updating topics of words and clusters of documents, the 10 topics are discovered perfectly as shown in Figure 4(c). The topic mixtures π_1 and π_2 of two clusters are shown in Figure 4 (d). π_1 only has large weights on horizontal bars while π_2 only has large weights on vertical bars.

4. Application and Experimental Results

After computing the low-level visual features as described in Section 2, we divide our video sequence into 10 second long clips, each treated as a document, and feed these documents to the hierarchical models described in Section 3. In this section, we explain how to use the results from hierarchical Bayesian models for activity analysis. We will mainly show results from *HDP*, since *HDP* automatically decides the number of topics, while *LDA* needs to know that in advance. If the topic number is properly set in *LDA*, it provides similar results.

4.1. Summary of Typical Atomic Activities and Interactions

In scene analysis, people often ask "what are the typical activities and interactions in this scene?" The parameters estimated by our hierarchical Bayesian models provide a good answer to this question.

The topics of words are actually a summary of typical atomic activities in the scene. Each topic has a multinomial distribution over words (i.e., visual motions), specified by β in *LDA* and $\{\phi_k\}$ in *HDP* (ϕ_k can be easily estimated given the words assigned to topic k after sampling). We use the term "atomic" to indicate that the activity is usually performed by the agent coherently without break. The motions caused by the same kind of atomic activity often co-occur in the same video clip, thus being grouped into the same topic. Our *HDP* model automatically discovered 29 atomic activities in this traffic scene. In Figure 5, we show the motion distributions of these topics. The topics are sorted by size (the number of words assigned to the topic) from large to small. Topics 1 and 4 explain "vehicles pass road *d* from left to right". This activity is broken into two topics because

