Global Behaviour Inference using Probabilistic Latent Semantic Analysis

Jian Li, Shaogang Gong, Tao Xiang
Department of Computer Science
Queen Mary College, University of London, London, E1 4NS, UK
{jianli, sgg, txiang}@dcs.qmul.ac.uk

Abstract

We present a novel framework for inferring global behaviour patterns through modelling behaviour correlations in a wide-area scene and detecting any anomaly in behaviours occurring both locally and globally. Specifically, we propose a semantic scene segmentation model to decompose a wide-area scene into regions where behaviours share similar characteristic and are represented as classes of video events bearing similar features. To model behavioural correlations globally, we investigate both a probabilistic Latent Semantic Analysis (pLSA) model and a two-stage hierarchical pLSA model for global behaviour inference and anomaly detection. The proposed framework is validated by experiments using complex crowded outdoor scenes.

1 Introduction

For automatic dynamic scene analysis, anomaly detection is a challenging task especially given a scene consisting of complex correlated activities of multiple objects in an outdoor setting. Until now, most research has been focused on modelling and detecting anomalies of isolated or independent individual behaviours. For example, with tracking-based techniques [5, 6], each individual object's trajectory is compared to a set of known trajectory model templates and if the difference in trajectories is large, the corresponding behaviour is considered as being abnormal. However, examining individual object's behaviours in isolation is insufficient for describing potentially global anomaly involving multiple objects in a complex scene, where each object's behaviour is intrinsically affected by other objects either in the vicinity or further away. We consider that modelling and inferring global behavioural correlations shall provide a more meaningful mechanism for inferring global behaviour pattern and detecting anomaly in complex scenes.

Recently, a number of approaches have been proposed on modelling correlated behaviours of multiple objects. Xiang and Gong [10] proposed to cluster local events into categories by feature similarity. Activities are represented as sequential relationships among event groups using Dynamic Bayesian Networks. Their extended work was shown to have the capability of detecting suspicious behaviour in front of a secured entrance [11]. However, the types of activities modelled were restricted to a small set of events in a small local region without considering any true sense of global context. Brand and Kettnaker [1] attempted modelling scene activities using a Multi-Observation-Mixture+Counter Hidden Markov Model (MOMC-HMM). A traffic circle at a crossroad is modelled as sequential

states and each state is a mixture of multiple activities (observations). However, their abnormality detection is based only on how an individual behaves in isolation. How activities interact in a wider context is not considered. Wang et al [9] proposed modelling behaviour by grouping low-level motion features into topics using hierarchical Bayesian models. Since only simple local motion features are considered for behaviour representation, their method has limited ability to model behaviour correlations between moving and stationary objects, and ignored any global context for modelling complex behaviours in a wide-area scene.

In this work, we develop a framework for global behaviour inference and anomaly detection based on a novel model for multi-object behavioural correlation. In particular, object behaviours are represented as classes of spatio-temporal atomic video events. Each event class corresponds to behaviours of a group of objects with a certain size and specific motion directions. Without the need to track targets, such a representation is more robust for analysing crowded scenes. Behaviours are inherently context-aware, exhibited through constraints imposed by scene layout and the temporal nature of activities in a given scene. In order to constrain the number of meaningful behavioural correlations from potentially a very large number of all possible correlations of all the objects appearing everywhere in the scene, we first segment semantically a scene into different spatial regions by the spatial distribution of atomic video events in the entire scene. In each region, events are then re-clustered into different groups with ranking on both event types and their dominating features to represent how objects behave locally in each region. For modelling behaviour correlations within and across the segmented semantic regions, the probabilistic Latent Semantic Analysis (pLSA) model [3] is studied. The pLSA model was initially proposed for extracting semantic topics of linguistic words in text documents, [4]. More recently, the model and its derivatives have been employed in computer vision for extracting object categories [2] and recognising single object actions [7]. In this work, we first formulate a standard pLSA model for behaviour correlation modelling without considering any semantic context of a given scene. We then develop a novel two-stage hierarchical pLSA model based on semantic scene decomposition in order to improve the robustness of behaviour modelling against noise resulting in reduced false alarms in anomaly detection. Specifically, at the first stage, local behaviour correlations within each region are modelled. The inferred local behaviour patterns are then fed into the second stage for global behaviour inference and anomaly detection. The strength and weakness of both models are studied through extensive experiments carried out using complex crowded outdoor scenes. The results validate the effectiveness of the proposed framework.

