Action Recognition Using Topic Models

Xiaogang Wang

Abstract In this book chapter, we will introduce approaches of using topic models for action recognition. Topic models were originally developed in language processing. In recent years, they were applied to action recognition and other computer vision problems, and achieved great success. Topic models are unsupervised. The models of actions are learned through exploring the co-occurrence of visual features without manually labeled training examples. This is important when there are a large number of actions to be recognized in a large variety of scenes. Most topic models are hierarchical Bayesian models and they jointly model simple actions and complicated actions at different hierarchical levels. Various knowledge and contextual information can be well integrated into topic models as priors. We will explain how topic models can be used in different ways for action recognition in different scenarios. For examples, the scenes may be sparse or crowded. There may be a single camera view or multiple camera views. The camera settings may be near-field or far-field. In different scenarios, different features, such as trajectories, local motions and spatial-temporal interest points, are used for action recognition.

Keywords: action recognition, topic models, hierarchical Bayesian models, clustering

1 Introduction

Action recognition from video sequences has a wide variety of applications in both public and private environments, such as homeland security [1, 2], crime prevention [3, 4, 5], traffic control [6, 7, 8], accident prediction and detection [9], and monitoring patients, elderly and children at home [10]. These applications include a

Xiaogang Wang
Department of Electronic Engineering, The Chinese University of Hong Kong, e-mail: xgwang@ee.cuhk.edu.hk

large variety of scenes such as airports, train stations, highways, parking lots, stores, shopping malls and offices. Due to the fast growing of cheap sensors and video data and also a growing need for security and efficient information retrieval, there are increasing demands on automatic action recognition. Over the past decade, significant work has been reported on this topic. Literature reviews can be found in [11, 12].

Some existing approaches [13, 14, 15] required manually labeling examples to train classifiers or discriminative models for action recognition. Some of them required training different classifiers or models for different scenes. Because of the large number of different action categories to be recognized and the large variety of different scenes, people prefer algorithms [16, 17, 18] which automatically learn the models of the actions in the target scenes without supervision.

Many approaches [19, 17, 16, 20, 2, 18] directly used motion feature vectors to describe video clips without tracking objects. For example, Zelnik-Manor and Irani [16] modeled and clustered video clips using multi-resolution histograms. Zhong et al. [17] also computed global motion histograms and did word-document analysis on videos. Their words were frames instead of moving pixels. They clustered video clips through the partition of a bipartite graph. Without object detection and tracking, a particular activity cannot be separated from other activities simultaneously occurring in the same clip, as is common in crowded scenes. These approaches treated a video clip as an integral entity and tagged the whole clip as normal or abnormal. They were often applied to simple data sets where there was only one kind of activity in a video clip.

In some approaches, objects (or their parts) were first detected, tracked and classified into different classes. Their tracks were used as features to model activities [21, 22, 23]. 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 [14, 24]. 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.

In recent years, inspired by the great success of topics models, such as Probabilistic Latent Semantic Analysis (pLSA) [25] and Latent Dirichlet Allocation (LDA) [26], in the applications of language processing, they have been also applied to action recognition. Significant progress has been made. Topic models recognize actions through exploring the co-occurrence of features at different hierarchical levels. Compared with other approaches, topic models have some attractive features. Firstly, they are unsupervised and learn the models of actions without requiring manually labeling training examples. They can also separate co-occurring actions without human intervention. Secondly, topic models allows the models of different action classes to share features and training data. Therefore the models of action classes can be learned more robustly avoiding the overfitting problem. Thirdly, most topic models are hierarchical Bayesian models, which allows to jointly model simple and complicated actions at different levels. Various knowledge and constraints can be nicely added into Bayesian frameworks as priors. Thus they can better solve

problems which are difficult for nonBayesian approaches such as jointly modeling actions in multiple views and dynamically updating the models of actions. Lastly, they can be well integrated with nonparametric Bayesian models, which use Dirichlet Processes (DP) [27] as priors to automatically learn the number of action classes from data without being manually specified.

In this chapter, we will introduce three types of approaches of using topic models for action recognition in different scenarios based on different types of features. The approaches introduced in Section 3 assume that cameras are stationary and scenes are parse. The trajectories of objects are available by tracking objects and are used as features for action recognition. The approaches in Section 4 assume that the cameras are stationary, scenes are crowded and there different types of actions simultaneously happening. It is very difficult to detect and track objects in crowded scenes because of frequent occlusions. Local motions (such as moving pixels) are used as features to model actions without tracking objects. Topic models are able to separate co-occurring actions and jointly model simple actions and complicated global behaviors of the videos. In both Section 3 and Section 4, topic models can recognize actions in single camera views or multiple camera views. In Section 5, cameras are not necessary to be stationary and interest points are used as features to recognition human actions in near fields.

2 Topic models

Under topic models in language processing, words, such as "professor" and "university", often co-existing in the same documents are clustered into the same topic, such as "education". A document is modeled as a mixture of topics and each word is generate from a single topic. pLSA [25] and LDA [26] are two popular topic models. Their graphical models are shown in Figure 1. Suppose there are M documents in the data set. Each document j has N_j words. Each word w_{ji} is assigned one of the K topics according to its topic label z_{ji}. Under pLSA the joint probability $P(\{w_{ji}\},\{d_j\},\{z_{ji}\})$ has the form of the graphical model shown in Figure 1(a). The conditional probability $P(w_{ji}|d_j)$ marginalizing over topics z_{ji} is given by

$$P(w_{ji}|d_j) = \sum_{k=1}^{K} P(z_{ji}=k|d_j)P(w_{ji}|z_{ji}=k). \tag{1}$$

$P(z_{ji}=k|d_j)$ is the probability of topic k occurring in document d_j. $P(w_{ji}|z_{ji}=k)$ is the probability of word w_{ji} occurring in topic k and is the model of topic k. Fitting the pLSA model involves determining $P(w_{ji}|z_{ji})$ and $P(z_{ji}=k|d_j)$ by maximizing the following objective function using the Expectation Maximization (EM) algorithm:

$$L = \prod_{j=1}^{M}\prod_{i=1}^{N_j} P(w_{ji}|d_j) \tag{2}$$

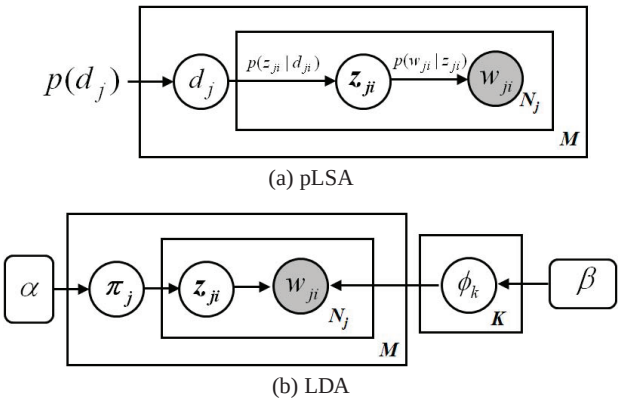


Fig. 1 Graphical models of pLSA and LDA.

pLSA is a generative model only for training documents but not for new documents. pLSA does not provide probabilistic model at the level of documents. Each model is represented by a list of numbers $p(z|d_j)$, but these numbers are not generated from a probabilistic model. This shortcoming has been addressed by LDA, whose graphical model is shown in Figure 1(b). The generative process of LDA is described as following.

