J Inf Process Syst, Vol.16, No.3, pp.612~628, June 2020
https://doi.org/10.3745/JIPS.02.0134

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Abnormal Behavior Recognition Based on Spatio-temporal Context

Yuanfeng Yang*, Lin Li**, Zhaobin Liu***, and Gang Liu***

Abstract

This paper presents a new approach for detecting abnormal behaviors in complex surveillance scenes where anomalies are subtle and difficult to distinguish due to the intricate correlations among multiple objects' behaviors. Specifically, a cascaded probabilistic topic model was put forward for learning the spatial context of local behavior and the temporal context of global behavior in two different stages. In the first stage of topic modeling, unlike the existing approaches using either optical flows or complete trajectories, spatio-temporal correlations between the trajectory fragments in video clips were modeled by the latent Dirichlet allocation (LDA) topic model based on Markov random fields to obtain the spatial context of local behavior in each video clip. The local behavior topic categories were then obtained by exploiting the spectral clustering algorithm. Based on the construction of a dictionary through the process of local behavior topic clustering, the second phase of the LDA topic model learns the correlations of global behaviors and temporal context. In particular, an abnormal behavior recognition method was developed based on the learned spatio-temporal context of behaviors. The specific identification method adopts a top-down strategy and consists of two stages: anomaly recognition of video clip and anomalous behavior recognition within each video clip. Evaluation was performed using the validity of spatio-temporal context learning for local behavior topics and abnormal behavior recognition. Furthermore, the performance of the proposed approach in abnormal behavior recognition improved effectively and significantly in complex surveillance scenes.

Keywords

Abnormal Behavior Recognition, Cascade Model, Spatio-temporal Context, Topic Model

1. Introduction

In dynamic surveillance scenes, especially in the case of complex interactive behaviors among multiple moving objects, the traffic abnormal behavior recognition of the vehicle is a very challenging problem in the field of computer vision. According to the behavioral features used, the current existing approaches for traffic scene analysis can be divided into two categories: motion trajectory-based methods and local motion feature vector-based methods. Most of the reported approaches are based on the trajectory analysis of the objects. This type of methods first learns the trajectory patterns of moving objects to establish behavioral models, and then the trajectories of each moving object are matched with the learned behavior models. If the differences between the trajectories and behavioral models exceed the threshold,

Manuscript received December 11, 2018; first revision February 27, 2019; second revision April 16, 2018; accepted May 8, 2019.

Corresponding Author: Yuanfeng Yang (yfyangsz@hotmail.com)

* Jiangsu Province Support Software Engineering R&D Center for Modern Information Technology Application in Enterprise, Suzhou, China (yfyangsz@hotmail.com)

** School of Computer and Information Engineering, Xiamen University of Technology, Xiamen, China (linli@xmut.edu.cn)

*** School of Computer Engineering, Suzhou Vocational College, Suzhou, China ({liuzhaobin, liugang}@jssvc.edu.cn)

the behaviors corresponding to those trajectories will be considered abnormal behaviors.

There have been many research studies on such trajectory-based abnormal behavior recognition methods. Commonly used methods are decision tree [1], hidden Markov model [2,3], neural network [4], support vector machine [5], Bayesian [6], etc. Nonetheless, the quality of these methods is highly dependent on the robust tracking of objects, which is inherently difficult in general due to the complexity of surveillance scenes, change of illumination, and occlusion between objects. Methods based on the local motion feature vector [7-10] need not acquire the trajectories of moving objects, directly describing the video clip using the motion feature vector. Zelnik-Manor and Irani [7] constructed a similarity matrix by the distance measurement method based on the multiple time-scale features of each video clip. The model of video clip behaviors was then automatically established by spectral clustering. Zhong et al. [8] also divided video sequences into clips. By treating the video clip as a document, each video frame is treated as a word, and clustering of video clip was consequently converted into document clustering. The methods above could not identify multiple behaviors in the same video clip, as they are applicable only to simple data sets with only one behavior in the video clip. Similarly, some studies also modeled video clip as an indivisible whole, which could only determine whether the entire video clip had abnormal behavior [9,10].

Due to the complexity of the surveillance scene, however, many types of abnormal behavior recognition make sense only when considering the spatio-temporal context of behaviors, i.e., the behavioral relevance of different moving objects. Fig. 1 shows a typical anomalous behavior that can be identified only by considering the spatio-temporal context of behaviors in the scene.

Fig. 1(a) shows the monitoring scene of normal traffic behavior. The motion of the fire engine in Fig. 1(b) can be identified as normal traffic behavior if considered in isolation. Considering the spatio-temporal context of behaviors in the scene, however, the horizontal motion of the fire engine interrupts the vertical traffic flow, which should be identified as anomalous behavior.

(a) (b)

Fig. 1. Typical abnormal behavior: monitoring scenes of (a) normal behavior and (b) abnormal behavior.

Many scholars have done useful explorations on the analysis of interactive behaviors among multiple moving objects. Xiang and Gong [11] proposed a method for modeling the interactive behaviors among multiple moving objects based on the BIC (Bayesian Information Criterion) framework model. The continuity studies in [12] showed that the model can perform abnormal behavior detection and normal behavior recognition in a short time. Nonetheless, this method of modeling behaviors was limited to a small set of discrete events in a local region, and it did not implement global behavior modeling in the true sense.

Rand and Kettnaker [13] used MOMC-HMM (multi-observation-mixture+counter hidden Markov model) to model the behaviors in the surveillance scenes, and they did not consider the behavioral relevance in the global scope. Wang et al. [14] proposed the hierarchical Bayesian model modeling interactive behaviors. These three elements (bottom optical flow field characteristics, middle atomic activity, and high-level interaction behavior) were connected by such hierarchical Bayesian model. In another study [15], three levels of video events were connected by the hierarchical Dirichlet process (HDP) model: low-level visual features, simple atomic activities, and multi-agent interactions. By combining generative models (HDP models) and discriminative ones (GP models), the HDP models learned the activity patterns in an unsupervised manner, and the GP models accomplished activity recognition and anomaly detection. Still, the two methods make it difficult for the model to extend other different types of features for modeling the correlation of behaviors. If more features are added to the model, the complexity of the model will be greatly improved.