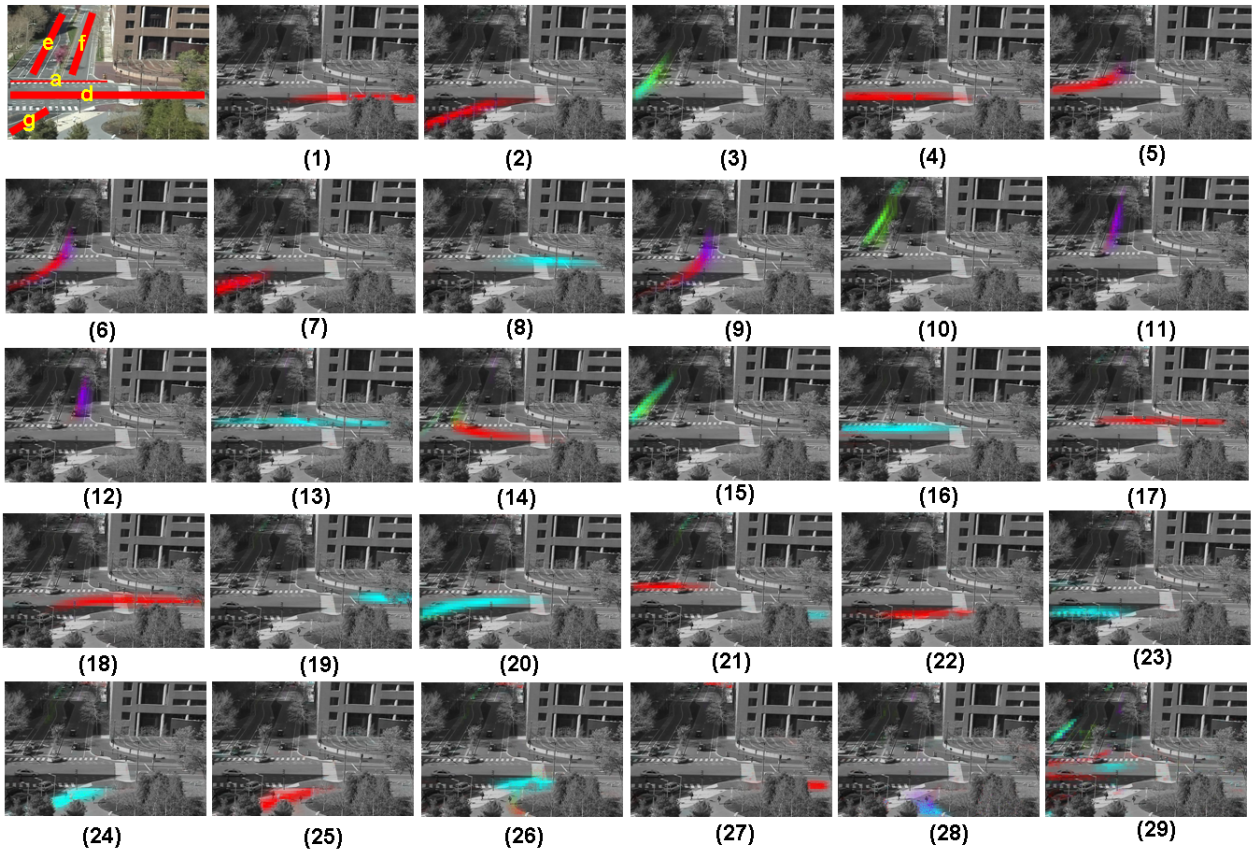


Figure 5. Motion distributions of some topics discovered by our *HDP* model. The motion is quantized into four directions represented by four colors as shown in Figue 1(a). The topics are sorted according to how many words in the corpus are assigned to them (from large to small). See all 29 topics in our supplementary video. For convenience, we label roads and crosswalks as $a, b, \ldots$ in the first image.

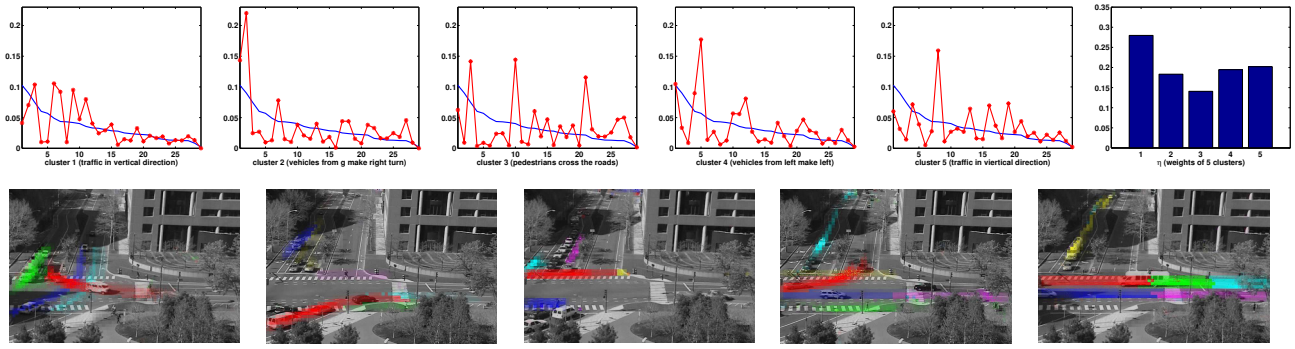
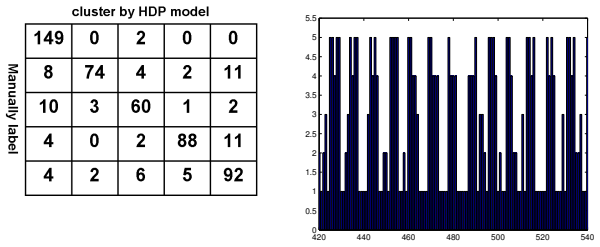


Figure 6. The short video clips are grouped into five clusters. **In the first row**, we plot the prior distribution over 29 topics as a red line for each cluster. For comparison, the blue line in each figure is the average topic distribution of the whole corpus. The large bar figure plots the cluster mixture weights (η in Figure 3 (b)). **In the second row**, we show a video clip as an example for each type of interaction and mark the motions of the five largest topics in that video clip. Notice that colors distinguish different topics in the same video (the same color may correspond to different topics in different video clips) instead of representing motion directions as in Figure 5.

when vehicles from g make a right turn (see topic 2) or vehicles from road e make a left turn (see topic 14), they also share the motion in 4. From topic 10 and 19, we find vehicles stopping behind the stop lines during red lights. Topics 13, 17, 21 explain that pedestrians walk on crosswalks. When people pass the crosswalk a, they often stop at the divider beween roads e and f waiting for vehicles to pass by. So this activity breaks into two topics 17 and 21.



(a) (b)

Figure 7. Results of video segmentation. (a) confusion matrix; (b) segmentation result of 20 minutes video.