1. $\{\phi_k\}$ are models of topics and are discrete distributions over the codebook of words. They are generated from a Dirichlet prior $Dir(\phi_k;\beta)$ given by β.
2. Each document j has a multinomial distribution π_j over K topics and it is generated from a Dirichlet prior $Dir(\pi_j;\alpha)$ given by α.
3. Each word i in document j is assigned to one of the K topics and its label z_{ji} is sampled from a discrete distribution $Discrete(z_{ji};\pi_j)$ given by π_j.
4. The observed word w_{ji} is sampled from the model of its topic: $Discrete(w_{ji}|\phi_{z_{ji}})$.

α and β are hyperparameters. ϕ_k, π_j and z_{ji} are hidden variables to be inferred. The joint distributions of the LDA model is

$$p(\{w_{ji}\},\{z_{ji}\},\{\phi_k\},\{\pi_j\}|\alpha,\beta)=\prod_{k=1}^{K}p(\phi_k|\beta)\prod_{j=1}^{M}p(\pi_j|\alpha)\prod_{i=1}^{N_j}p(z_{ji}|\pi_j)p(w_{ji}|\phi_{z_{ji}}). \tag{3}$$

Under LDA, if two words often co-occur in the same documents, one of the topic models will have large distributions on both of them.

Both pLSA and LDA require the number of object classes to be known in advance. As an extension of LDA, Hierarchical Dirichlet Process (HDP) proposed by Teh et al. [28] could automatically learn the number of topics from data using Dirichlet Processes [27] as priors.

When topic models are applied to action recognition, words, documents and topics will be mapped to some specific concepts in the context of action recognition. Under topic models, visual features are quantized into visual words. Topic models

will explore the co-occurrence of visual words to learn the models of action classes. Although the co-occurrence of visual words is commonly observed in actions and can be used to learn the models of action classes, the "bag-of-words" assumption loses the spatial and temporal relationship among visual words, which is also very important for action recognition. Therefore, when topic models are applied to action recognition, they are modified to incorporate such information. Topic models also have been widely applied other computer vision problems such as object segmentation [29] and scene categorization [30]. One of the major advantages of topic models is their unsupervised nature. This is very important for discovering classes of actions from large amounts of video surveillance data and videos collected from the web. Topic models are hierarchical Bayesian models, under which topics are middle-level representations. Through sharing topics among documents, training data is shared, which avoids the overfitting problem to some extend. Topic models are hierarchical Bayesian models, which can flexibly include spatial and temporal information as priors.

2.1 Inference on topic models

Doing inference on the hidden variables of topic models is a big challenge. For example, in LDA the posteriors of hidden variables need to be computed,

$$p(\{\pi_j\},\{\phi_k\},\{z_{ji}\}|\{w_{ji}\},\alpha,\beta) = \frac{p(\{\pi_j\},\{\phi_k\},\{z_{ji}\},\{w_{ji}\}|\alpha,\beta)}{p(\{w_{ji}\}|\alpha,\beta)}. \quad (4)$$

However, this posterior distribution is intractable to compute. A variety of approximate inference algorithms were proposed. Blei et al. [26] proposed a variational inference algorithm on LDA. It considers a family of lower bounds, indexed by a set of variational parameters. The target posterior distribution is approximated by found the tightest lower bound by an optimization procedure. Variational inference methods for Dirichlet processes are also proposed [31, 32]. A drawback of variational inference is that it is not clear how big the gap is between the found lower bound and the target distribution.

Another type of inference methods are based on Markov chain Monte Carlo (MCMC) [33]. Griffths et al. [34] proposed a collapsed Gibbs sampling algorithm for the inference on LDA. It generates a sequence of samples from the distribution 4 in iterative steps. At each step, a hidden variable is sampled given other variables sampled from previous steps. The hidden variables π_j and ϕ_k can be analytically marginalized without being sampled and the sampling efficiency can be improved. Only z_{ji} needs to be sampled from the following distribution,

$$p(z_{ji}|\{z_{j'i'}\}^{-ji},\alpha,\beta) \propto \frac{m_{k,w_{ji}}^{-ji}+\beta}{m_{k\cdot}^{-ji}+W\cdot\beta}\cdot\frac{n_{jk}^{-ji}+\alpha}{K\cdot\alpha+n_{j\cdot}^{-ji}} \quad (5)$$

where W is the size of the dictionary, m_{kw} is the number of words assigned to topic k with value w, $m_{k\cdot}$ is the number of words assigned to topic k, n_{jk} is the number of words assigned to topic k in document j, $n_{j\cdot}$ is the total number of words in document j. $m^{-ji}_{kw_{ji}}$, $m^{-ji}_{k\cdot}$, n^{-ji}_{jk} and $n^{-ji}_{j\cdot}$ are the same statistics as m_{kw}, $m_{k\cdot}$, n_{jk} and $n_{j\cdot}$ except that they have excluded the word i in document j. ϕ_k and π_j can be estimated from from $\{z_{ji}\}$,

$$\hat{\phi}_k = \frac{m_{kw} + \beta}{m_{k\cdot} + W \cdot \beta} \tag{6}$$

$$\hat{\pi}_j = \frac{n_{jk} + \alpha}{K \cdot \alpha + n_{j\cdot}}. \tag{7}$$

Teh et al. [28] used collapsed Gibbs sampling to do inference on HDP. One of the drawbacks of MCMC is its low efficiency. Also there is hard to get the theoretical justification on the convergence of a MCMC sampling algorithm. In order to improve the efficiency of inference, distributed inference for topic models are proposed [35].

3 Far-field action recognition based on trajectories of objects

3.1 Single camera view

In far-field video surveillance, if there is only one camera view and the scene is sparse, objects can be detected and tracked. In far-field settings, the captured videos are of low resolution and poor quality, and therefore it is difficult to compute more complicated visual features. Usually only positions of objects are recorded along the tracks, which are called trajectories. The majority of visible actions of objects are distinguished by the patterns of objects moving from one location to another and trajectories are used as features for action recognition. Many approaches were proposed to cluster trajectories into different action classes and detect abnormal trajectories. New trajectories were classified into one of the existing clusters. Most of existing approaches [36, 23, 12] first defined the pairwise similarities between trajectories and the computed similarity matrix was input to some standard clustering algorithms.