In recent years, topic models have been used in the field of computer vision such as image segmentation, object detection, scene understanding, image annotation, etc. In the field of motion object behavior analysis and understanding research, Wang et al. [16,17] regarded the trajectory as a document, and the trajectory points were regarded as words in the document. The dual-HDP method for modeling a semantic scene was proposed to analyze the moving objects' behaviors. Nonetheless, this method required complete trajectories to model, which limited the application in cases wherein complete trajectories could not be acquired. Zhou et al. [18] used the random field topic model to cluster the motion trajectories of objects for analyzing the semantic regions where the object motion direction was consistent. In the process of modeling, the correlation between trajectory fragments was established according to the multiple spanning trees, which increased the complexity of the model. All of these methods still could not solve the problems raised in this paper. Kaviani et al. [19] presented an unsupervised approach based on fully sparse topic models (FSTM) to model activities and interactions in complex scenes of traffic video. This method first temporally segmented the video into non-overlapping clips, which were considered as documents. The optical flow extracted from each pair of consecutive frames was then quantified as words according to the position and motion direction. Similar to literature [9,10], this method could only judge whether all video clips had abnormal behavior.

The basic topic model has been upgraded in a more straightforward manner. A two-level motion pattern mining approach [20] was used to learn behaviors in a dynamic scene. The first-level LDA (latent Dirichlet allocation) learned single-agent motion patterns, whereas the second-level LDA used the single-agent motion patterns as words to learn interaction patterns. This hierarchy enabled interaction pattern detection for every video frame rather than for clips. Nonetheless, this method also required complete trajectories to model. Similar to the cascaded topic model structure proposed in this paper, Li et al. [21,22] first used a semantic scene segmentation model to segment the surveillance scene into multiple regions. Each region used PLSA (probability latent semantic analysis) to learn local behaviors in the region. The two-level hierarchical PLSA model was then used to model cross-region interactions. Loy et al. [23] decomposed complex global behavior patterns based on temporal features or spatio-temporal visual context (there were also steps to segment the scene into multiple regions), with the decomposed behaviors using cascaded dynamic Bayesian networks to perform modeling, global behavioral reasoning, and abnormal behavior detection. The methods above all needed to segment the scene (considered an image segmentation problem) as the basis of subsequent behavior modeling. The interactive modeling of global behaviors was embodied

in the construction and analysis of the co-occurrence matrix of local behavioral topics across regions. Moreover, the assumption that there was correlation between the local behaviors across regions was based on the segmentation of the region, which actually limited the contextual learning of behaviors.

As for the main contributions of our work, (1) we propose an abnormal behavior recognition method based on the cascaded topic model in complex surveillance scenes, which has obvious advantages in simplifying complex global behavior modeling. (2) In the first phase of the cascaded structure, the trajectory segments within the video clip are treated as a single document for topic modeling, which avoids the situation wherein complete trajectories cannot be acquired in complex scenes. At the same time, there are no defects using the local motion feature vector for analyzing the behaviors of moving objects in complex scenes without object detection and tracking. (3) The abnormal behavior recognition method adopts a top-down strategy, which not only has the ability to recognize different types of abnormal behavior but can also recognize anomalies in the case wherein there are complex interactions among multiple moving objects' behaviors at the same time.

The rest of this paper is organized as follows: Section 2 presents the overall framework of the cascaded topic model; Section 3 details the topic model modeling of the two phases as well as how to combine spatio-temporal context learning for behavior; Section 4 presents the specific abnormal behavior recognition method; Section 5 discusses the experimental results; finally, we conclude our work in Section 6.

2. Overall Framework of the Cascaded Topic Model

In the research on vehicle behavior analysis, we find that, when the topic model is used to model the trajectory in the surveillance scenes, the topic represents the semantic region shared between the trajectories. The same type of behavior should go through the same combination of semantic regions, and it has a prior distribution of semantic regions. These trajectories, which are clustered into the same behavior category, share such prior distribution. This type of property is manifested as the spatial similarity of behavior. Blei and Lafferty [24] believed that the topic evolved along the time axis and satisfied the first-order Markov hypothesis. Based on this hypothesis, in 2006, dynamic topic models (DTM) was proposed. Such properties are manifested in the behavioral analysis as the limitations of time. In other words, a certain type of behavior can only be combined through a specific semantic region within a certain period of time; in another period of time, the topic (semantic region) will evolve, and the behaviors will also change. In the field of traffic monitoring, surveillance video can be segmented into different video clips according to traffic signals, and each video clip has different moving object behavior. When different video clips use the topic model to model the behaviors of moving objects, the topics learned are different, and they have a different prior distribution. This is consistent with the spatio-temporal characteristics of behaviors. In addition, standard topic models cannot simultaneously model and differentiate between local behavior (within video clip) and global behavior (cross-video clips). Therefore, this section considers the decomposition of complex global behaviors in a hierarchical structure in spatio-temporal organization mode.

As shown in Fig. 2, the global behavior pattern analysis method based on the cascaded topic model proposed in this section mainly adopts a two-stage topic model, and it is organized in a cascaded manner. The first stage of the cascaded topic model obtains the spatial context of local behavior through the inference of implicit local behavior topics within each video clip. Specifically, due to the complexity of

the scene and failure of tracking, the trajectory of the same moving object is divided into multiple trajectory segments. This paper regards a trajectory segment in a video clip as a document. The trajectory points in the trajectory segment are quantized into motion words according to the position of the rectangular region and direction of motion following discretization. The concept correspondence relationship is described in [16]. All trajectory segments within a video clip form a corpus of local behavioral learning.

The second phase of topic modeling is used to learn the temporal context of global behavior. A video clip in this stage corresponds to a document, and all video clips within the surveillance video form a corpus of global behavioral learning. The local behavioral topics (semantic regions) inferred from the first-stage topic modeling build a dictionary through the process of local topic clustering. Each type of local behavior topic corresponds to a word, and the size of the dictionary is the number of local behavior topic categories. The local behavioral topics within each video clip are labeled as categories of local behavioral topics, and the video clips are transformed into "documents" composed of words.