Since a video clip often includes several topics, this can be explained by an interaction which is a combination of atomic activities. In our hierarchical Bayesian models, the video clips are automatically clustered into different interactions. The topics mixtures ($\{\alpha_c\}$ in *LDA* and $\{\pi_c\}$ in *HDP*) as priors on clusters provide a good summary of interactions. Figure 6 plots the topic mixtures π_c of five clusters under our *HDP* model. Cluster 1 explains traffic moving in a vertical direction. Vehicles from e and g move vertically, crossing road d and crosswalk a. 3, 6, 7, 9 and 11 are major topics in this interaction, while other topics related to horizontal traffic(1, 4, 5, 8, 16, 20), and pedestrians walking on crosswalk a and b (13, 17, 21, 23), are very low. Cluster 2 explains "vehicles from road g make right turn to road a while there is not much other traffic". At this time, vertical traffic is forbidden while there are no vehicles traveling horizontally on road d, so these vehicles from g can make right turn. Cluster 3 is "pedestrians walk on the crosswalks while there is not much traffic". Several topics (21, 13, 17) related to pedestrian walking are much higher that the average distribution on the whole corpus. Topics 10 and 15 are also high because they explain that vehicles on road e stop behind the stop line. Cluster 4 is "vehicles on road d make left turn to road f". Topics 5, 11 and 12 related to this activity are high. Topics 1 and 4 are also high since horizontal traffic from left to right is allowed at this time. However topics 8, 16 and 20 are very low, because traffic from right to left conflicts with this left turn. Cluster 5 is horizontal traffic. During this interaction, topics 13, 17 and 21 are also relatively high, since pedestrians are allowed to walk on a.

4.2. Video Segmentation

Given a long video sequence, we want to segment it based on different types of interaction occuring, and also detect single-agent activities both temporally and spatially. Our models provide a natural way to complete this task since video clips are labeled into clusters (interactions) and motion features are assigned to topics (activities). To evaluate the clustering performance, we manually label the 540 video clips into five typical interactions in this scene as described in Section 4.1. The confusion matrix is shown in Figure 7 (a). The average accuracy is 85.74%. Figure 7 shows the labels of video clips in the last 20 minutes. We can observe that each traffic cycle lasts around 85 seconds. In the second row of Figure 6, we show an example video clip for each type of interaction. In each video clip, we choose the five largest topics and mark motions belonging to different topics by different colors.

4.3. Abnormality Detection

We want to detect abnormal video clips and localize abnormal activities in the video clip. Under the Bayesian models, abnormality detection has a nice probabilistic explanation since every video clip and motion can be assigned a marginal likelihood, rather than by comparing similarities between samples. Computing the likelihoods of documents and words under *LDA* has been described in Section 3.1 (see Eq 5). Computing the likelihood under *HDP* is not straightforward. We need to compute the likelihood of document j given other documents, $p(\mathbf{x}_j|\mathbf{x}^{-j})$, where $\mathbf{x}^{-j}$ represents the whole corpus excluding document j. Since we have already drawn M samples $\{\mathbf{k}^{-j(m)}, \{\pi_l^{(m)}\}, \pi_0^{(m)}\}_{m=1}^M$ from $p(\mathbf{k}^{-j}, \{\pi_l\}, \pi_0|\mathbf{x})$ which is very close to $p(\mathbf{k}^{-j}, \{\pi_l\}, \pi_0|\mathbf{x}^{-j})$, we approximate $p(\mathbf{x}_j|\mathbf{x}^{-j})$ as

$$\frac{1}{M}\sum_m \sum_{c_j} \int_{\omega_j} \sum_{\mathbf{k}_j} \sum_i p(x_{ji}|k_{ji}, \mathbf{k}^{j(m)}, \mathbf{x}^{-j}) \\ p(\mathbf{k}_j|\omega_j) p(\omega_j|\pi_{c_j}^{(m)}) \eta_{cj} d\omega_j$$

$p(k_j|\pi_{c_j}^{(m)})$ is a Dirichlet distribution. If $(u_1, \ldots, u_T)$ is the Dirichlet prior on ϕ_k, $p(x_{ji}|k_{ji}, \mathbf{k}^{j(m)}, \mathbf{x}^{-j}) = (u_{x_{ji}} + n_{x_{ji}})/(\sum_{t=1}^T (u_t + n_t))$ is a multinomial distribution, where n_t is the number of words in $\mathbf{x}^{-j}$ with value t assigned to topic k_{ji}(see [12]). The computation of $\int_{\omega_j} \sum_{\mathbf{k}_j} p(x_{ji}|k_{ji}, \mathbf{k}^{j(m)}, \mathbf{x}^{-j}) p(\mathbf{k}_j|\omega_j) p(\omega_j|\pi_{c_j}^{(m)})$ is intractable, but can be approximated with a variational inference algorithm as in [1].

