From Local Temporal Correlation to Global Anomaly Detection

Chen Change Loy, Tao Xiang, Shaogang Gong

From Local Temporal Correlation to Global Anomaly Detection

Chen Change Loy, Tao Xiang, and Shaogang Gong

Department of Computer Science
Queen Mary, University of London, London E1 4NS, UK
{ccloy,txiang,sgg}@dcs.qmul.ac.uk

Abstract. In this paper, we propose a novel framework tailored towards global video behaviour anomaly detection in complex outdoor scenes involving multiple temporal processes caused by correlated behaviours of multiple objects. Specifically, given a complex wide-area scene that has been segmented automatically into semantic regions where behaviour patterns are represented as discrete local atomic events, we formulate a novel cascade of Hidden Markov Models to model behaviours with complex temporal correlations by utilising combinatory evidences collected from local atomic events. Using a cascade configuration not only allows for accurate detection of video behaviour anomalies, more importantly, it also improves the robustness of the model in dealing with the inevitable presence of errors and noise in the behaviour representation resulting less false alarms. We evaluate the effectiveness of the proposed framework on a real world traffic scene. The results demonstrate that the framework is able to detect not only anomalies that are visually obvious, but also those that are ambiguous or supported only by very weak visual evidence, e.g. those that can be easily missed by a human observer.

1 Introduction

Automated anomaly detection has drawn increasing attention in computer vision research [1–3] due to its potential applications in security and visual surveillance, video highlight extraction and video summarisation. These applications have traditionally relied on human operators to monitor the videos continuously and determine whether behaviours being monitored are normal or abnormal. Anomaly detection systems are anticipated to ease the burden of human operators by automatically sending alerts on abnormal behaviours that should be given further examination.

Abnormal behaviours may involve either single object or multiple objects. Anomalies in the former case are mostly caused by a large spatial deviation and/or an obvious change of visual appearance or motion characteristic triggered by objects in isolation, e.g. a person running whilst others walking, or a car doing U-turn on a one way street. They are visually more distinctive and can be easily detected either by human operators or most of the existing techniques such as template/trajectory matching [4, 5]. Detection of abnormal behaviours involving multiple objects is more challenging, especially when the anomalies are caused by abnormal correlations among behaviours of multiple objects, e.g. a vehicle changes lane at a wrong time forcing other vehicles to slow down. These anomalies are more subtle to detect because they often take place in a larger spatial and temporal context. Conventional methods that consider only behaviour

of individual object in isolation are inadequate in detecting this type of anomalies [6, 7] because each individual object's behaviour can be perfectly normal without being considered in a global context, e.g. in the aforementioned example, neither a vehicle changing lane or a vehicle slowing down is abnormal when viewed in isolation.

Moreover, the characteristics of different types of abnormal behaviour involving multiple objects are visually more difficult to define and quantify. Until now, there is little if any work done to address this problem systematically. To that end, we consider that multi-object anomalies can be grouped into three categories based on their visual distinctiveness and their frequency of occurrence in the training set. The first category (*Category-A*) of abnormal correlations often causes an apparent visual changes which are very different from what are observed from the training set (e.g. an accident in a traffic junction causing interruptions to the overall traffic flow). *Category-B* corresponds to anomalies that are visually ambiguous due to their rare occurrence in the training set. The third category (*Category-C*) of anomalies are supported only by very weak visual evidence, i.e., featured with very subtle deviation from the normal temporal order/durations of different correlated temporal processes (e.g. a train arrives slightly ahead of schedule on a platform). Anomalies in both Category-A and Category-C refer to those that have never occurred in the training set, whilst anomalies in Category-B are those that appear in the training set but are statistically under-represented. From the perspective of a human observer, anomalies in Category-A are visually obvious thus easy to detect. In contrast, both anomalies in Category-B and Category-C are more likely to be missed due to the lack of visual evidence.