An approach of clustering trajectories using topic models was proposed in [37]. In this approach, trajectories are treated as documents, observations (points) on trajectories are treated as words, and semantic regions are treated as topics. Observations are quantized into words according to a feature codebook based on their locations and moving directions. To build the feature codebook, the 2D space of the scene in uniformly divided into small cells and the moving direction is quantized into four. In the physical world, objects move along some paths. We refer

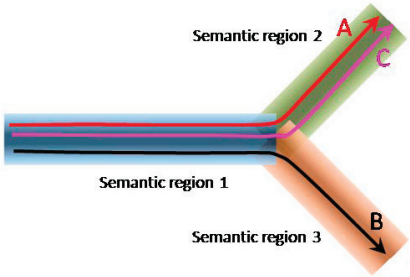


Fig. 2 Example to explain the modeling of semantic regions and actions. There are three semantic regions (indicated by different colors) which form two paths. Both trajectories A and C pass through regions 1 and 2, so they are clustered into the same action class. Trajectory B passes through regions 1 and 3, so it is clustered into a different action class.

to the subsets [1] of paths commonly taken by objects as semantic regions, i.e. two paths may share one semantic region as shown in Figure 2. Semantic regions are modeled as discrete distributions over the quantized space of the scene and moving directions. Their models are learned from the co-occurrence of features. When we track objects, identity co-occurrence of feature values can be observed. Identity co-occurrence means that two feature values are observed on the same trajectory and they are related to the same object. If two locations are on the same semantic regions, they are connected by many trajectories and both of them will have large weights on one of the models of semantic regions learned by the topic model. On the other hand, if two trajectories pass through the same combination of semantic regions, they are on the same path and thus they belong to the same action class. A Dual Hierarchical Dirichlet Processes (Dual-HDP) model was proposed in [37] to jointly learns the models of semantic regions and cluster trajectories into different paths (action classes). Dual-HDP is a non-parametric extension of the LDA mixture model, whose graphical model is shown in Figure 3. To simplify the description, we only explain the LDA mixture model below. The advantage of Dual-HDP is to automatically learn the number of semantic regions and the number of paths from data using Dirichlet processes as priors.

The LDA model in Figure 1 (b) does not model clusters of trajectories (documents). All the trajectories share the same Dirichlet prior α. In action recognition, we assume that trajectories of the same action class are on the same path and pass through the same set of semantic regions (topics). So they would be grouped into the same clusters and share the same Dirichlet prior over semantic regions. In the LDA mixture model shown in Figure 3, the M trajectories are grouped into L clusters. Each cluster c has its own Dirichlet prior α_c. For a trajectory j, its cluster label c_j is first drawn from a discrete distribution η. The joint distribution of hidden variables

[1] If a path is viewed as a set of quantized spatial locations and moving directions, semantic regions are subsets of paths and they can obtained through the operations of intersection and set difference between paths.

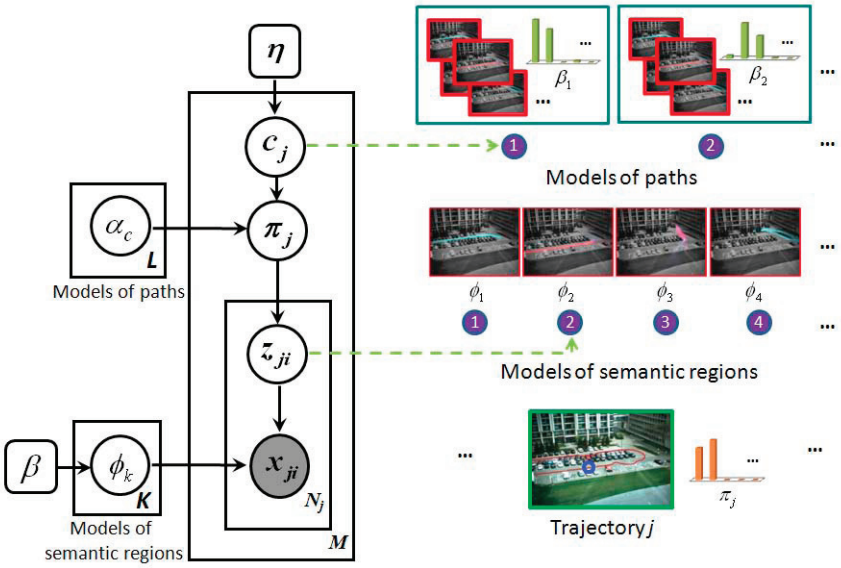


Fig. 3 Graphical model of the LDA mixture model for action recognition based on trajectories.

c_j, π_j, z_{ji} and observed words x_{ji} is given by,

$$\begin{aligned} & p(\{x_{ji}\},\{z_{ji}\},\{\pi_j\},\{c_j\},\{\phi_k\}|\{\alpha_c\},\beta,\eta) \\ =& \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{j=1}^{M} p(c_j|\eta) p(\pi_j|\alpha_{c_j}) \prod_{i=1}^{N_j} p(z_{ji}|\pi_j) p(x_{ji}|z_{ji},\phi_{z_{ji}}) \end{aligned} \tag{8}$$

z_{ji}, π_j, c_j, ϕ_k and α_c can be learned using variational methods [38] or collapsed Gibbs sampling [37]. Once the models ($\{\phi_k\}$) of semantic regions and the models $\{\alpha_c\}$ of paths are learned and fixed, they can be used to classify a new trajectory into one of the unsupervisedly learned paths (action classes). They also be used to detect abnormal trajectories which do not fit any of the existing path models.

In [37], experiments were conducted on a large scale trajectory data set including more than 40,000 trajectories collected from a parking lot scene. Some results are shown in Figure 4. These semantic regions and trajectory clusters represent some typical activities happening in this scene. For example, the first cluster in Figure 4 explains vehicles entering and leaving the parking lot. The last cluster explains pedestrians entering the parking lot and crossing the grass field. Some abnormal trajectories are detected. Many of them are pedestrian walking along the grass field. Some are pedestrian crossing the parking lot and turning back in the middle. These activities are uncommonly seen in this scene. They do not fit topic models, have low likelihoods and therefore are detected as abnormalities.