Figure 8 shows the top five abnormal video clips detected by *HDP*. The red color highlights the regions with abnormal motions in the video clips. There are two abnormal activities in the first video. A vehicle is making a right-turn from road d to road f. This is uncommon in this scene because of the layout of the city. Actually there is no topic explaining this kind of activity in our data (topics are summaries of typical activities). A person is approaching road f, causing abnormal motion. In the successive video clip, we find that the person is actually crossing road f outside the crosswalk region. This video clip ranked No.4 in abnormality. In the second and third videos, bicycles are crossing the road abnormally. The fifth video is another example of a pedestrian

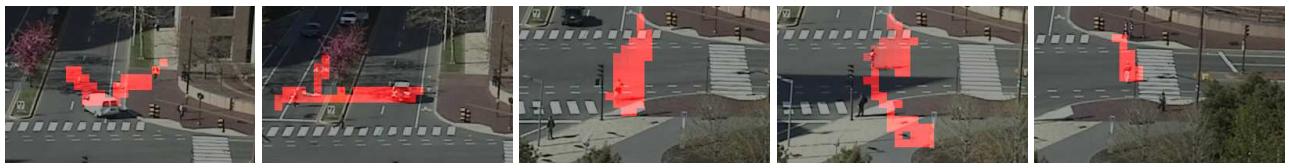
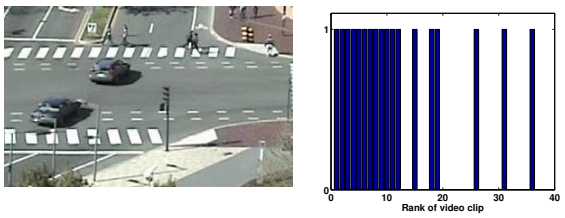


Figure 8. Results of abnormality detection. We show the top five video clips with the highest abnormality (lowest likelihood). In each video clip, we highlight the regions with motions of high abnormality.



(a) (b)

Figure 9. Query result of jay-walking(co-occurence of topics 6 and 13). (b) shows the top 40 retrieval results. If the video clip is correct, it is labeled as 1 otherwise 0.

crossing the road outside the crosswalk.

4.4. High-Level Semantic Query

In our framework, it is convenient to use mid-level representations of atomic activities as tools to query for examples of interactions of interest. Suppose a user wants to detect jay-walking. This is not automatically discovered by the system as a typical interaction. Thus, the user simply picks topics involved in the interaction, e.g. topic 6 and 13, and specifies the query distribution q ($q(6) = q(13) = 0.5$ and other mixtures are zeros). The topic distributions $\{p_j\}$ of video clips in the data set match with the query distribution using relative entropy between q and p_j. Figure 9 (b) shows the result of querying examples of "pedestrians walk on crosswalk *a* from right to left while vehicles are approaching in vertical direction". All the video clips are sorted by matching similarity. The true instance will be labeled 1 ortherwise it is labeled as 0. There are totally 18 instances in this data set and they are all found among the top 37 examples. The top 12 retrieval results are all correct.

5. Conclusion

We have proposed an unsupervised framework adopting hierarchical Bayesian models to model activities and interactions in complicated scenes. We improve *LDA* and *HDP* models by co-clustering both words and documents. Our system summarizes typical activities and interactions in the scene, segments the video sequence both temporally and spatially, detects abnormal activities and supports high-level semantic queries on activities and interactions.

6. Acknowledgement

The authors wish to acknowledge DARPA and DSO National Laboratories of Singapore for partially supporting this research.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. 2, 3, 4, 7

[2] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2004. 2

[3] N. Ghanem, D. Dementhon, D. Doermann, and L. Davis. Representation and recognition of events in surveillance video using petri net. In *CVPR Workshop*, 2004. 2

[4] S. Honggeng and R. Nevatia. Multi-agent event recognition. In *Proc. ICCV*, 2001. 2

[5] F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, 2005. 2

[6] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *JCAI*, pages 674–680, 1981. 2

[7] J. C. Niebles, H. Wang, and F. Li. Unsupervised learning of human action categories using spatial-temporal words. In *Proc. BMVC*, 2006. 2

[8] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Trans. on PAMI*, 22:831–843, 2000. 2

[9] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proc. ICCV*, 2005. 2

[10] P. Smith, N. V. Lobo, and M. Shah. Temporalboost for event recognition. In *Proc. ICCV*, 2005. 2

[11] C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on PAMI*, 22:747–757, 2000. 2

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet process. *Journal of the American Statistical Association*, 2006. 2, 4, 5, 7

[13] X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *Proc. ECCV*, 2006. 2

[14] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *Proc. CVPR*, 2001. 1

[15] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *Proc. CVPR*, 2004. 1, 2