2 Semantic Scene Segmentation

Behaviour Representation: We represent a behaviour using a set of low-level atomic video events of similar spatio-temporal features. To detect atomic video events, we first perform background subtraction and detect image events as blobs of foreground pixels, each of which is represented by a vector of 10 features as:

$$\mathbf{v}_f = [x, y, w, h, r_s, r_p, u, v, r_u, r_v], \tag{1}$$

where (x, y) and (w, h) are the centroid position and the width and height of a rectangular bounding box respectively, $r_s = w/h$ is the ratio between width and height, r_p is the percentage of foreground pixels in a bounding box, (u, v) is the mean optic flow vector

for the bounding box, $r_u = u/w$ and $r_v = v/h$ are the scaling features between motion information and blob shape. Instead of performing clustering directly to image events as proposed in [10], we derive a set of atomic events from these image events to reduce measurement noise. First, a video is temporally segmented into non-overlapping clips with equal length. Second, in each clip, image events are clustered using K-means and the number of clusters are set as the average number of image events across all the frames in that clip. We then regard each cluster of image events in a clip as an atomic event which is represented by a 20 components feature vector:

$$\mathbf{v} = [\bar{\mathbf{v}}_f, \bar{\mathbf{v}}_s], \tag{2}$$

where $\bar{\mathbf{v}}_f = mean(\mathbf{v}_f)$ and $\bar{\mathbf{v}}_s = var(\mathbf{v}_f)$, $\mathbf{v}_f$ given by Eqn. (1). Third, for all the atomic events that can be extracted from a video, a Gaussian Mixture Model (GMM) is employed for clustering with the number of clusters automatically determined using BIC [8]. Each cluster of atomic events is then defined as a type of behaviour. However, such a behaviour representation is based on a global clustering of all the atomic video events detected in the entire scene without any spatial or temporal restriction. It thus does not provide a good model for capturing behaviour correlations more selectively, both in terms of spatial locality and temporal dependency. In order to represent behaviours more accurately in context, we segment a scene semantically into regions according to event distribution.

Scene Segmentation: We treat this as an image segmentation problem. However, instead of representing each pixel location by RGB values or texture features, each pixel is assigned an event feature vector. The length of each vector is equal to the types of atomic video events detected in the entire scene, and each component corresponds to the number of occurrence of a specific type of event at that pixel location. For segmentation, a spectral clustering algorithm is deployed based on a modification of the method proposed by Zelnik-Manor and Perona [12]. The original Zelnik-Manor and Perona (ZP)'s algorithm automatically determines the scaling factors for measuring feature similarities and the number of segments. However, we find that the original ZP algorithm suffers from severe under-fitting given our data. To yield meaningful segmentation, instead of computing the feature scaling factor σ_i by measuring the distance between the current feature and the feature from a specific neighbour, we compute σ_i as the standard deviation of feature distances between the current location and all locations within a given radius r. The scaling factor σ_x is computed as the mean of the distances between all locations and the center of radius r. Given the feature similarity measurements, an affinity matrix can be constructed. The original ZP algorithm is then applied to the affinity matrix to automatically select the number of segments and perform segmentation.

Local Behaviour Learning: Given the segmented regions, local atomic events are relearned using image events within each region. Specifically, the most relevant features out of the 10 features in Eqn. (1) are selected using entropy in each region separately. The events represented using the selected features are then grouped within each region using the same clustering procedure described earlier, which results in different types of local behaviours being discovered by different local event clusters.