A few approaches have been proposed for modelling activities involving multiple objects [6–9], little effort has been taken in detecting abnormal interactions that are either ambiguous or weakly supported by visual evidence. Previous work mainly focused on classifying interactive behaviours [7–9] or detecting anomalies with apparent visual evidence (Category-A) [6]. In contrast, we focus on the detection of abnormal interactions involving multiple objects (we refer to as global anomalies), particularly those belonging to Category-B and Category-C, for which modelling the correlations of multiple temporal processes in an effective and efficient manner becomes crucial. In particular, considering that in a busy public space different objects' behaviours are correlated either implicitly or explicitly with each other, we propose to exploit the *combinatory evidence* obtained from each individual temporal process to support the decision in global anomaly detection.

2 Related Work

A number of approaches have been proposed to model behaviours of multiple objects without considering the temporal correlations among multiple objects [6, 7]. The assumption of different temporal processes being independent is clearly invalid when dealing with behaviours involving multiple interacting objects. One of the early work in modelling correlations among objects was undertaken by Oliver et al. [8, 10], in which a Coupled Hidden Markov Model (CHMM) [11] is used to model people interactions based on their proximity and trajectories in an outdoor scene. Causal relationships among multiple temporal streams are taken into account by fully coupling pairs of Hidden Markov Models (HMMs) so that each hidden state is conditionally dependent on all

past states of all processes in the previous time instance. An alternative solution based on Dynamically Multi-Linked HMM (DML-HMM) is introduced by Gong and Xiang [9]. In contrast to CHMM, causal temporal relationships are established only among a subset of hidden states with high relevance across multiple temporal processes. Both CHMM and DML-HMM have been shown to be effective in modelling correlated behaviours. However, the behaviours being modelled in [8,9] are limited to small local regions. Temporal correlations in a larger spatial context are not investigated. In addition, there are no attempt to detect ambiguous anomalies and subtle anomalies supported with weak evidence. In contrast, our cascade HMMs are designed to learn efficiently from combinatory evidence extracted from local temporal processes, thus providing a mechanism to detect abnormal correlations in a complex scene with a large spatial context. Moreover, thanks to the cascade model structure, our model is less susceptible to noise and errors in behaviour representation compared with an alternative single stage model.

More recently, a static model based on a hierarchical probabilistic latent semantic analysis (pLSA) is introduced by Jian et al. [12] for global phase inference and anomaly detection in traffic scenes. A scene segmentation algorithm [13] is used to decompose a scene into semantic regions, where each of them encloses an area that contains a set of correlated atomic events. The bottom layer of the hierarchical pLSA is employed to model the local atomic events, whilst the top layer is used to model the global correlations among them. Despite the proposed method has been shown to be capable in detecting abnormal behaviours in a global context, it is limited to modelling static causal relationships without taking the temporal ordering and duration aspects of behaviours into account. The model is thus unable to detect anomalies embedded in the temporal structure of correlation.

In this work, we propose a novel framework based on a cascade of HMMs to model the multiple temporal processes in busy outdoor scenes. The use of cascade structure offers a number of unique advantages over the existing Dynamic Bayesian Network (DBN) based models: (1) it provides a mechanism that exploits evidences from local temporal processes to detect video behaviour anomalies supported only by weak visual evidence; (2) it provides effective temporal correlation modelling, enabling it to detect anomalies embedded in the temporal structure of correlation; (3) the cascade structure is less susceptible to noise compared with a single model that learns directly from the observational space. Our approach overcomes the problems associated with a static model such as Jian et al. [12], as this is a dynamic temporal model. Experiments are carried out on a busy traffic scene. The results show that the framework is able to detect not only anomalies that are visually obvious, but also those that are ambiguous or supported only by very weak visual evidence, i.e., those that can be easily missed even by a human observer.

3 Methodology

3.1 Behaviour Representation

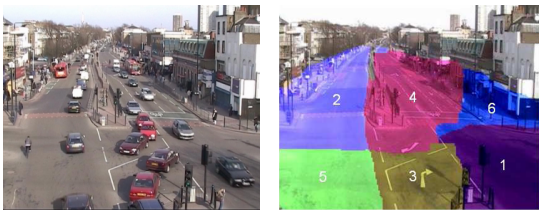
Since reliable tracking is difficult and discontinuity in object trajectories is prevalent in a crowded outdoor scene, trajectory-based representation is not employed in this study.