Tracking failures in complex scenes are unavoidable due to the noise and errors of the underlying visual features. By constructing a cascaded topic model structure, each phase can take advantage of the inference results of the previous phase, and topic modeling in the second phase of the cascaded structure will greatly reduce the effects of noise and errors of the underlying visual features. In addition, a single complex model usually has scalability problems [14]; the cascade structure can largely avoid such problems by decomposing complex global behaviors.

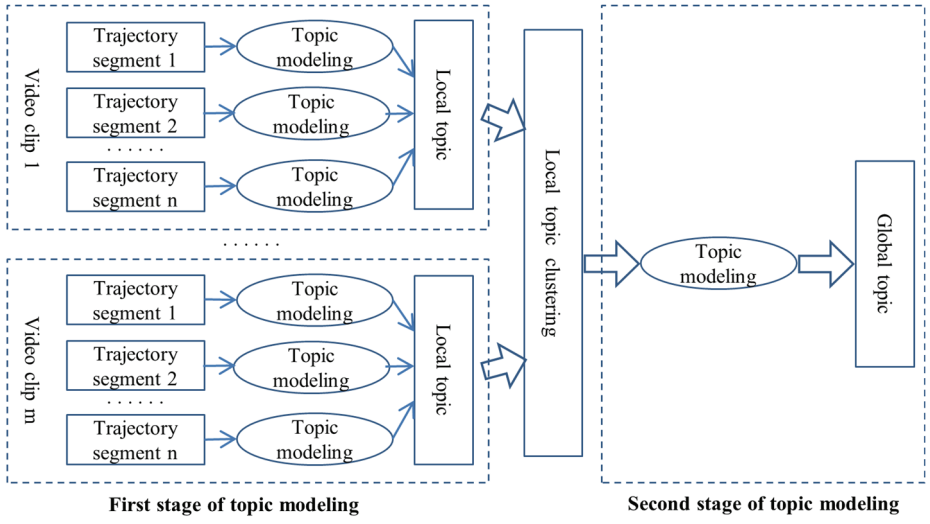


Fig. 2. Structure of cascaded topic model.

3. Behavioral Spatio-temporal Context Learning

3.1 Local Behavioral Pattern Learning

Given a section of traffic video $V = \{v_1, v_2, \cdots, v_T\}$, V can be divided into t-segment video clips, $1 \le t \le T$. The trajectories generated by the moving objects within each video clip contain trajectory

segments due to factors such as complexity of the scene and failure of tracking. These trajectory segments and video clips are treated as documents to train the cascaded topic models during different topic modeling phases. In the first stage of topic modeling, a trajectory segment within a video clip is treated as a document. The trajectory points in the trajectory segment are quantized into motion words according to the position of the rectangular area where the trajectory points are located and the motion direction following discretization. Each trajectory segment is represented as a random mixture of *K* local behavior topics, where *K* represents the number of local behavior topics (semantic regions in the scene within the video clip). The local behavior topics are essentially semantic regions passed by the trajectory segments. Since multiple trajectory segments may belong to the same moving object, the standard topic model cannot model the relationship between documents. Thus, a MRF-LDA (Markov random fields - latent Dirichlet allocation) model is proposed [18] as shown in Fig. 3. MRF is used to model the relationship between documents.

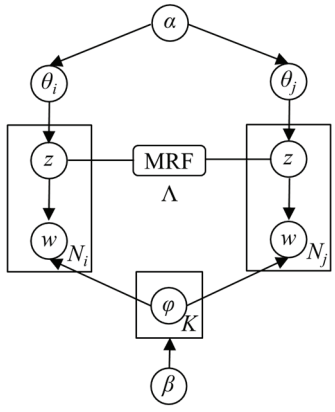


Fig. 3. MRF-LDA graph model.

Λ describes the MRF connection of adjacent trajectory segments, and $\varepsilon(i)$ is defined as the set of trajectory segments close to trajectory segment TS_i. Such can be viewed as MRF structure. Trajectory segment TS_i is represented by triple variables (x_u^i, y_u^i, t_u^i) representing the two-dimensional spatial position point of moving object t_u^i at time (x_u^i, y_u^i). Therefore, trajectory segment TS_i can be formally represented as $\{(x_p^i, y_p^i, t_p^i), (x_{p+1}^i, y_{p+1}^i, t_{p+1}^i), \cdots, (x_q^i, y_q^i, t_q^i)\}(t_p^i \prec t_q^i)$. In other words, trajectory segment TS_i starts at moment t_p^i and ends at moment t_q^i. The starting and ending positions are (x_p^i, y_p^i) and (x_q^i, y_q^i), respectively, and the speeds at the two positions are $v_p^i = (v_p^{ix}, v_p^{iy})$ and $v_q^i = (v_q^{ix}, v_q^{iy})$, respectively.

If trajectory segment TS_j satisfies Eqs. (1), (2), and (3), it is then considered to be associated with TS_j, $j \in \varepsilon(i)$.

$$t_q^i \prec t_p^j \prec t_q^i + \Delta t \tag{1}$$

$$\left|x_q^i - x_p^j\right| + \left|y_q^i - y_p^j\right| \prec \Delta s \tag{2}$$

$$\frac{V_q^i \bullet V_p^j}{\left\|V_q^i\right\| \left\|V_p^j\right\|} \succ c \tag{3}$$

The equations above indicate that associated trajectory segments TS_j and TS_i are temporally and spatially close, maintaining a consistent direction of motion. This paper considers trajectory segments that may belong to the same moving object to be associated. For example, Eq. (1) indicates that a pair of trajectory segments overlapping in time may not belong to the same moving object; thus, the pair of trajectory segments is not associated. Eqs. (2) and (3) illustrate the adjacency of the spatial position and consistency of the motion direction, respectively. If the conditions above are met, and $z_{in_1} = z_{jn_2}$, then the MRF connection is defined as Equation (4).

$$\Lambda(z_{in_1}, z_{jn_2}) = \exp\left(\frac{V_q^i \bullet V_p^j}{\left\|V_q^i\right\|\left\|V_p^j\right\|} - 1\right) \tag{4}$$

Otherwise, $\Lambda(z_{in_1}, z_{jn_2}) = 0$.