3 Global Interaction Modelling and Anomaly Detection

3.1 pLSA

The pLSA is a generative model which aims to find a latent topic $Z \in \mathscr{Z} = \{Z_1, \cdots, Z_{N_Z}\}$ from a vocabulary $\mathscr{W} = \{W_1, \cdots, W_{N_W}\}$ given a set of documents $\mathscr{D} = \{D_1, \cdots, D_{N_D}\}$ [3].

An explicit graphic representation is shown in Figure 1. Given observable variables $\mathscr{W}$

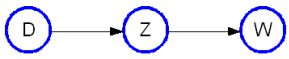


Figure 1: Standard pLSA.

and $\mathscr{D}$, an $N_D \times N_W$ dimensional co-occurrence matrix $\mathbf{M}$ can be built in which each entry $m(D_j, W_i)$ corresponds to the count of occurrence of word W_i in document D_j. In the pLSA, the joint probability between a word and a document can be expressed as:

$$P(D_j, W_i) = P(W_i|D_j)P(D_j), \tag{3}$$

where $P(W_i|D_j)$ is computed as:

$$P(W_i|D_j) = \sum_{k=1}^{N_Z} P(W_i|Z_k)P(Z_k|D_j). \tag{4}$$

The conditional probabilities of word, document given a latent topic $P(W_i|Z_k)$ and $P(D_j|Z_k)$ can be learned using an EM algorithm to maximise $\prod_i \prod_j P(D_j, W_i)^{m(D_j,W_i)}$ where the E-step is shown as:

$$P(Z_k|D_j, W_i) = \frac{P(Z_k)P(D_j|Z_k)P(W_i|Z_k)}{\sum_{k'=1}^{N_Z} P(Z_{k'})P(D_j|Z_{k'})P(W_i|Z_{k'})}, \tag{5}$$

and the M-step is shown as:

$$P(W_i|Z_k) \propto \sum_{j=1}^{N_D} m(D_j, W_i)P(Z_k|D_j, W_i), \tag{6}$$

$$P(D_j|Z_k) \propto \sum_{i=1}^{N_W} m(D_j, W_i)P(Z_k|D_j, W_i), \tag{7}$$

$$P(Z_k) \propto \sum_{j=1}^{N_D} \sum_{i=1}^{N_W} m(D_j, W_i)P(Z_k|D_j, W_i). \tag{8}$$

3.2 pLSA for Correlation Modelling

To modelling behavioural correlations using pLSA, we consider a video clip as a document in which a specific set of local behaviours/atomic event classes may occur. The classes of local behaviours learned from all regions are regarded as visual words. Any information on how different local behaviours are correlated is embedded in the document-word co-occurrence matrix $\mathbf{M}$, and is considered as interesting hidden topics to be discovered. Since we are only concerned with the occurrence of each type of local behaviour rather than the occurrence frequency, the elements of $\mathbf{M}$ is assigned to binary values:

$$m(D_j, W_i) = \begin{cases} 1 & \text{if } W_i \text{ occur} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

With this co-occurrence matrix, a pLSA model can be learned which is then used to infer the hidden topics. Given our definition of documents and words, the hidden global behaviour topics correspond to specific behaviour correlation structures and can be used to segment video clips into different temporal phases. In particular, during different phases, different correlations of local behaviours are expected. To infer the global behaviour topic given a learned pLSA behaviour correlation model and a video clip D_j, we compute

$$P(Z_k|D_j) = \frac{P(D_j|Z_k)P(Z_k)}{P(D_j)}. \tag{10}$$

where $P(D_j|Z_k)$ and $P(Z_k)$ are obtained using Eqn. (7) and Eqn. (8) respectively, and $P(D_j)$ is computed as:

$$P(D_j) = \sum_{k=1}^{N_Z} P(D_j|Z_k)P(Z_k). \tag{11}$$

The topic/phase for clip D_j is then determined as:

$$\text{Topic}(D_j) = \max_k P(Z_k|D_j). \tag{12}$$

Our behaviour pLSA model can also be readily used for abnormal behaviour detection via examining whether the behavioural correlations detected in a video clip are expected by the model. Specifically, we compute an abnormality score for each clip D_j as the joint probability of all local behaviour classes:

$$\log P(m(D_j,W_1),\cdots,m(D_j,W_{N_W})) = \sum_{i=1}^{N_W} m(D_j,W_i)\log P(W_i|D_j), \tag{13}$$

where $m(D_j,W_i) = 1$ indicates behaviour class W_i occurred in clip D_j whereas $m(D_j,W_i) = 0$ means W_i did not happen in D_j. A lower score indicates higher anomaly in this clip. Once an abnormal clip (document) is detected, the specific abnormal behaviour classes (words) that caused the abnormality (unusual topics) can be located by examining $P(W_i|D_j)$.