Instead, we adopted a discrete event based approach introduced by Jian et al. [13] to represent global behaviours in a wide-area outdoor scene. We provide a concise description of the approach here to facilitate explanations in other sections appeared later. Interested readers may refer to the original paper [13] for more details.

A continuous video sequence $\mathbf{V}$ is first segmented uniformly into T non-overlapping video clips $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_t, \ldots, \mathbf{v}_T\}$, with each video clip $\mathbf{v}_t$ containing N_f frames. Foreground blobs with similar spatio-temporal features are then clustered using k-means into a set of atomic events (e.g. vehicles stop at the middle of intersection waiting for right-turning) across all the frames in a clip $\mathbf{v}_t$. The spatio-temporal features employed in the clustering include the centroid position and the width and height of the object bounding box, the ratio of width and height, the occupancy ratio, the optical flow vectors for the bounding box and the scaled optical flow vectors according to the bounding box size. Upon obtaining atomic events for all clips, global clustering based on Gaussian Mixture Model (GMM) is performed to group the atomic events into coarse global event classes. Based on the distributions of the clusters, spatial scene segmentation is carried out to decompose a scene $\mathbf{S}$ into R semantic regions, where $S = \{s_1, \ldots, s_r, \ldots, s_R\}$. Specifically, two pixels are considered to be similar and grouped together if similar classes of events occurred there. Consequently, object behaviours within each segmented region are similar to each other whilst being different from those in other regions. To detect the atomic events more accurately, the aforementioned clustering method is repeated within each region with automatic feature selection to group foreground blobs into finer regional event classes. As a result, a video clip $\mathbf{v}_t$ is spatially represented by segmented regions (see Fig. 1(b)), where each region contains a set of correlated regional atomic events. To construct the input features for the proposed cascade model, we represent the behaviours captured in a video clip $\mathbf{v}_t$ using R binary vectors $\{\mathbf{y}_t^1, \ldots, \mathbf{y}_t^r, \ldots, \mathbf{y}_t^R\}$, which correspond to the occurrence of regional events in each region. Specifically, the binary vector $\mathbf{y}_t^r$ is given as $\mathbf{y}_t^r = [O_1, \ldots, O_m, \ldots, O_{M_r}]$, where

$$O_m = \begin{cases} 1 & \text{if atomic event } e_m^r \text{ occurs, } 1 \leq m \leq M_r \\ 0 & \text{otherwise} \end{cases} . \tag{1}$$

where e_m^r is the atomic event belonging to the mth regional event class in region s_r and M_r is the total number of regional event classes in region s_r.



(a) A traffic scene (b) Six segmented regions

Fig. 1. Semantic scene segmentation.

3.2 Cascade Model Structure

The proposed framework combines two stages of Multi-Observation HMMs (MOHMMs) [9] in a cascade, and the model is represented as a DBN [14]. The first stage of the model is composed of multiple MOHMMs, each of which is used to model the temporal evolvement of regional atomic events within each region. The second stage consists of a single MOHMM for modelling the state sequence inferred from the first stage, and is responsible for learning the global correlations among temporal processes across regions. In contrast to conventional HMM that emits a deterministic symbol in a given state, a MOHMM allows a state to produce multiple symbols in every time step. It is thus ideal for modelling multiple atomic events temporally in each region in the first-stage cascade, and modelling the temporal correlations of behaviours across multiple regions in the second-stage cascade.

The MOHMMs in both stages are ergodic models and characterised by a tuple $\lambda = (Q, O, \pi, A, B)$ where Q is the state variable, O is the set of observation symbols, π is the initial state distribution, A the state transition distribution ($a_{ij} = P(Q_{t+1} = q_j|Q_t = q_i)$), and B is observational distribution ($b_j(m) = P(O_t = o_m|Q_t = q_j)$). In the following text, the MOHMMs at the first stage are known as local models and denoted as $\mathbf{\Lambda}^l$, where $\mathbf{\Lambda}^l = \{\lambda_1, \ldots, \lambda_r, \ldots, \lambda_R\}$, and R is the number of regions. The second stage MOHMM is known as global model denoted as $\mathbf{\Lambda}^g$. We denote the state space of hidden variable Q^{l^r} of a local model as $q_i^{l^r}$ where $1 \leq r \leq R$, and the state space of hidden variable Q^g of a global model is denoted as q_i^g, where i is the state index. Figure 2 illustrates the structure of the proposed framework. Observation nodes are shown as shaded squares and hidden nodes are shown as clear squares.