In the first stage of topic modeling, each trajectory segment is modeled using MRF-LDA. Each topic z is modeled as polynomial distribution $\varphi_k = [\varphi_{k1}, \varphi_{k2}, \cdots, \varphi_{kV}]$ in the motion dictionary, i.e., the mixture ratio $\varphi \sim Dirichlet(\beta)$ of various motion words. Polynomial distribution $\theta_i = [\theta_{i1}, \theta_{i2}, \cdots, \theta_{iK}]$ over K topics is generated by the Dirichlet distribution $Dirichlet(\theta_i|\alpha)$. For each motion word w_j on trajectory TS_i, topic $z_{ij} = k$ is determined by probability parameter θ_{ik} , and $\varphi_{z_{ij}}$ determines the generation of motion word w_{ij} .

Given α and β, the joint probability distribution of topic mixture θ_i, motion word mixture φ , topic variable $z_i = \{z_{ij}\}$, and motion words $w_i = \{w_{ij}\}$ is shown as Eq. (5).

$$p\left(\theta_i, z_i, \varphi, w_i \middle| \alpha, \beta\right) = p\left(\theta_i \middle| \alpha\right) p\left(\varphi \middle| \beta\right) \prod_{j=1}^{N_i} p\left(z_{ij} \middle| \theta_i\right) p\left(w_{ij} \middle| z_{ij}, \varphi\right) \tag{5}$$

N_i is the number of motion words on trajectory TS_i. According to the properties of the topic model, the terms co-occurring in the document are classified into the same topic. In other words, if two position points in the scene are connected by multiple trajectory segments, the two position points will belong to the same semantic region, and $p\left(z_{ij} \middle| \theta_i\right)$ will be defined by Eq. (6).

$$p\left(Z \middle| \theta\right) \propto \exp\left(\sum_i \log\theta_i + \sum_{j \in \varepsilon(i)} \sum_{n_1, n_2} \Lambda(z_{in_1}, z_{jn_2})\right) \tag{6}$$

The Gibbs Sampling derivation formula is shown as Eq. (7).

$$\begin{aligned} & p\left(z_{ij} \middle| w, z_{\neg ij}, \alpha, \beta\right) \\ & \propto \frac{n_{k,\neg ij}^v + \beta}{\sum_{v=1}^V (n_{k,\neg ij}^w + \beta)} \frac{n_{i,\neg j}^k + \alpha}{\sum_{k=1}^K (n_{i,\neg j}^k + \alpha)} \exp\left(\sum_{j \in \varepsilon(i)} \sum_{n_1, n_2} \Lambda(z_{in_1}, z_{jn_2})\right) \end{aligned} \tag{7}$$

3.2 Local Behavioral Topic Clustering

After learning local behavior patterns (the first phase of topic modeling), the spatial context of local behavior within each video clip can already be inferred. All similar local behaviors within the same video clip share the same topic. Although different video clips may have similar local behaviors, they do not share topics across video clips. That is because the topics learned within different video clips are not the

same even if the local behaviors are similar. Therefore, the inference results of the first-stage topic modeling cannot be directly outputted to the second stage of topic modeling but need to be pre-processed before being inputted to the next stage.

The essence of the temporal context of global behavior is the evolution of local behavior within the video clip on the time axis. Because of the deviation of the local behavioral topics learned in each video clip, this study clustered these local behavioral topics prior to global behavioral pattern learning so that (from the perspective of the entire traffic video sample) the behavioral local topic is fixed and unified. What is changed is the distribution of local behavior topic categories within each video clip. Therefore, in order to learn the global behavior patterns, the clustering of local behavior topics is first completed, and the local behavior topics in each video clip are marked as categories of local behavior topics.

In this study, the spectral clustering algorithm was used to cluster local behavioral topics with high spatial similarity so that the global behavior patterns of moving objects can be learned through local behavioral topics. The local behavior topic vector is represented by distribution $p\left(w_v \middle| z_i\right)$ of topics in the V-dimensional word space, and the similarity between the local behavior topic vectors is calculated by Equation (8).

$$TopSim\left(z_i \middle| z_j\right) = TopSim\left(\beta_i \middle| \beta_j\right) = \frac{\sum_{v=1}^{V} \beta_{iv} \bullet \beta_{jv}}{\sqrt{\sum_{v=1}^{V} (\beta_{iv})^2 (\beta_{jv})^2}} \tag{8}$$

The measure of similarity between local topic vectors disregards the time factor and considers only the spatial position and direction of motion of the moving objects, which is determined by the construction process of the motion dictionary.

We define local behavior topic vector set $TopVec = \left\{z_1^1, \cdots, z_1^{K_1}, z_2^1, \cdots, z_2^{K_2}, \cdots, z_T^1, \cdots, z_T^{K_T}\right\}$. z_t^k is represented as the k-th local topic vector learned in the t-th video clip. The spectral clustering algorithm first calculates the similarity between any two vectors in local behavior topic vector set $TopVec$ and constructs similarity matrix $A \in R^{N \times N}$. Matrix element $A_{ij} = TopSim\left(z_i \middle| z_j\right)$; when $i = j$, $A_{ii} = 0$. After the Laplacian matrix is constructed according to the similarity matrix, the eigenvalues and eigenvectors of the Laplacian matrix can be calculated. Finally, the appropriate eigenvectors are selected to cluster different local behavioral topics.

3.3 Global Behavioral Pattern Learning

The second phase of topic modeling uses the LDA topic model [25] to learn the temporal context of global behavior. The process of local behavior topic clustering constructs the dictionary of this stage. Each word in the dictionary corresponds to the index of a kind of local behavior topics. The size of the dictionary is the number C of local behavior topic categories. All video clips within the surveillance video form a corpus of global behavioral learning. The global behavior topics learned in this stage correspond to the temporal context structure of behaviors.