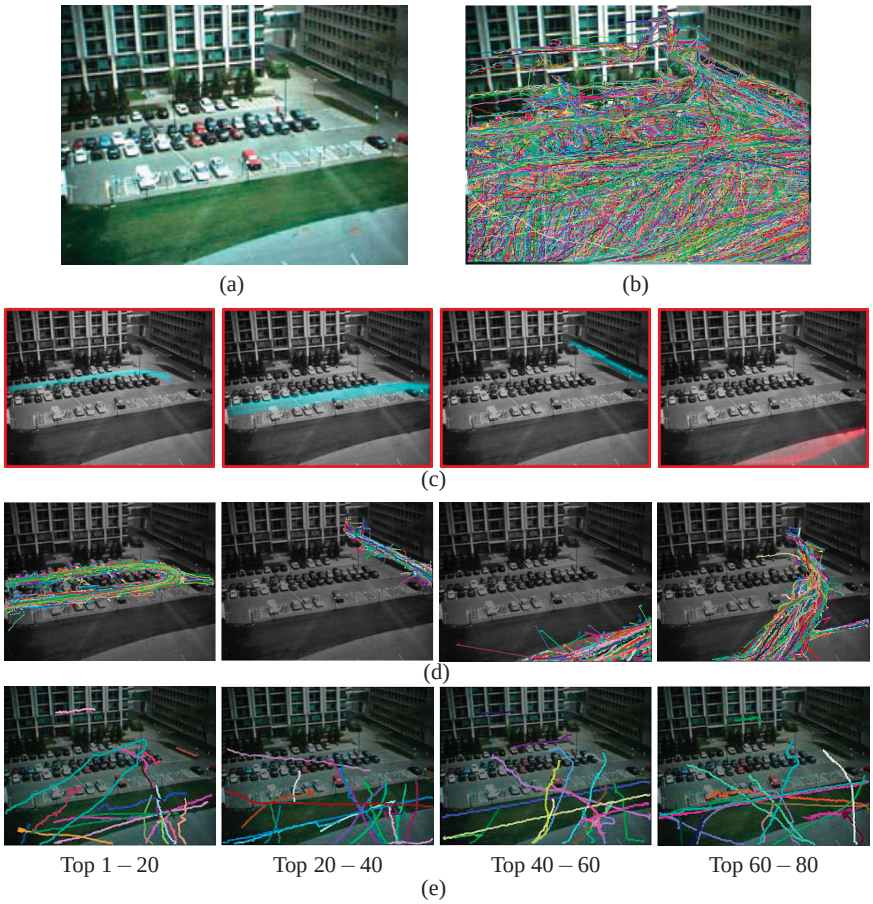


Fig. 4 Experimental results of the approach in [37] on more than 40,000 trajectories collected from a parking lot scene. (a) The background image of the parking lot scene. (c) Trajectories collected from the parking lot scene within one week. (c) Learned models of semantic regions. Colors represent different moving directions: $\rightarrow$ (red), $\leftarrow$ (cyan), $\uparrow$ (magenta), and $\downarrow$ (blue). (d) Clusters of trajectories. They represent different classes of actions. (e) Detected top 80 abnormal trajectories.

3.2 Multiple camera views

The approach proposed in [37] was only applicable to a single camera view. In [39, 40], this topic model was extended to jointly model actions in multiple camera views. Many existing approaches [41, 42] of action recognition in multiple camera views required inference on the topology of a camera network and a solution to the correspondence problem, i.e. tracking objects across camera views. Some had constraints on the topology of camera views (e.g. camera views must have significant

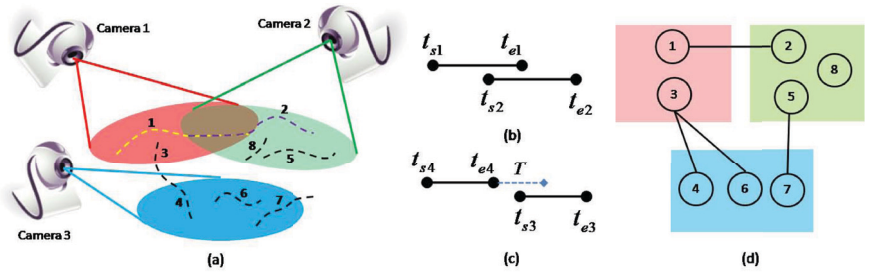


Fig. 5 Example of building a network connecting trajectories in multiple camera views. (a) Trajectories in three camera views. (b) The temporal extents of trajectories 1 and 2. (c) The temporal extents of trajectories 3 and 4. (d) The network connecting trajectories. See text for details.

overlaps) and required a lot of human effort. The approach proposed in [40] recognized actions in multiple camera views without doing inference on the topology of camera views and without solving the correspondence problem. It assumed that the topology of camera views was unknown and arbitrary, and the cameras were not calibrated. The camera views might be disjoint. Objects were first tracked in each camera view independently without being tracked across camera views. The goal was to learn the model of an action with distributions in all the camera views and cluster trajectories across camera views without supervision. As an extension of [37], [40] assumed that if two trajectories were observed in two camera views around the same time, they were more likely to be the same object and thus should have a higher probability to be in the same action class under the model to be learned. A smoothness constraint, which required that two trajectories with strong temporal correlation should have the same action label to avoid penalty, was added as prior in the hierarchical Bayesian model to cluster trajectories across camera views.

A network is built connecting trajectories observed in multiple camera views based on their temporal extents. Each trajectory is a node on the network. Let t_{si} and t_{ei} be the starting and ending time of trajectory i. T is a positive temporal threshold. It is roughly the maximum transition time of objects moving between adjacent camera views. If trajectories a and b are observed in different camera views and their temporal extents are close,

$$(t_{sa} \leq t_{sb} \leq t_{ea} + T) \vee (t_{sb} \leq t_{sa} \leq t_{eb} + T), \tag{9}$$

then a and b will be connected by an edge on the network. This means that a and b may be the same object since they are observed by cameras around the same time. There is no edge between two trajectories observed in the same camera view. An example can be found in Figure 5. As shown in (a), the views of cameras 1 and 2 overlap and are disjoint with the view of camera 3. Trajectories 1 and 2 observed by cameras 1 and 2 correspond to the same object moving across camera views. Their temporal extents overlap as shown in (b), so they are connected by an edge on the network as shown in (d). Trajectories 3 and 4 observed by cameras 1 and

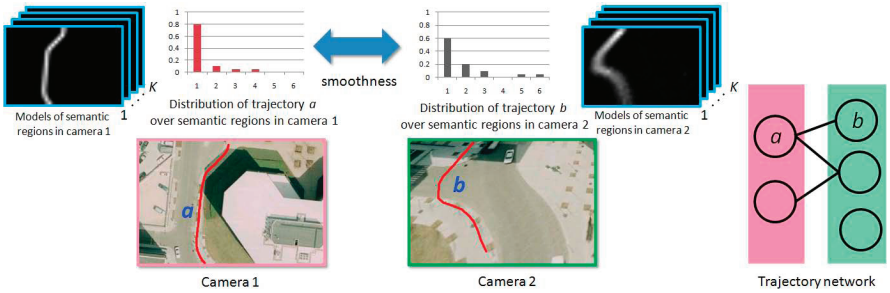


Fig. 6 Example to describe the high level picture of our model. See detail in the text.

3 correspond to an object crossing disjoint views. Their temporal extents have no overlap but the gap is smaller than T as shown in (c), so they are also connected. Trajectories 3 and 6, 5 and 7 do not correspond to the same objects, but their temporal extents are close, so they are also connected on the network. A single trajectory 3 can be connected to two trajectories (4 and 6) in other camera views. An edge on the network indicates a possible correspondence candidate only based on the temporal information of trajectories. But we do not really solve the correspondence problem when building the trajectory network, since many edges are actually false correspondences. The network simply keeps all of the possible candidates.

In Figure 6, we use an example to describe the high-level picture of this approach. Trajectories *a* and *b* are observed in different camera views and connected by an edge on the trajectory network. Points on trajectories are assigned to semantic regions by fitting models of semantic regions. The model of a semantic region has joint distributions in all the camera views. Both *a* and *b* have distributions over semantic regions. The smoothness constraint requires that their distributions over semantic regions are similar in order to have small penalty. In this example, both trajectory *a* and *b* have a larger distribution on semantic region 1, so the models of semantic region 1 in two different camera views can be related to the same action class.

Examples of trajectory clusters obtained by the approach in [40] on a street scene with four cameras are shown in Figure 7. Action 1 captures vehicles moving on the road. It is observed by all of the four cameras. Vehicles first move from the top-right corner to the bottom-left corner in the view of camera 4. Then they enter the bottom region in the view of camera 1 and move upward. Some vehicles disappear at the exit points observed in the views of cameras 2 and 3, and some move further beyond the view of camera 3. In action 2, pedestrians first walk along the sidewalk in the view of camera 1, and then cross the street as observed by camera 4.