3.3 Hierarchical pLSA for Correlation Modelling

A novel two-stage hierarchical pLSA is formulated to overcome two shortcomings of the standard pLSA models for behaviour correlation modelling: 1) local behaviour detection are noisy in crowded scene due to image noise and occlusions. Using them directly as input makes pLSA vulnerable to noise; 2) global behaviour context embedded in semantic scene decomposition is ignored. The model structure is illustrated in Figure 2. The model consists of two stages. In the first stage, we treat each segmented region as a document and learn the local behaviour correlations. In the second stage, the local topics/phases obtained from each region are regarded as visual words for modelling global correlations. Compared to a standard pLSA, the proposed model uses the local behaviour topics inferred from the first stage pLSAs, instead of the detected noisy local behaviours, as model input for global behaviour inference. It is thus less sensitive to noise in behaviour representation. Furthermore, it seamlessly integrates the semantic scene decomposition result into model structure, which makes the model more suitable for complex behaviour modelling in a wide area.

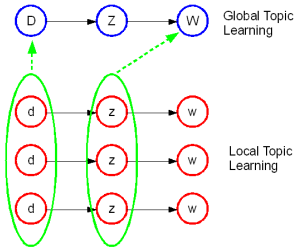


Figure 2: Hierarchical pLSA framework.

Suppose a scene is decomposed into Q regions, a video clip D_j is spatially split into Q sub-clips $D_j = \{d_j^1,\cdots,d_j^Q\}$. In the temporal domain, the corpus for a region q, where $1 \le q \le Q$, can be represented as $\mathscr{D}_q = \{d_1^q,\cdots,d_{N_D}^q\}$. Meanwhile, if N_w^q local behaviour

classes have been identified in region q, we consider the vocabulary of visual words in region q as $\mathscr{W}_q = \{w_1^q, \cdots, w_{N_w^q}^q\}$. Given the observable variables $\mathscr{D}_q$ and $\mathscr{W}_q$ in region q, we use a standard pLSA in the first stage to extract N_z^q local behaviour topics/phases: $\mathscr{Z}_q = \{z_1^q, \cdots, z_{N_z^q}^q\}$. In particular, we are interested in labelling a regional clip d_j^q with a dominant topic. This is achieved by firstly computing $P(z_k^q|d_j^q)$ for all possible values of k and then determining the local topic for d_j^q as:

$$\text{Topic}(d_j^q) = \max_k P(z_k^q|d_j^q). \tag{14}$$

In the second stage pLSA, we model behaviour correlations across regions. The local behaviour topics inferred from the first stage pLSAs are used as visual words in the second stage pLSA. More precisely, the global vocabulary of visual words can now be written as: $\mathscr{W} = \{z_1^1, \cdots z_{N_z^1}^1, \cdots, z_1^Q, \cdots z_{N_z^Q}^Q\}$ and the number of regional topics in the scene is denoted as N_W where $N_W = \sum_{q=1}^{Q} N_z^q$. Given a set of training video clips $\mathscr{D} = \{D_1, \cdots, D_{N_D}\}$, we can construct an $N_D \times N_W$ dimensional binary co-occurrence matrix $\mathbf{M}$ so that:

$$m(D_j, z_k^q) = \begin{cases} 1 & \text{if } \ z_k^q = arg\max_k P(z_k^q|d_j^q), \quad k = 1, \cdots N_z^q, \quad q = 1, \cdots, Q, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