Since global behavior pattern learning only cares about the co-occurrence of each type of local behavior topic rather than the co-occurrence frequency, the document of the LDA topic model in this stage is represented as a binary C-dimensional (dictionary size) feature vector wherein each binary value element (0 or 1) indicates whether the corresponding word (a type of local behavior topic) exists in the document (video clip).

Given a piece of traffic video sample $V = \{v_1, v_2, \cdots, v_T\}$, V is split into t-segment video clips ($1 \le t \le T$) constituting corpus $D = \{v_j\}$, $1 \le j \le T$. Assuming K-type global correlation behaviors (corresponding to K-type global behavior topics), the polynomial distribution parameters that need to be learned are the $K \times C$-dimensional matrix modeled in the dictionary of this stage. That represents the mixture ratio of various words on each global behavior topic, which is represented as $\varphi_{k,c} = p(w_c | z_k)$ and $\sum_{c=1}^{C} \varphi_{k,c} = 1$. Its essential meaning is that the co-occurrence probability of the C-type local behavior topic in the video clip corresponds to the context information of the global behavior.

Although the input to the model in this stage is a binary feature vector rather than a count of words, the sampling process of words has not been changed. The generation process of word w_{ji} on the corresponding video clip v_j is as follows:

1) For each word w_{ji}, sample its corresponding topic type z_{ji} : $z_{ji} \sim Multinomial(\theta_j)$.

2) Determine motion words w_{ji} from the dictionary by conditional probability $p\left(w_{ji} | z_{ji}, \varphi_{z_{ji}}\right)$.

After two stages of topic model training, the cascaded LDA topic model can be used to interpret behavioral patterns (local behavior patterns and global behavior patterns) in test videos. The local behavior patterns reveal the spatial context of local behaviors within the video clip, with the global behavior patterns demonstrating the relevance or co-occurrence of local behaviors in the temporal context.

4. Abnormal Behavior Recognition

Through the cascaded LDA topic model described above, local behavior patterns and global behavior patterns are learned, and complex global behaviors can be decomposed according to the spatio-temporal characteristics of the behaviors. In the case wherein multiple moving objects occur simultaneously and there are interactive behaviors (such as traffic intersections), the key is to know when and where the abnormal behaviors will occur. At this time, the spatio-temporal context information of the learned behaviors cannot recognize and explain the occurrence of abnormal behavior due to the lack of location information of the topics (or words). Based on the spatio-temporal context information of behaviors, this section proposes a new anomalous behavior recognition method. The method not only recognizes video clips with abnormal behaviors but can also locate the moving object's trajectories that cause abnormal behaviors within the video clip. In particular, the cascaded LDA topic model corresponds to the identification of anomalous video clips and anomalous behaviors of moving objects through two-stage topic modeling. The specific identification method adopts a top-down strategy, and it is divided into two stages of identification.

The first stage performs the recognition of abnormal video clips. Given a piece of traffic video test sample consisting of t non-overlapping video clips, each video clip is treated as a document to be checked for abnormal behaviors during the second phase of topic modeling. In the training phase, the corresponding LDA topic models need to be trained independently for each type of video clip. After Gibbs Sampling converges, the corresponding $\hat{\theta}_l$ and $\hat{\varphi}_l$ are obtained statistically. For the tested video clips, the reasoning process and the training process of the topic modeling are basically similar, and $\hat{\varphi}_l$ in the Gibbs Sampling formula is considered to remain stable and is provided by the topic model during the

training phase. During the sampling process, only topic distribution θ_{test} of the video clip needs to be estimated. Then, for video clips of category l, the likelihood values of the topic distribution in the tested video clip are calculated by Eq. (9).

$$ClipSim\left(v_{test}\middle|v_l\right)=ClipSim\left(\theta_{test}\middle|\hat{\theta}_l\right)=\frac{\theta_{test}\bullet\hat{\theta}_l}{\sqrt{(\theta_{test})^2(\hat{\theta}_l)^2}} \tag{9}$$

The lower the likelihood value is, the higher the likelihood of anomalous behavior in the video clip. The video clip abnormality scoring function is defined as Eq. (10).

$$abf=\arg\max_{L} ClipSim\left(v_{test}\middle|v_l\right) \tag{10}$$

where L represents a collection of video clip categories, $l\in L$. If video clip abnormality score *abf* is lower than threshold *TH*, it is judged that there is abnormal behavior in the video clip.

The second stage performs anomalous behavior recognition within the video clip. Once video clip v_{test} is recognized as abnormal, the abnormal trajectory of the moving object begins to be recognized. Specifically, the top-down model first determines the abnormal words (local behavior topic categories) in the video clip through the second-stage topic model, and then uses the first-stage topic model to locate anomalous moving object trajectories. The main steps for determining anomalous words are described below.

1) Word w_i ($1\le i\le n$, n is the number of words in the video clip) is first removed in turn, so n new video clips $v^*_{\neg i}$ are obtained.
2) Then, word w_i is saved into the candidate abnormal word set for the corresponding video clip, and abnormality score *abf* for each video clip is calculated simultaneously.
3) If all the abnormality scores are lower than threshold *TH*, step 1) is performed. Otherwise, the following steps are performed:
   (a) If the video clip whose abnormality score is higher than threshold *TH* is unique, it can be judged that the words in the corresponding abnormal word set are abnormal words.
   (b) Otherwise, the word in the abnormal word set corresponding to the highest abnormality score is judged as abnormal word.

The abnormal words determined at this time substantially correspond to the local behavior topic categories obtained by the clustering of the local behavior topics learned in the first-stage topic modeling. By calculating the similarity between the local behavior topic vector in the video clip and such abnormal local behavior topic vector by Eq. (8), the abnormal local behavior topic can be easily determined.

In the first stage of topic modeling, distribution θ_i of topics in each trajectory (trajectory segment) can be estimated. If the trajectory (or trajectory segment) generated by the moving object in the abnormal video clip contains an abnormal local behavior topic, the trajectory can be considered to be abnormal.