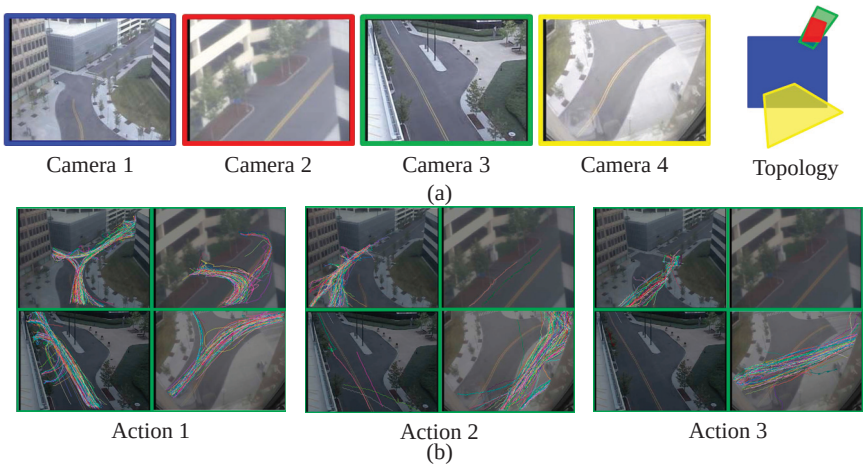


Fig. 7 Experimental results of the approach in [40] on a street scene. (a) Camera views and their topology in a street scene. When the topology of camera views is plotted, the fields of camera views are represented by different colors: blue (camera 1), red (camera 2), green (camera 3), yellow (camera 4). However, the approach approach in [40] does not require knowledge of the topology of the cameras in advance. (b) Examples of trajectory clusters which correspond to different action classes in the street scene.

4 Far-field action recognition based on local motions

The approaches introduced in Section 3 do not work well in crowded scenes where it is difficult to detect and track objects because of frequent occlusions. In [38, 43], an approach was proposed to jointly detect single-agent actions and global behaviors of video clips in crowded scenes using Dual-HDP. In crowded scenes, different types of single-agent actions often happen simultaneously and it is difficult to separate them without detecting and tracking objects. Global behaviors are characterized by the combinations of different types of single-agent actions co-occurring in the video clips. Although some approaches [16, 17] used motion feature vectors of video clips to model global behaviors of whole video clips without tracking objects, they had difficulty of separating co-occurring activities. The approach in [43] used moving pixels to drive the representation of actions and global behaviors without tracking objects. It assumed that motion features related to the same action class had temporal correlation because an action typically generates continuous motions in time. It leaned action models over motion features by exploring temporal co-occurrence information of features and separated co-occurring actions without human labeling effort using topic models. These action models were used as components to model more complicated global behaviors of video clips. This approach is introduced below.

It computes local motions as low-level features. Moving pixels are detected in each frame as follows: it computes the intensity difference between two successive frames, on a pixel basis. If the difference at a pixel is above a threshold, that pixel is

detected as a moving pixel. The motion direction at each moving pixel is obtained by computing optical flow [44]. 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 is uniformly divided into cells. The motion of a moving pixel is quantized into four directions. Thus each detected moving pixel is assigned a word from the codebook based on rough position and motion direction. The whole video sequence is uniformly divided into nonoverlapping short clips, e.g. each video clip lasts 10 seconds. In this approach, video clips are treated as documents, moving pixels are treated as words, and actions are treated as topics.

The LDA mixture model was used in [38] and it was extended to Dual-HDP in [43]. The LDA mixture model is shown in Figure 8 and will be explained. Suppose that a long video sequence is divided into M short video clips. Video clip j has N_j moving pixels. x_{ji} is the observed feature value of moving pixel i in video clip j. All the moving pixels are clustered into K actions. Actions are shared by all the video clips. ϕ_k is the discrete distribution of an action class over locations and moving directions. In the example shown in Figure 8, there are four action classes. Action class 1 and 3 are vehicles crossing the street intersection. Action class 2 is vehicle turning left and action class 4 is vehicles turning right. All the video clips are clustered into L global behaviors. Each global behavior c has a different prior distribution α_c over action classes. In this example, there are two global behaviors. Global behavior one has a larger weight on action class three than other action classes. Global behavior 2 has larger weights on action classes 1, 3 and 4. Each video clip j choose one of the global behaviors from a distribution η. c_j is the global behavior indicator. Video clip j samples a distribution π_j over action classes from the prior given by its global behavior. Each moving pixel i in video clips j chooses an action class from distribution π_j and sample its feature values x_{ji} from the distribution given by its action class. z_{ji} is the action class indicator. Under this model, if two motion features often co-occur in the same video clips, they have strong temporal correlation and will be grouped into the same action model. Video clips belong to the same global behavior have similar sets of co-occurring actions. This model can be used for action detection (since moving pixels are labeled as different action classes) and temporal segmentation of video sequences (since video clips are labeled as different global behaviors). If video clips and moving pixels do not fit the learned LDA mixture model, they are detected as abnormalities.

In [43], the approach was tested on a 90 minutes long video sequence taken from a traffic scene. 29 models of action classes of vehicles and pedestrians were learned. Examples of the models of action classes are shown in Figure 9. They represent the actions of vehicles crossing the intersection, vehicles turning left, vehicles turning right and pedestrians walking on crosswalks. Five types of global behaviors are learned in this scene. They correspond to different types of traffic modes. Their distributions over action classes are shown in Figure 10. Figure 11 shows the confusion matrix of assigning short video clips to different five different global behaviors compared with manually labeled ground truth. Figure 12 shows the top five detected abnormal video clips. The red color highlights the regions with abnormal motions in

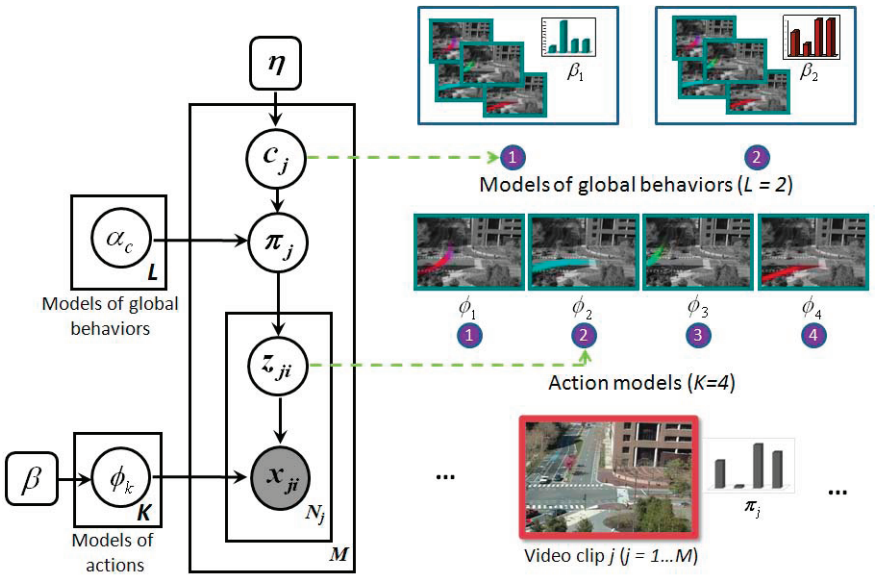
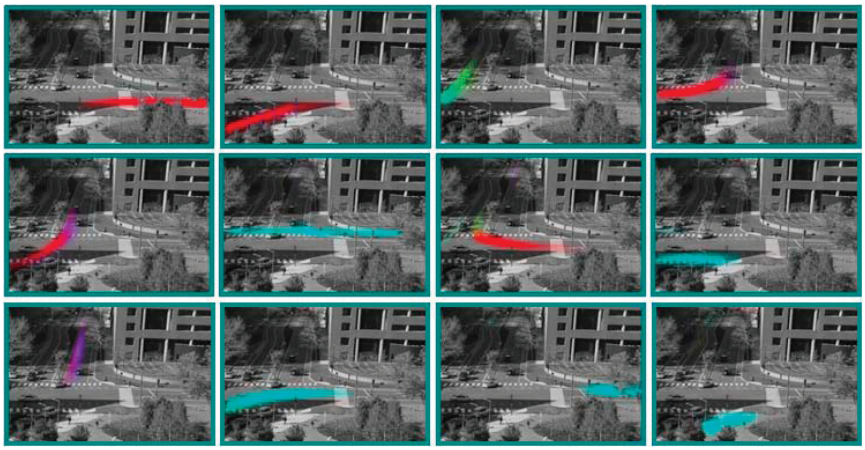