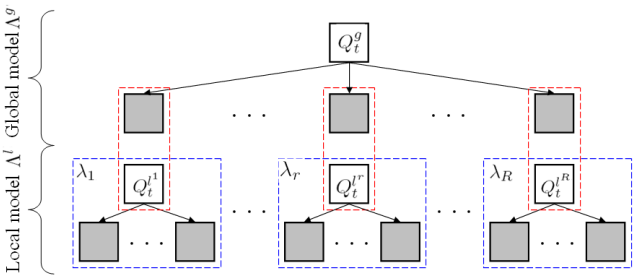


Fig. 2. Dynamic Bayesian Network (DBN) representation of the proposed cascade model with one time slice unrolled.

3.3 Model Training

Model training is accomplished through two steps, i.e., the local models training and global model training. We use training sequence $\mathbf{y}_{1:T}^r = \{\mathbf{y}_1^r, \ldots, \mathbf{y}_t^r, \ldots, \mathbf{y}_T^r\}$ to estimate parameters of the model λ_r through Expectation Maximisation (EM) algorithm. However, it is well known that EM algorithm may converge to a poor local optimum solution if the parameters are not initialised properly. One possible solution for this is to perform k-means clustering on the training data for model parameter initialisation, with

the number of clusters k equals to the number of hidden states N_s. However, k-means clustering itself converges to numerous local minima depending on the initial cluster centroids. To overcome this problem, we repeat the clustering procedure for 100 times with the initial cluster centroids being selected randomly. We then select the cluster formation that minimises the sum of squared error (SSE) distances between the data and the centroid of their clusters. Specifically, the SSE is defined as:

$$SSE = \sum_{i=1}^{k} \sum_{\mathbf{y} \in \mathbf{c}_i} (\mathbf{y}_i - \overline{\mathbf{c}_i})^2. \tag{2}$$

where $\overline{\mathbf{c}_i}$ is the centroid of cluster $\mathbf{c}_i$, where $1 \leq i \leq k$.

Subsequently, we initialise the model λ_r based on the result obtained from k-means clustering. Specifically, the observation probability distribution $B = \{b_i(m)\}$ is estimated by calculating the frequency of occurrence of the symbol in the corresponding cluster, which can be written as:

$$\begin{aligned} b_i(m) &= P(O_m \text{ at } t | Q_t^{l^r} = q_i^{l^r}) \\ &= \frac{\#\, O_m \text{ in } \mathbf{c}_i}{T}. \end{aligned} \tag{3}$$

where $1 \leq i \leq N_s$ and $1 \leq m \leq M_r$. We use uniform estimates for the initial state distribution $\pi = 1/N_s$ and state transition distribution A. Upon determining the initial values of the parameters, learning proceeds to estimate the parameters through EM algorithm.

After the training of local models $\mathbf{\Lambda}^l$, we proceed to the training of the global model $\mathbf{\Lambda}^g$. First, the most probable explanation (MPE) through a local model λ_r is obtained by using Viterbi decoding, given as:

$$\mathcal{Q}_{1:T}^{l^r} = \underset{Q_{1:T}^{l^r}}{\operatorname{argmax}}\, P(Q_{1:T}^{l^r} | \mathbf{y}_{1:T}^{r}). \tag{4}$$

The MPE obtained across all first-stage models represents the understanding of the learned models regarding the phases of individual temporal processes. With the help of Viterbi algorithm, the temporal and causal relationships learned during the model training can be employed to explain away the errors in the observational space. This would reduce the influence of noise during second-stage model learning. The MPE obtained across all local models are then augmented, forming an intermediate observational vector, given as:

$$\mathbf{z}_t = [\mathcal{Q}_t^{l^1}, \ldots, \mathcal{Q}_t^{l^r}, \ldots, \mathcal{Q}_t^{l^R}]. \tag{5}$$

The input is then transmitted to the global model $\mathbf{\Lambda}^g$ for training. Parameters initialisation and estimation of the global model are carried out with the similar steps applied to the local models.