5. Experiments and Analysis

5.1 Dataset

In order to test and verify the effectiveness of the vehicle abnormal behavior recognition method based on the spatio-temporal context proposed in this paper, QMUL Street Intersection Dataset [26] was

employed in our experiments. This data set contained 45 minutes of 25 fps video of a busy street intersection. There are four types of traffic flow patterns controlled by traffic lights at traffic intersections as shown in Fig. 4. Traffic flow pattern A represents two opposite vertical traffic flows. Traffic flow pattern B is a traffic flow wherein two opposite vertical traffic flows turn left and right. Traffic flow patterns C and D indicate the directions of left and right traffic flows, respectively. The order of traffic flow patterns occurring depends on how busy the vertical traffic flow pattern is. Traffic flow pattern B will only start after traffic flow pattern A is completed, and traffic flow patterns C and D will occur one after the other. The order of occurrence of the four traffic flow patterns is A, B, C, and D.

The data set contains about 75,000 video frames. The traffic video was first divided into 250 segments of video clip that do not overlap in time by an equal length of 300 frames. A total of 73 video clips (including 21,900 frames) were extracted from the data set as training data for modeling the cascaded topic models. The remaining 177 video clips (including 53,100 frames) were used for testing. In the construction of the motion dictionary of the first-stage topic model, the 360×288 surveillance scene was divided into units of size 9×9, and the direction in which each unit may move was discretized into four directions perpendicular to each other. Thus, the size of the dictionary is 40×32×4.

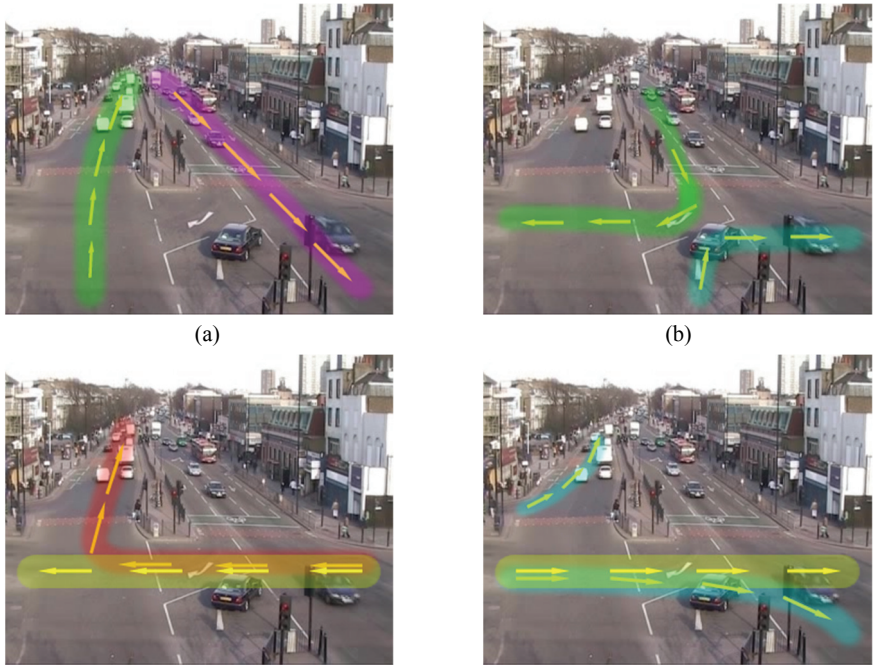


Fig. 4. Traffic flow patterns at a street intersection: (a) pattern A, (b) pattern B, (c) pattern C, and (d) pattern D.

5.2 Experimental Results

Test 1: Spatial context learning for local behavioral topics

In local behavior pattern learning (first-stage topic modeling), the trajectory segments in the video clip are regarded as documents. The vehicle motion trajectory points are mapped to the corresponding motion words in the motion dictionary in this stage according to the position of the rectangular region and the

discretized motion direction. At this time, the trajectory segments will be encoded as a sequence of motion words. Set the model parameters $\alpha=K/50$, $\beta=0.01$, and the topics number $K=30$. In the experiment, 30 local behavior topics are learned through the MRF-LDA topic model. The learned local behavior topics essentially represent the semantic regions in the scene that the trajectory segments pass by in the video clip.

In the process of local behavior topic clustering, the local behavioral topics learned in each video clip are clustered into 20 local behavior topic categories by spectral clustering algorithm. The distribution of visual local behavior topic categories is shown in Fig. 5. Each ellipse represents a local behavior topic category, with the ellipse center corresponding to the average position of all local behavior topics belonging to that category.

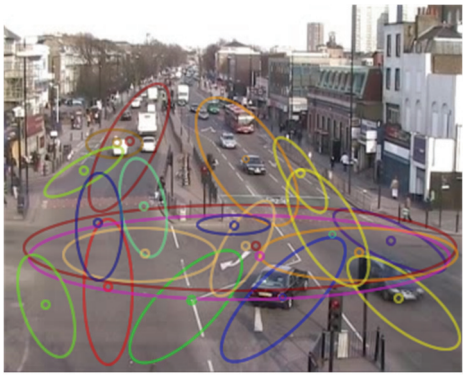


Fig. 5. Local behavior topic categories.

Test 2: Abnormal behavior recognition

Artwork has no text along the side of it in the main body of the text. In the second phase of LDA topic modeling, the video clip is treated as a document, and the local behavior topic categories correspond to the words in the dictionary in this stage. Set the model parameters $\alpha=K/50$, $\beta=0.01$, and the topics number $K=4$. Each learned topic corresponds to one type of commonly observed concurrent object behaviors under a specific traffic phase. In the process of identifying abnormal behavior, threshold value *TH* is set to 0.15, with the likelihood value $ClipSim(v_{test}|v_l)$ of the topic distribution in the video clip first calculated according to formula (9). Video clip abnormality scoring function value *abf* is then calculated according to formula (10). If *abf* is lower than threshold value *TH*, it can be judged that there is abnormal behavior in the video clip. Finally, according to the method for locating anomalous moving object trajectories in Section 4.4, the abnormal words in the video clip with abnormal behavior are captured, and the abnormal moving object trajectory can then be located. After 177 tested video clips were manually marked as normal or abnormal, 34 abnormal video clips were identified according to the video clip abnormality scoring function, and the moving object causing the abnormality was located. Fig. 6 shows part examples of identified abnormal behaviors.