Fig. 8 Graphical model of the LDA mixture model for activity analysis based on moving pixels.

the video clips. The detected abnormal actions are pedestrians and bicycles crossing the road abnormally.

Besides [43], some other approaches of action recognition using topic models based on local motions were also proposed in recent years. Li et al. [45] first segmented the scene into different spatial regions, called semantic regions, by the spatial distribution of atomic video events in the scene. Within each semantic region, video events were clustered to extract visual words. These visual words represented how object behave locally in each region. The behavior correlations within and across the segmented semantic regions were modeled by a proposed hierarchical pLSA model. At the first stage, local behavior correlations within each region were modeled. Then the local behavior patterns were used as the input of the second stage for global behavior inference and abnormality detection. The models discussed above do not model the temporal dependency between video clips. A Markov clustering topic model was proposed in [46] to model the temporal dependency. It integrated the dynamic Bayesian network and LDA. Visual events were clustered into activities, these activities were clustered into global behaviors, and behaviors were correlated over time. In [47], a Temporal Order Sensitive LDA (TOS-LDA) model was proposed to discover behavior global correlations over a distributed camera network. TOS-LDA encoded temporal orders among visual words and could represent both long-scale behavior co-occurrences and short-scale temporal order dynamics in a single model. [48] also treated local motions and video clips as words and documents, and cluster local motions based on their co-occurrence in video clips.

(a)



(b)

Fig. 9 Examples of models of action classes learned by the approach proposed in [43] in a crowded traffic scene. (a) Background image of the traffic scene. (b) Models of action classes. Colors represent different moving directions. Their meanings are the same as in Figure 4.

However, instead of using topic models, diffusion maps embedding were used based on the measure of conditional entropy. Motion patterns were detected as different spatial and temporal scales. All these models are unsupervised and applicable to crowded scenes.

Most of the approaches discussed in this section model actions through exploring the co-occurrence of motion features. They worked well for scenes, such as traffic scenes, where at different time different subsets of activities were observed. However, they may fail in scenes where all types of actions happen simultaneously most of time with significant temporal overlaps. In this type of scenes, without tracking objects, the temporal co-occurrence information alone is not discriminative enough and the models of different action classes may not be well separated. The semantic regions learned from local motions also tend to be in short range compared with those learned from trajectories. These are the limitations of this type of approaches to be addressed in the future work.

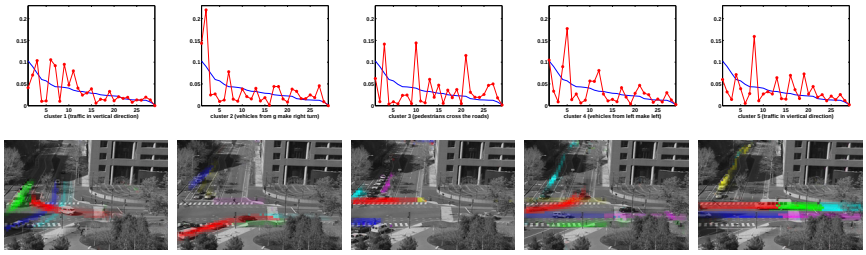


Fig. 10 The short video clips are grouped into five global behaviors. **In the first row**, we the mixtures plot $\{\pi_c\}$ over 29 action classes as prior of each global behaviors represented by red curves. For comparison, the blue curve in each plot is the average topic mixture over the whole data set. The x-axis is the index of action classes. The y-axis is the mixture over action classes. **In the second row**, we show a video clip as an example for each type of global behaviors and mark the motions of the five largest actions in that video clip. Notice that colors distinguish different action classes in the same video (the same color may correspond to different topics in different video clips) instead of representing motion directions as in Figure 9.

Manually label \ cluster by HDP model					
	149	0	2	0	0
	8	74	4	2	11
	10	3	60	1	2
	4	0	2	88	11
	4	2	6	5	92

Fig. 11 The confusion matrix of assigning short video clips into different global behaviors.

5 Near-field action recognition based on interest points

Topic models have also been used to recognize human actions in near-field settings. Niebles et al. [49] proposed an approach of extracting spatial-temporal words from space-time interest points and using pLSA to capture the co-occurrence of spatial-temporal words. A video sequence is a collection of spatial-temporal words and is treated as a document. Topic models corresponded to human action classes. This approach is applicable to moving cameras.

Space-time interest points are detected using the approach proposed in [50]. Let I be a video sequence. Gaussian filters and Gabor filters are applied to the video sequence to obtain the responses R as following,

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2.$$

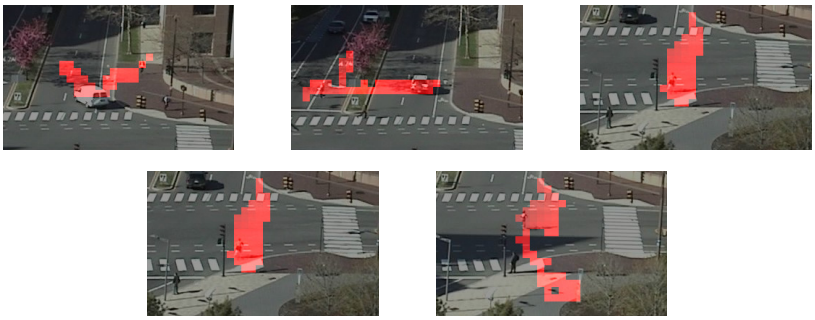


Fig. 12 Results of abnormality detection using the approach in [43]. The top five video clips with the highest abnormality (lowest likelihood) are shown. In each video clip, the regions with motions of high abnormality are highlighted.

$$h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-\frac{t^2}{\tau^2}}$$

$$h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-\frac{t^2}{\tau^2}}$$

$g(x,y;\sigma)$ is a Gaussian smoothing kernel with the standard deviation σ. It is applied to the spatial dimensions x and y. h_{ev} and h_{od} are 1D Gabor filters and are applied to the time dimension t. ω is chosen as $4/\tau$. With these filters, regions undergoing complex motions induces large responses. Therefore, space-time interest points are extracted at the local maxima of the responses. Examples of detected interest points are shown in Figure 13. Around each interest point, a visual descriptor is calcualted by concatenating the gradients into a vector. These descriptors are projected to a lower dimensional space using PCA and quantized into spatial-temporal words by k-means clustering. The model of an action class is a discrete distribution over the codebook and is learn as a topic model under pLSA.