3.4 Anomaly Detection

On-line abnormality detection can be achieved by examining the log-likelihood of a given sequence with respect to the model over time. A low log-likelihood value may

signify an occurrence of anomaly. Given an unseen sequence $\mathbf{y}^r_{1:t}$, where $1 \leq t \leq T$, the normalised log-likelihood $LL^{l^r}_t$ at clip t with respect to the local model λ_r and the normalised log-likelihood LL^g_t at clip t with respect to the global model $\mathbf{\Lambda}^g$ can be computed as follows:

$$LL^{l^r}_t = \frac{1}{t} \log P(\mathbf{y}^r_{1:t}|\lambda_r). \tag{6}$$

$$LL^g_t = \frac{1}{t} \log P(\mathbf{z}_{1:t}|\mathbf{\Lambda}^g). \tag{7}$$

Both log-likelihood values can be used for abnormality detection. However, since we are interested in global behaviour anomalies that are defined in the correlations of multiple temporal processes, the use of global log-likelihood LL^g_t for anomaly detection would be more appropriate.

To obtain LL^g_t in an on-line manner so that anomalies can be detected on the fly, we need to estimate the local hidden state $Q^{l^r}_t$ instantly at every clip t. By on-line estimation we mean the past evidence before the current clip t need not be presented to the model for the estimation of most likely state. This results in a constant computational time in computing LL^g_t over time. To estimate $Q^{l^r}_t$, the marginal probabilities $P(Q^{l^r}_t = q^{l^r}_i|\mathbf{y}^r_{1:t})$ are first computed, where $q^{l^r}_i$ represents the ith hidden state and $\mathbf{y}^r_{1:t}$ is the unseen sequence between times 1 and t. The most likely hidden state is then determined by choosing the hidden state that yields the highest marginal probability:

$$\mathcal{Q}^{l^r}_t = \underset{q^{l^r}_i}{\operatorname{argmax}}\ P(Q^{l^r}_t = q^{l^r}_i|\mathbf{y}^r_{1:t}). \tag{8}$$

The computation of marginal probability takes constant space and time at every time instance by using recursive estimation to compute $P(Q^{l^r}_t|\mathbf{y}^r_{1:t})$ as a function of current input $\mathbf{y}^r_t$ and prior belief state $P(Q^{l^r}_{t-1}|\mathbf{y}^r_{1:t-1})$. Specifically,

$$\begin{aligned} P(Q^{l^r}_t|\mathbf{y}^r_{1:t}) &\propto P(\mathbf{y}^r_t|Q^{l^r}_t, \mathbf{y}^r_{1:t-1}) P(Q^{l^r}_t|\mathbf{y}^r_{1:t-1}) \\ &\propto P(\mathbf{y}^r_t|Q^{l^r}_t)[\textstyle\sum_{Q^{l^r}_{t-1}} P(Q^{l^r}_t|Q^{l^r}_{t-1}) P(Q^{l^r}_{t-1}|\mathbf{y}^r_{1:t-1})]. \end{aligned} \tag{9}$$

where the constant of proportionality is $1/P(\mathbf{y}^r_t|\mathbf{y}^r_{1:t-1})$. The procedure described here is traditionally known as "filtering", because we are filtering out the noise from the observations [14].

The most likely hidden state sequences obtained from each local model at clip t are augmented according to (5) to form an observational input for global model. Subsequently, we can compute the LL^g_t following (7). Global anomaly detection can be achieved by comparing LL^g_t with a pre-defined threshold TH. Specifically, if $LL^g_t < TH$, the unseen clip is detected as abnormal.