It can be observed that the behavior of a single moving object shows very weak abnormal information, and it is normal to treat these behaviors in isolation. As the essence of such anomalous behavior, however, the behavior of the moving object occurs at the wrong place and time, resulting in an abnormal correlation with other moving objects in the scene.

(a)

(b)

Fig. 6. Abnormal behavior examples: (a) Example 1 and (b) Example 2.

It can be observed that the behavior of a single moving object shows very weak abnormal information, and it is normal to treat these behaviors in isolation. As the essence of such anomalous behavior, however, the behavior of the moving object occurs at the wrong place and time, resulting in an abnormal correlation with other moving objects in the scene.

Test 3: Comparison of different methods

We compared the performance of the cascaded topic model proposed in this paper with two types of methods. One type of method first performs segmentation of the scene, and then uses a two-level hierarchical model to model the global behavior patterns across regions, such as Cas-PLSA [22], CasDBNS [23], and Cas-LDA [27]. Another type of method uses only a single layer of LDA or PLSA models. Specifically, the ROC (receiver operating characteristic) curve and AUROC (area under the ROC curve) value are obtained by changing threshold *TH*. ROC space is defined by true positive rate (TPR) and false positive rate (FPR) as x and y axes, respectively, depicting relative trade-offs between true positive (benefits) and false positive (costs). The statistical results of TPR and FPR are required to draw the ROC curve.

We defined TPR to measure the proportion of correctly identified abnormal behaviors to all abnormal behaviors. True positive (TP) is the number of correctly identified abnormal behaviors, false negative (FN) is the number of abnormal behaviors misidentified as normal behaviors, and TP+FN is the total number of abnormal behaviors.

$$TPR = TP \big/ (TP + FN) \tag{11}$$

FPR was defined to measure the proportion of normal behaviors misidentified as abnormal behaviors to all normal behaviors. False positive (FP) is the number of normal behaviors misidentified as abnormal behaviors, True negative (TN) is the number of correctly identified normal behaviors, and FP+TN is the total number of normal behaviors.

$$FPR = FP / (FP + TN) \tag{12}$$

Thus, the ROC curves of different abnormal identification methods can be created by plotting the TPR against the FPR at various threshold settings as shown in Fig. 7.

From the ROC curves, the AUROC values of different abnormal identification methods can be obtained by computing the area under the ROC curves corresponding to their own as shown in Table 1.

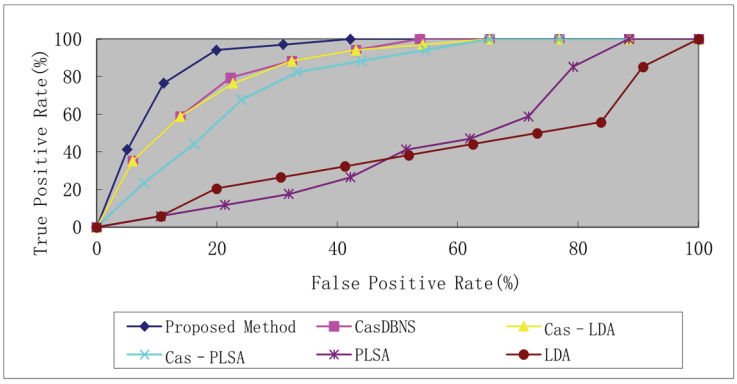


Fig. 7. Performance comparison of abnormal behavior recognition using ROC curves.

Table 1. AUROC values for different abnormal identification methods

Proposed method	CasDBNS	Cas-LDA	Cas-PLSA	PLSA	LDA
0.9153	0.8540	0.8472	0.7894	0.4353	0.3945

Generally speaking, a higher AUROC value indicates better performance. It can be seen from the experimental results that the two-level hierarchical model including the cascaded topic model proposed in this paper outperforms the single-layer LDA and PLSA model in terms of abnormal behavior recognition. The cascaded topic model proposed in this paper has the highest AUROC value of 0.9153, whose performance is superior to that of the other three two-level hierarchical models (Cas-PLSA, CasDBNS, and Cas-LDA). The performance of the three two-level hierarchical models is relatively close, the AUROC values are between 0.7 and 0.9, and there is certain recognition accuracy. Among them, Cas-PLSA has the lowest performance among the three. It can also be observed in the graph that the AUROC values of the single-layer LDA and PLSA models are all below 0.5, because the single-layer model cannot model the global correlation behavior and the application value is not high.

6. Conclusion

The abnormal behavior recognition method proposed in this paper decomposes complex global behaviors according to the spatio-temporal characteristics of behavior. It need not segment the surveillance scene in advance to obtain local behavior, and it has obvious advantages in simplifying complex global behavior modeling. More importantly, cascaded structure-based modeling and complex global behavioral decomposition naturally reflect complex behavioral spatio-temporal context structures,

which enable surveillance video to detect anomalous behavior more effectively. Finally, the abnormal behavior recognition method adopts a top-down strategy based on the consideration of global behavior correlation, which can not only recognize the video clips with abnormal behaviors but also locate the motion trajectories that cause abnormal behaviors within the video clip. The experimental results show that, in addition to the ability to identify different types of anomalous behaviors, the proposed method can also identify anomalies in complex behaviors that cannot be identified when considering individual object behaviors in isolation.

As future work, we would like to explore the parallel sampling algorithm of the topic model; on the other hand, we would like to study how to exploit the sparsity of the LDA model to accelerate the algorithm and save memory. Simultaneously, adjusting the parameters to optimize the model quality, optimizing hyper parameters α and β, and intelligently training the number of topics require further research. With the complexity of surveillance scenes and scopes, the coordination of multi-camera monitoring functions and the fusion of video data information will also be the focus of further research.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (No. 61672372, 61472211), Outstanding Science-Technology Innovation Team Program of Colleges and Universities in Jiangsu, Industrial Technology Innovation Project of Suzhou City (No. SYG201710), and Suzhou Vocational University Innovation Foundation (No. SVU2016CGCX06, SVU2018CX10).