Let $P(w|z)$ be the distribution of topic z over the codebook and its is learned from the training set. A new video sequence d_{test} is projected on the simplex spanned by the learned $P(w|z)$. The mixing coefficients $P(z|d_{test})$ is estimated by minimizing the KL divergence between the measured empirical distribution $\tilde{P}(w|d_{test})$ and $P(w|d_{test}) = \sum_{k=1}^{K} P(z=k|d_{test})P(w|z_k)$. The optimization problem is solved by the EM algorithm. In order to localize multiple actions in a single video sequence, each interest point with word value w is assigned to one the action classes k by finding the maximum posterior,

$$P(z=k|w,d_{test}) = \frac{P(w|z=k)P(z=k|d_{test})}{\sum_{l=1}^{K} P(w|z=l)P(z=l|d_{test})}.$$

In [49], this approach was tested on the KTH human motion data set [51], which includes 598 video sequences of 6 action classes. Some examples of recognized actions are shown in Figure 13. Recognition accuracies of different methods on this data set are reported in Table 1. The topic model proposed in [49] outperforms other

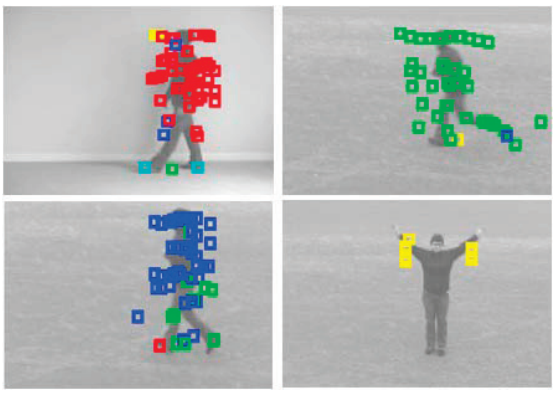


Fig. 13 Human action recognition using the approach proposed in [49]. Rectangles indicate detected interest points. Different colors represent different topics (action classes). Red: walking; blue: jogging; green: running; and yellow: hand waving. The figure is reproduced from [49].

Methods	Niebles et al. [49]	Dollar et al. [50]	Schuldt et al. [51]	Ke et al. [52]
Accuracy (%)	81.50	81.17	71.72	62.96

Table 1 Recognition accuracies of different methods on the KTH human motion data set. The table is reproduced from [49].

methods. Moreover, the method in [49] is unsupervised without requiring manually labeling training data, while other methods in comparison are supervised. The method in [49] assumes that there are multiple actions in a video sequence, while other methods assume that there is only one action in a video sequence.

The method in [49] completely ignored the geometric relationship among spatial-temporal words with the "bag-of-words" assumption. This limits its discriminative power. In [53], an approach was proposed to combine the constellation model, which captured the geometric relationship of different parts of objects, and the "bag-of-words" model. The proposed approach model human actions at different hierarchical levels. At the higher level, the human action is model as a constellation of P parts. At the lower layer, each part is associated to a bag of spatial-temporal features.

6 Further Reading

In computer vision, besides action recognition, topic models have also been applied to scene categorization [54], object recognition [55, 56] and image semantic segmentation [57, 29, 58] and image search [59]. The original topic models developed in language processing have been extended to incorporate spatial and temporal information to better solve computer vision problems [56, 29, 60, 61]. In recent years, many other topic models, such as dynamic topic models [62], author-topic model

[63], HMM-LDA [64], polylingual topic models [65], correlated topic models [66], and supervised topic model [67] were proposed. It would be also interesting to see how these models can be applied to solve computer vision problems. Topic models are hierarchical Bayesian models, which have some nice properties of jointly modeling simple and complicated actions at different hierarchical levels and effectively addressing the overfitting problem through modeling the dependency among parameters. To better understand hierarchical Bayesian models, [68] is recommended. Topic models can be well integrated with nonparametric Bayesian models, which use Dirichlet processes as priors to automatically learn the number of clusters driven by data. Both LDA and the LDA mixture model have their nonparametric versions, HDP [28] and Dual-HDP [43]. More advanced models based on HDP can be found in [69, 70, 71, 72].

7 Conclusion and Discussion

In this chapter, we introduce different types of approaches using topic models to recognize actions in different scenarios. In these approaches, words, documents and topics are mapped into different concepts under different contexts of action recognition. When the scene is sparse and objects can be well tracked, trajectories of objects are treated as documents, observations on trajectories are treated as words, and semantic regions are treated as topics. In crowded scenes where object are untrackable, short video clips are treated as documents, moving pixels are treated as words and action classes are treated as topics. In near-field action recognition, video sequences are treated as documents, spatial-temporal interest points are treated as words and action classes are treated as topics. Topic models can extended from a single camera views to multiple camera views. The models of action classes are unsupervisedly learned through exploring the identity co-occurrence or temporal co-occurrence of visual features. Topic models can also be well integrated nonparametric Bayesian models to automatically learn to the number of action classes.

As a direction of the future work, topic models need to better capture the spatial and temporal relationship of words and documents. In far-field action recognition, both trajectories and local motions used as low-level features have their limitations. It worth to explore topic models based on tracklets (fragments of trajectories) when the scenes are crowded and all types of actions happen simultaneously with significant temporal overlap. Topic models have been applied to recognize actions over a small camera network (with fewer than ten camera). However, some video surveillance applications (such as monitoring the traffic flows in large cities) require action recognition over a very large camera network with hundreds or even thousands of cameras. How to apply topic models in these scenarios is another research direction to be explored.