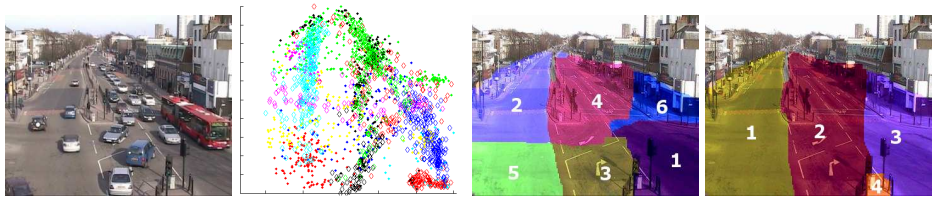
The global behaviour topics/phases $\mathscr{Z} = \{Z_1, \cdots, Z_{N_Z}\}$ are then inferred using the learned second stage pLSA. Correspondingly, the score for anomaly detection in each video clip is now computed as:

$$\log P(m(D_j, z_1^1), \cdots, m(D_j, z_{N_z^Q}^Q)) = \sum_{q=1}^{Q} \sum_{k=1}^{N_z^q} m(D_j, z_k^q) \log P(z_k^q|D_j). \tag{16}$$

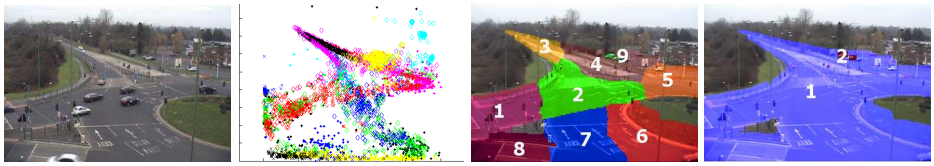
4 Experiments

Data Sets - We evaluated the performance of the proposed framework using video data captured from two busy traffic-light controlled road junctions (referred as Scene-1 and Scene-2 respectively). Example frames are shown in Figure 3 (a) and (e). Both videos were recorded at 25Hz and have a frame size of 360×288 pixels. In Scene-1, 2117 global atomic video events were extracted from 22000 frames (73 non-overlapping clips) used for training. The global atomic video events were automatically grouped into 13 clusters. In Scene-2, 43900 frames were used for training consisting of 146 non-overlapping clips. The extracted 4182 global atomic video events were grouped into 19 clusters. The clustering results are shown in Figure 3 (b) and (f) where clusters are distinguished by colour and labels. Our testing data consist of 12000 frames (39 clips) from Scene-1 and 44500 frames (148 clips) from Scene-2 respectively. There is no overlap between the training and the testing data.

Spatial Scene Segmentation - As can be seen in Figure 3 (c) and (g), Scene-1 and Scene-2 were segmented into 6 and 9 regions respectively using the modified ZP algorithm proposed in this paper. For comparison, the original ZP algorithm yields 4 regions and 2 regions respectively (see Figure 3 (d) and (h)). It is evident that the original ZP algorithm suffered from under-fitting severely and was not able to segment those scenes correctly according to spatial distribution of behaviours. In contrast, our approach provides a more meaningful semantic segmentation of both scenes.



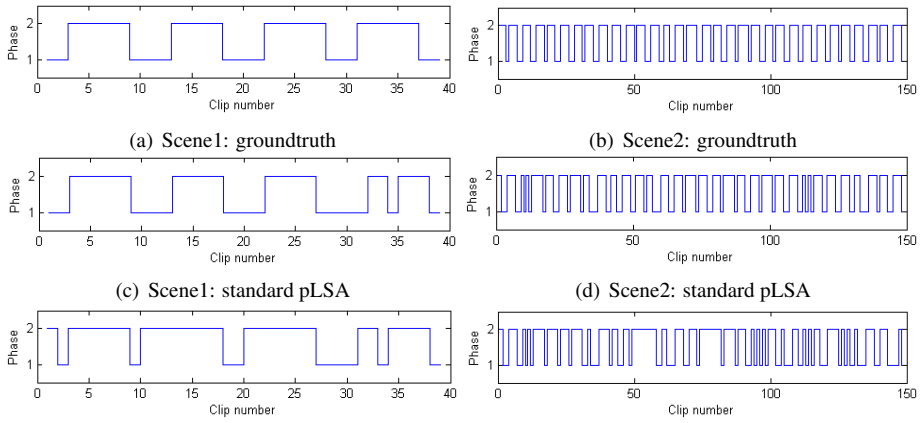
(a) Scene-1 (b) Event classification (c) Modified ZP (d) Original ZP