References

[1] C. Piciarelli and G. L. Foresti, "On-line trajectory clustering for anomalous events detection," *Pattern Recognition Letters*, vol. 27, no. 15, pp. 1835-1842, 2006.

[2] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi, "Traffic monitoring and accident detection at intersections," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 2, pp. 108-118, 2000.

[3] Y. Sun, L. Sun, H. Zhu, and X. Zhou, "Activity anomaly detection based on vehicle trajectory of automatic number plate recognition system," *Journal of Computer Research and Development*, vol. 52, no. 8, pp. 1921-1929, 2015.

[4] C. Micheloni, L. Snidaro, and G. L. Foresti, "Exploiting temporal statistics for events analysis and understanding," *Image and Vision Computing*, vol. 27, no. 10, pp. 1459-1469, 2009.

[5] C. Piciarelli, C. Micheloni, G. L. Foresti, "Trajectory-based anomalous event detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1544-1554, 2008.

[6] H. Y. Hu, Q. N. Wang, Z. W. Qu, and Z. H. Li, "Spatial pattern recognition and abnormal traffic behavior detection of moving object," *Journal of Jilin University (Engineering and Technology Edition)*, vol. 41, no. 6, pp. 1598-1602, 2011.

[7] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Kauai, HI, 2001, pp. 123-130.

[8] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, 2004, pp. 819-826.

[9] T. Xiang and S. Gong, "Video behaviour profiling and abnormality detection without manual labelling," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV)*, Beijing, China, 2005, pp. 1238-1245.

[10] Y. Wang, H. Jiang, M. S. Drew, Z. N. Li, and G. Mori, "Unsupervised discovery of action classes," in *Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, 2006, pp. 1654-1661.

[11] T. Xiang and S. Gong, "Beyond tracking: modelling activity and understanding behavior," *International Journal of Computer Vision*, vol. 67, no. 1, pp. 21-51, 2006.

[12] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 893-908, 2008.

[13] M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 844-851, 2000.

[14] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception by hierarchical Bayesian models," in *Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, 2007, pp. 1-8.

[15] M. Y. Yang, W. Liao, Y. Cao, and B. Rosenhahn, "Video event recognition and anomaly detection by combining Gaussian process and hierarchical Dirichlet process models," *Photogrammetric Engineering & Remote Sensing*, vol. 84, no. 4, pp. 203-214, 2018.

[16] X. Wang, K. T. Ma, G. W. Ng, and W. E. Grimson, "Trajectory analysis and semantic region modeling using a nonparametric Bayesian model," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVRP)*, Anchorage, AK, 2008, pp. 1-8.

[17] X. Wang, K. T. Ma, G. W. Ng, and W. E. L. Grimson, "Trajectory analysis and semantic region modeling using nonparametric hierarchical Bayesian models," *International Journal of Computer Vision*, vol. 95, no. 3, pp. 287-312, 2011.

[18] B. Zhou, X. Wang, and X. Tang, "Random field topic model for semantic region analysis in crowded scenes from tracklets," in *Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI 2011, pp. 3441-3448.

[19] R. Kaviani, P. Ahmadi, and I. Gholampour, "Incorporating fully sparse topic models for abnormality detection in traffic videos," in *Proceedings of 2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, Iran, 2014, pp. 586-591.

[20] L. Song, F. Jiang, Z. Shi, and A. K. Katsaggelos, "Understanding dynamic scenes by hierarchical motion pattern mining," in *Proceedings of 2011 IEEE International Conference on Multimedia and Expo*, Barcelona, Spain, 2011, pp. 1-6.

[21] J. Li, S. Gong, and T. Xiang, "Scene segmentation for behaviour correlation," in *Computer Vision – ECCV 2008.* Heidelberg: Springer, 2008, pp. 383-395.

[22] J. Li, S. Gong, and T. Xiang, "Global behaviour inference using probabilistic latent semantic analysis," in *Proceedings of the British Machine Vision Conference, Leeds, UK*, 2008.

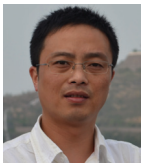
[23] C. C. Loy, T. Xiang, and S. Gong, "Detecting and discriminating behavioural anomalies," *Pattern Recognition*, vol. 44, no. 1, pp. 117-132, 2011.

[24] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006, pp. 113-120.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[26] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *International Journal of Computer Vision*, vol. 98, no. 3, pp. 303-323, 2012.

[27] J. Li, S. Gong, and T. Xiang, "Learning behavioural context," *International Journal of Computer Vision*, vol. 97, no. 3, pp. 276-304, 2012.



Yuanfeng Yang **https://orcid.org/0000-0003-4881-0523**

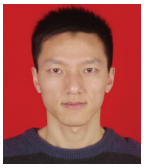
He received his Ph.D. degree in computer science and technology from Soochow University. He is currently an associate professor in the School of Computer Engineering, Suzhou Vocational University. His research interests include computer vision, pattern recognition and image processing.

Lin Li **https://orcid.org/0000-0001-7257-7659**

He received his Ph.D. degree at Yuan Ze University, Taiwan. He is currently an associate professor in the School of Computer and Information Engineering, Xiamen University of Technology. His research interests include data mining, decision analysis, cloud computing, and pattern recognition.

Zhaobin Liu **https://orcid.org/0000-0003-4632-4740**

He received his M.S. degree in computer science and technology from Xi'an Jiaotong University. He is currently a professor in the School of Computer Engineering, Suzhou Vocational University. His research interests include wireless sensor network and pervasive computing.



Gang Liu **https://orcid.org/0000-0001-7532-8586**

He received his Ph.D. degree in computer science and technology from Nanjing University of Science and Technology. He is currently a lecturer in the School of Computer Engineering, Suzhou Vocational University. His research interests include wireless sensor network, software engineering and information security.