References

1. N. D. Bird, Masoud O., N. P. Papanikolopoulos, and A. Isaacs. Detection of loitering individuals in public transportation area. *IEEE Trans. on Intelligent Transportation Systems*, 6:167–177, 2005.
2. T. Xiang and S. Gong. Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67:21–51, 2006.
3. D. Ayers and M. Shah. Monitoring human behavior from video taken in an office environment. *Image and Vision Computing*, 19:833–846, 2001.
4. J. Dever, N. V. Lobo, and M. Shah. Automatic visual recognition of armed robbery. In *Proc. Int'l Conf. Pattern Recognition*, 2002.
5. A. Datta, M. Shah, N. Da, and V. Lobo. Person-on-person violence detection in video data. In *Proc. Int'l Conf. Pattern Recognition*, 2002.
6. H. Veeraraghavan, O. Maoud, and N. Papanikolopoulos. Computer vision algorithms for intersection monitoring. *IEEE Trans. on Intelligent Transportation Systems*, 4:78–89, 2003.
7. P. Kumar, S. Ranganath, W. Hu, and K. Sengupta. Framework for real-time behavior interpretation from traffic video. *IEEE Trans. on Intelligent Transportation Systems*, 6:43–53, 2005.
8. R. Khoshabeh, T. Gandhi, and M. M. Trivedi. Multi-camera based traffic flow characterization and classification. In *Proc. IEEE Conf. Intelligent Transporation Systems*, 2007.
9. S. Atev, H. Arumugam, O. Masaoud, R. Janardan, and N. P. Papanikolopoulos. A vision-based approach to collision prediction at traffic intersections. *IEEE Trans. on Intelligent Transportation Systems*, 6:416–423, 2005.
10. C. Lin and Z. Ling. Automatic fall incident detection in compressed video for intelligent homecare. In *Proc. IEEE Int'l Conf. Computer Communications and Networks*, 2007.
11. W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, Man, Cybernetics-Part C: Applications and Reviews*, 34:334–352, 2004.
12. B. T. Morris and M. M. Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Trans. on Circuits and Systems for Video Technology*, 18:1114–1127, 2008.
13. S. S. Intille and A. F. Bobick. A framework for recognizing multi-agent action from visual evidence. In *Proc. National Conf. Artificial Intelligence*, 1999.
14. S. Honggeng and R. Nevatia. Multi-agent event recognition. In *Proc. Int'l Conf. Computer Vision*, 2001.
15. Y. Ke, R. Suckthanlar, and M. Hebert. Event detection in crowded videos. In *Proc. Int'l Conf. Computer Vision*, 2007.
16. L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2001.
17. H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2004.
18. Y. Wang, T. Jiang, M. S. Drew, Z. Li, and G. Mori. Unsupervised discovery of action classes. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2006.
19. J. W. Davis and A. F. Bobick. The representation and recognition of action using temporal templates. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 1997.
20. P. Smith, N. V. Lobo, and M. Shah. Temporalboost for event recognition. In *Proc. Int'l Conf. Computer Vision*, 2005.
21. N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Trans. on PAMI*, 22:831–843, 2000.
22. C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on PAMI*, 22:747–757, 2000.
23. X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *Proc. European Conf. Computer Vision*, 2006.
24. N. Ghanem, D. Dementhon, D. Doermann, and L. Davis. Representation and recognition of events in surveillance video using petri net. In *CVPR Workshop*, 2004.

25. T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence*, 1999.
26. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
27. T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230m, 1973.
28. Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet process. *Journal of the American Statistical Association*, 2006.
29. X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *Proc. Neural Information Processing Systems Conf.*, 2007.
30. L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2005.
31. D. M. Blei and M. I. Jordan. Variational methods for the dirichlet process. In *Proc. Int'l Conf. Machine Learning*, 2004.
32. D. M. Blei and M. I. Jordan. Variational inference for dirichlet process mixtures. *ba*, 1:121–144, 2005.
33. M. Jordan. *Learning in Graphical Models*. MIT Press, 1999.
34. T. L. Griffiths and M. Steyvers. Finding scientific topics. In *Proc. of the National Academy of Sciences of the United States of America*, 2004.
35. D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed inference for latent dirichlet allocation. In *Proc. Neural Information Processing Systems Conf.*, 2007.
36. W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *IEEE Trans. on PAMI*, 28:1450–1464, 2006.
37. X. Wang, K. T. Ma, G. Ng, and E. Grimson. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2008.
38. X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception by hierarchical bayesian models. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2007.
39. X. Wang, K. Tieu, and E. Grimson. Correspondence-free multi-camera activity analysis and scene modeling. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2008.
40. X. Wang, K. Tieu, and E. Grimson. Correspondence-free activity analysis and scene modeling in multiple camera views. *IEEE Trans. on PAMI*, 2009.
41. L. Lee, R. Romano, and G. Stein. Monitoring activities from multiple video streams: Establishing a common coordinate frame. *IEEE Trans. on PAMI*, 22:758–768, 2000.
42. D. Thirde, M. Borg, J. Ferryman, J. Aguilera, and M. Kampel. Distributed multi-camera surveillance for aircraft servicing operations. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2005.
43. X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Trans. on PAMI*, 31, 2009.
44. B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *Proc. Int'l Joint Conf. Artificial Intelligence*, pages 674–680, 1981.
45. J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. In *Proc. European Conf. Computer Vision*, 2008.
46. T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In *Proc. Int'l Conf. Computer Vision*, 2009.
47. J. Li, S. Gong, and T. Xiang. Discovering multi-camera behaviour correlations for on-the-fly global activity prediction and anomaly detection. In *Proc. of IEEE Int'l Workshop on Visual Surveillance*, 2009.
48. Y. Yang, J. Liu, and M. Shah. Video scene understanding using multi-scale analysis. In *Proc. Int'l Conf. Computer Vision*, 2009.
49. J. C. Niebles, H. Wang, and F. Li. Unsupervised learning of human action categories using spatial-temporal words. In *Proc. British Machine Vision Conference*, 2006.
50. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. 2005.

51. C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Proc. Int'l Conf. Pattern Recognition*, 2004.
52. Y. Ke, R. Sukthankar, and M. Hebert. Hebert. In *Proc. Int'l Conf. Computer Vision*, 2005.
53. J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2007.
54. L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2005.
55. J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proc. Int'l Conf. Computer Vision*, 2005.
56. E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, 77:291–330, 2007.
57. B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2006.
58. L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proc. Int'l Conf. Computer Vision*, 2007.
59. K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. M. Blei, and M. J. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.
60. J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2007.
61. G. Passino, I. Patras, and E. Izquierdo. Latent semantics local distribution for crf-based image semantic segmentation. In *Proc. British Machine Vision Conference*, 2009.
62. D.M. Blei and J.D. Lafferty. Dynamic topic models. In *Proc. Int'l Conf. Machine Learning*, 2006.
63. M. Rosen-Zvi, T. Griffths, M. Steyvers, and P. Smyth. Probabilistic author-topic models for information discovery. In *Proc. of ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2004.
64. T.L. Griffths, M. Steyvers, D.M. Blei, and J.B. Tenenbaum. Integrating topics and syntax. In *Proc. Neural Information Processing Systems Conf.*, 2004.
65. D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. MaCallum. Polylingual topic models. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2009.
66. D.M. Blei and J.D. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1:17–35, 2007.
67. D. M. Blei and J. D. McAuliffe. Supervised topic models. In *Proc. Neural Information Processing Systems Conf.*, 2007.
68. A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2004.
69. N. Srebro and S. Roweis. Time-varying topic models using dependent dirichlet processes. Technical report, Department of Computer Science, University of Toronto, 2005.
70. L. Ren, D. B. Dunson, and L. Carin. The dynamic hierarchical dirichlet process. In *Proc. Int'l Conf. Machine Learning*, 2008.
71. A. Rodriguez, D. B. Dunson, and A. E. Gelfand. The nested dirichlet process. Technical report, Working Paper 2006-19, Duke Institute of Statistics and Decision Sciences., 2006.
72. J. E. Griffin and M. F. J. Steel. Order-based dependent dirichlet processes. *Journal of the American Statistical Association*, 101:179–194, 2006.