(e) Scene-2 (f) Event classification (g) Modified ZP (h) Original ZP

Figure 3: Semantic scene segmentation.

Global Behaviour Topic Inference - Given the segmented local regions, 30 classes of local behaviours were learned in Scene-1 and 52 classes were learned in Scene-2. The standard pLSA model and the hierarchical pLSA framework were used for modelling behaviour correlations and inferring global behaviour topics/phases. As behaviours occurred in both scenes are controlled largely by multiple traffic lights (up to 6), it is appropriate to set the number of global behaviour topics to 2 for both models, reflecting the number of traffic phases in each scene. For the hierarchical pLSA model, the number of local behaviour topics in each region was set to 6. This is because apart from the traffic lights, local behaviours are also controlled by additional local factors such as the distance of the vehicle in front; more hidden topics are thus needed.



(a) Scene1: groundtruth (b) Scene2: groundtruth

(c) Scene1: standard pLSA (d) Scene2: standard pLSA

(e) Scene1: hierarchical pLSA (f) Scene2: hierarchical pLSA

Figure 4: Temporal phase identification.

The global behaviour phases inferred using both models are shown in Figure 4. Ground truth was obtained by manually labelling each video clip in the testing data set into one of the two phases according to the traffic light phases. The accuracy of the global behaviour inference by both models was measured against the ground truth and shown in Table 1. Figure 4 and Table 1 indicate that both models achieve accurate global behaviour inference, with pLSA outperforming the hierarchical pLSA. It should be noted that the testing

video for Scene-1 contains clips with abnormal behaviour correlations and they may also affect the performance of temporal phase identification.

Accuracy	Scene-1	Scene-2
Standard pLSA	89.74 %	84.46 %
Hierarchical pLSA	76.92 %	72.30 %

Table 1: Global behaviour inference accuracy.

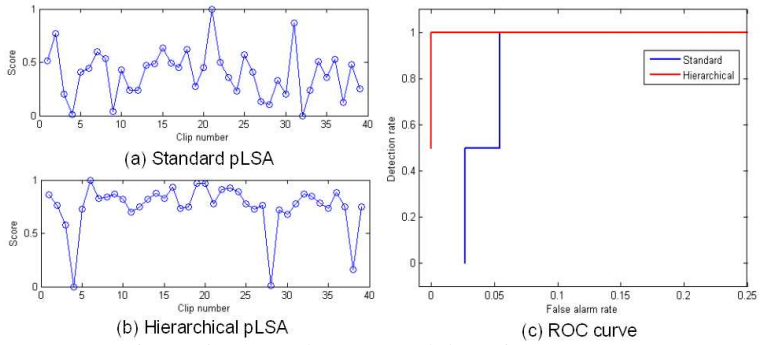


Figure 5: Anomaly score and detection accuracy.

Anomaly Detection - We examined the performance of anomaly detection of the proposed methods using a test video from Scene-1 consisting of 12000 frames or 39 clips. In the test video, two abnormal behaviours can be found in clip 4 and clip 28 respectively. Both were caused by the sudden occurrence of fire engines which interrupted the normal traffic flow (see Figure 6 and 7). The abnormality scores computed using Eqn. (13) and (16) were used for anomaly detection for the pLSA and hierarchical pLSA respectively. The lower the score is, the more likely it is that the clip contains abnormal behaviours. It can be seen from Figure 5 (a) and (b) that both models gave lowest scores for clip 4 and clip 28, indicating that both of them can be detected correctly. However, it can be seen from Figure 5 (a) that quite a few normal clips were also given low scores by the standard pLSA model, which would cause false alarms. For instance, clip 9 has an almost identical score to clips 4 and 28. To have a more detailed comparison of the two models, ROC curves are plotted which take into consideration both detection rate and false alarm rate. Figure 5 (c) shows clearly that the hierarchical pLSA yields better anomaly detection performance.