4 Experiments

4.1 Dataset

Experiments were conducted on a busy road traffic junction scene. The video footage was recorded at 25Hz and resized to a resolution of 360 $\times$ 288 pixels. The total duration of the recording is approximately 22 minutes, showing a busy road junction regulated by

traffic lights, dominated by four types of traffic flows as illustrated in Fig. 3. Specifically, Flow A corresponds to traffic in vertical directions. Flow B, C and D are regarded as traffic flows in horizontal directions. In particular, Flow B represents left-turning and right-turning traffics by vehicles from vertical directions. Flow C corresponds to rightward traffic and flow D corresponds to leftward traffic. The order of the traffic flow depends upon how busy the vertical traffics are. During most of the recording, the scene is extremely crowded. Consequently, Flow B can only take place after Flow A finishes and is followed by C and D, (i.e. the typical temporal order is A, B, C, D). However, it is noted that very occasionally, there is gap in Flow A which is big enough for Flow B to take place until the gap closes. In other words, Flow A and B can occur alternatively during the vertical traffic phase. This makes global behaviour modelling and anomaly detection challenging in this scene as vehicles do not interact simply following the traffic light cycles, but also driven by the traffic volume as well as the driving habits and reactions of drivers.

(a) Flow A (b) Flow B (c) Flow C (d) Flow D

Fig. 3. Traffic flows observed in the dataset.

A total of 112 non-overlapping clips were segmented from the video. In particular, 73 clips (22000 frames) were used for training, whilst 39 clips (12000 frames) were reserved for testing. Scene segmentation was applied on the dataset and a total of 30 local atomic events were discovered in six regions shown in Fig. 1(b).

Prior to the evaluation of the proposed method, ground truth was first obtained by manually performing exhaustive frame-wise examination on both the training and testing set. Consequently, 5 out of 39 testing clips were found to be abnormal, they are clips 3, 4, 10, 25 and 35. Among them, clips 3 and 4 were categorised in Category-A since they contain abnormal behaviours that were visually recognisable. In particular, the abnormality was caused by a fire engine approaching from the left entrance towards the right exit, causing interruptions to the vertical traffic at both directions (see Fig. 4(a) and 4(b)). The vehicles from vertical directions gave way to the fire engine which has priority, and the normal traffic flow resumed after the fire engine exited.

Clips 25 and 35 correspond to abnormal traffic flow where vehicles did not follow the typical temporal order of A-B-C-D. In particular, clip 25 shows a motorbike making a turn to the left exit from the upper direction during vertical traffic flow A. In clip 35, vehicles were using a gap in the middle of traffic flow A to make right turn and left turn at the same time interval. Both clips were grouped into Category-B because they belong to rare/unusual behaviour with low frequency of occurrence in the training set (out of 73 training clips, only 3 clips correspond to left-turn and 2 clips correspond to turning both ways).

Clip 10 was labelled as abnormal behaviours belonging to Category-C, where the abnormality is almost undetectable by human without comprehensive examination of the traffic cycle duration over time. Clip 10 (see Fig. 4(e)) shows a white van on its way from the left entrance to right exit. After careful examination on the traffic cycle duration and the distance between the van and the vehicles behind, it is discovered that the van ran the red light before traffic flow B finished its cycle. The ground truth is summarised in Table 1.

Category	Description	Clip
A	Abnormal behaviours that are visually obvious	3, 4
B	Rare and ambiguous behaviours	25, 35
C	Abnormal behaviours supported by weak evidence	10

Table 1. Ground truth was defined in accordance to the anomaly's visual distinctiveness and its frequency of occurrence in the training set.

(a) Clip 3 (Category-A) (b) Clip 4 (Category-A)

(c) Clip 25 (Category-B) (d) Clip 35 (Category-B) (e) Clip 10 (Category-C)

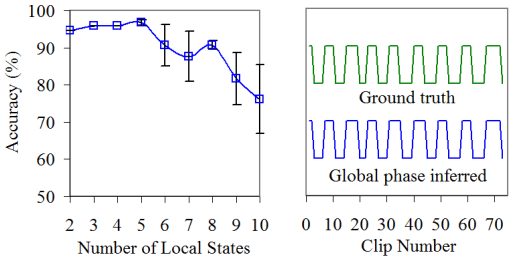
Fig. 4. Clips 3 and 4 are abnormal behaviours that are visually detectable (Category-A), whilst clips 25 and 35 are abnormal behaviours that are ambiguous (Category-B). Clip 10 contains subtle anomaly that is almost undetectable by human (Category-C). The corresponding objects that caused the anomalies are manually highlighted.