To locate the local behaviours that caused an anomaly using the standard pLSA, we computed $P(W_i|D_j)$, i.e. the probability of the occurrence of a behaviour in that specific clip, for each type of local behaviours occurred in a clip and then identified the five local behaviours with the lowest values of $P(W_i|D_j)$ as the cause for the anomaly. It is less straightforward for the hierarchical pLSA model. Specifically, we firstly computed $P(W_i|D_j)$ for the second stage pLSA to identify the three local regions that contribute to the lowest $P(W_i|D_j)$ values. We then considered each region to identify the local behaviours that should be blamed for the anomaly, using the first stage pLSAs for each region. Figure 6 and 7 show the local behaviours that caused clip 4 and 28 to be detected as anomalies using both models. It can be seen that both models correctly located mostly where and when an anomaly was taking place, with standard pLSA model giving more false alarms. By locating the cause of anomaly we can shed more light into the cause for

(a) Standard pLSA

(b) Hierarchical pLSA

Figure 6: Abnormal behaviour detection in clip 4. Different classes of local behaviours in each clip that caused the anomaly are shown using bounding boxes of different colours.

(a) Standard pLSA

(b) Hierarchical pLSA

Figure 7: Abnormal behaviour detection in clip 28.

Figure 8: The local behaviours that contribute to the false detection of clip 9 as an anomaly using the standard pLSA.

false alarms as well. It is evident in Figure 5 (a) that clip 9 is very likely to be falsely detected as an anomaly using the standard pLSA model. Figure 8 suggests that the erroneous atomic event detection caused by object occlusions is the reason for the false alarm. In contrast, clip 9 gives much higher score, indicating that the hierarchical pLSA model is more robust to noise and errors in behaviour representation.

5 Discussions and Conclusions

Our experimental results demonstrate that both the pLSA and the hierarchical pLSA models can effectively model behaviour correlations for global behaviour inference and anomaly detection. The results also suggest that the pLSA model is superior to the hierarchical pLSA models for global behaviour inference, whereas the latter has better performance on anomaly detection. Note that one of the main challenges for anomaly detection is to distinguish an anomaly from a noise contaminated normal behaviour; it is thus not surprising that the hierarchical pLSA model is better at anomaly detection due to its robustness to noise. This robustness is achieved by using behaviour topics/phases inferred at each semantically decomposed region as input for global behaviour inference. However, since not all the regions have a clear phase structure (e.g. some regions in Scene-1 contain pedestrians walking on pavements whose behaviours are not controlled by traffic lights), enforcing pLSAs at these regions will introduce uncertainties in the second stage pLSA for global behaviour topics/phases inference. This explains why the standard pLSA model gives better global behaviour topic estimation. The ongoing work is focused on how to automatically remove these regions from the hierarchical pLSA model.

References

[1] M. Brand and V. Kettnaker. Discovery and segmentation of activities in video. *PAMI*, 22 (8):844–851, 2000.

[2] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image serch. In *ICCV*, pages 1816–1823, Beijing, October 2005.

[3] T. Hofmann. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence*, 1999.

[4] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999.

[5] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *PAMI*, 28 (9):1450–1464, 2006.

[6] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. In *BMVC*, volume 2, pages 583–592, 1995.

[7] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, pages 1249–1258, 2006.

[8] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464., 1978.

[9] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception by hierarchical bayesian models. In *CVPR*, pages 1–8, Minneapolis, June 2007.

[10] T. Xiang and S. Gong. Beyond tracking: Modelling activity and understanding behaviour. *IJCV*, 67 (1):21–51, 2006.

[11] T. Xiang and S. Gong. Video behavior profiling for anomaly detection. *PAMI*, 30 (5):893–908, 2008.

[12] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, 2004.