4.2 Anomaly Detection

We employed a single MOHMM as a baseline method in this study. The number of hidden states were set to 2 with each of them represent the horizontal and vertical traffic flows, respectively. Unlike the proposed cascade model, the single MOHMM learns directly from the observational space and ignores the regional segmentation. Thus, each hidden state of the model consists of 30 observation nodes, where each node corresponds to an atomic event detected. Similar to the setting of cascade model, we used uniform initialisation for the prior distribution π and state transition distribution A for the single MOHMM. The observational distribution B was initialised according to (3).

For the cascade model, the number of hidden states in global model was set to 2 since the global traffic flows have two phases. To obtain the optimal number of hidden

states in each local model, we varied the number of local states from 2 to 10, and observed the matching accuracies of the global traffic phases inferred using the training data in each setting with the ground truth global phases. It was found that local model with five hidden states yielded the best accuracies. Figure 5(a) shows the plot of the matching accuracies averaged over 10 runs along with the standard deviations, against the number of hidden states. Figure 5(b) shows an example of the global phases inferred from the training data and the ground truth. As can be seen, the cascade model managed to learn the implicit temporal transition of the vertical and horizontal traffic phases accurately, achieving accuracy rate of 97.26% when the number of local states was set to 5.



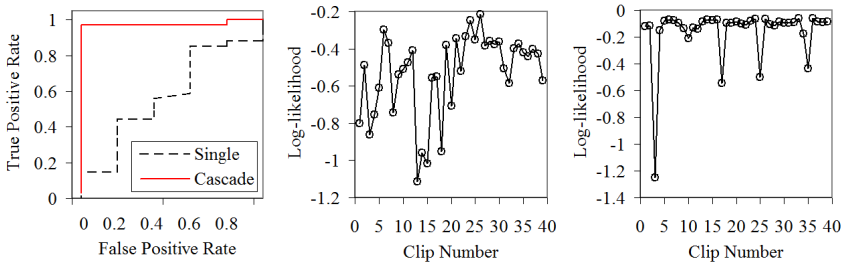
(a) (b)

Fig. 5. (a) The matching accuracies of global traffic phases inferred from training data with the ground truth global phases averaged over ten runs, along with the standard deviations being shown as error bars. (b) Global phases inferred from training data and the ground truth global phases.

We trained the single MOHMM and the cascade model using the training set with 20 iterations of EM learning. Both models were then evaluated using the test set to obtain the log-likelihood values (LL^g in the case of cascade model). The threshold variable TH were varied to obtain the Receiver Operating Characteristic (ROC) curves of the two models, as shown in Fig. 6(a). As can be seen from the ROC curves, the detection result of the proposed cascade model is significantly better than those obtained using the single MOHMM. In particular, the area under ROC (AUROC) achieved by using the single MOHMM was 0.5794, compared with 0.9765 obtained by using cascade MOHMMs.

To further investigate the miss detections and false alarms, the global log-likelihoods (LL^g in the case of cascade model) obtained from the both models are plotted as shown in Fig. 6(b) and 6(c), respectively. From the log-likelihood plots, we observed a high false alarm rate in anomaly detection with the single MOHMM. In particular, video clips 13, 15, 14, 18, 3, and 1 are those clips that returned the lowest log-likelihoods, despite the fact that most of them (except clip 3) are normal. The high false alarm rate suggests that the model is susceptible to noise since it learns directly from the observational space. For instance, in clip 13, vehicles in regions 1 and 4 were mistakenly grouped as a large blob across regions (see Fig. 7). The same error was also observed in clip 14. As a result of imperfect blob detection, the atomic events were grouped into the wrong clusters and consequently led to a false alarm using the single MOHMM. In contrast, clips 13 and 14 yielded much higher log-likelihood values using the cascade

model, indicating the model is able to cope with the erroneous input effectively. Apart from the noise problem, the large observational space per state of the single model implies that a single state has to represent implicitly multiple sources of atomic event variations at any given time instance. Unless we have sufficient amount of clean training instances, the poor structure would leads to ineffective learning of temporal correlation among atomic events. In contrast, as can be seen from Fig. 6(c), the log-likelihood plot of the cascade model is more selective and the model correctly picked up clips 3, 4, 10, 25 and 35 as abnormal behaviours. As expected, the cascade model was able to identified clips 3 and 4 as abnormal behaviour since these behaviours are supported with strong visual evidences. In addition, the proposed model was able to detect clip 25 and 35 as being abnormal, despite the fact that they are rare, ambiguous and harder to be identified without careful examination on the traffic cycles. More importantly, the proposed model successfully discover clip 10 as being abnormal, despite the behaviour is visually much less obvious. This demonstrates the advantage of our cascade model, i.e., although the evidence observed from a single atomic event is inadequate, the model is still able to detect global anomalies by exploiting combinatory temporal evidences from individual atomic events.



(a) ROC Curves (b) Single MOHMM (c) Cascade MOHMMs

Fig. 6. (a) ROC curves for single MOHMM and the proposed cascade model. Subfigures (b) and (c) show the log-likelihood plots for single MOHMM and cascade MOHMMs, respectively.

(a) Clip 13 (b) Clip 14

Fig. 7. Imperfect blob detections result in local atomic events being grouped into wrong clusters. The corresponding bounding box is marked in black colour.

5 Conclusions

We have presented a novel framework for anomaly detection in outdoor crowded scene by modelling temporal correlations of multiple objects. Experiments on complex traffic

scene have demonstrated the promising potential of the proposed model in detecting abnormal behaviours that are ambiguous or supported only by weak visual evidence. This is achieved by modelling temporal correlations among multiple atomic events and integrating lower-level evidences to increase the certainty in detecting anomalies. Our experimental results have also shown that cascade model is more accurate in detecting abnormal behaviours compared with single MOHMM model. The better performance is mainly due to the factorisation of observational space via semantic scene segmentation and the robustness to noise and errors in behaviour representation.

We should point out that different types of HMMs can be employed at both stages depending on detection requirement and scene complexity. The choice of models could range from a simple first-order HMM to more sophisticated models such as Hierarchical Hidden Markov Model (HHMM) or Hidden semi-Markov Model (HSMM). Our ongoing work is focused on investigating the incorporation of duration modelling capability into the framework to achieve a more accurate modelling of behaviour correlations with long-term temporal dependency.

References

1. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. TSMC **34** (2004) 334–352
2. Dee, H.M., Velastin, S.A.: How close are we to solving the problem of automated visual surveillance? Machine Vision and Applications (2007) 1–15 Spec. Issue Paper.
3. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. CVIU **104** (2006) 90–126
4. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. PAMI **23** (2001) 257–267
5. Johnson, N., Hogg, D.C.: Learning the distribution of object trajectories for event recognition. IVC **14** (1996) 609–615
6. Brand, M., Kettnaker, V.: Discovery and segmentation of activities in video. PAMI **22** (2000) 844–851
7. Nguyen, N.T., Venkatesh, S., Bui, H.H.: Recognising behaviours of multiple people with hierarchical probabilistic model and statistical data association. In: Proc. BMVC. (2007)
8. Oliver, N., Rosario, B., Pentland, A.: A bayesian computer vision system for modeling human interactions. PAMI **22** (2000) 831–843
9. Gong, S., Xiang, T.: Recognition of group activities using dynamic probabilistic networks. In: Proc. ICCV. (2003) 742–749
10. Oliver, N., Rosario, B., Pentland, A.: Statistical modeling of human interactions. In IEEE CVPR Workshop on the Interpretation of Visual Motion (1998)
11. Brand, M., Oliver, N., Pentland, A.: Coupled Hidden Markov Models for complex action recognition. In: Proc. CVPR. (1997) 994–999
12. Li, J., Gong, S., Xiang, T.: Global behaviour inference using probabilistic latent semantic analysis. BMVC (2008)
13. Li, J., Gong, S., Xiang, T.: Scene segmentation for behaviour correlation. ECCV (2008)
14. Murphy, K.P.: Dynamic Bayesian Networks: Representation, Inference and Learning. PhD thesis, Uni. of California at Berkeley, Computer Science Division (2002)