# Scene Segmentation for Behaviour Correlation

Jian Li, Shaogang Gong, and Tao Xiang

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

**Abstract.** This paper presents a novel framework for detecting abnormal pedestrian and vehicle behaviour by modelling cross-correlation among different co-occurring objects both locally and globally in a given scene. We address this problem by first segmenting a scene into semantic regions according to how object events occur globally in the scene, and second modelling concurrent correlations among regional object events both locally (within the same region) and globally (across different regions). Instead of tracking objects, the model represents behaviour based on classification of atomic video events, designed to be more suitable for analysing crowded scenes. The proposed system works in an unsupervised manner throughout using automatic model order selection to estimate its parameters given video data of a scene for a brief training period. We demonstrate the effectiveness of this system with experiments on public road traffic data.

## 1 Introduction

Automatic abnormal behaviour detection has been a challenging task for visual surveillance. Traditionally, anomaly is defined according to how individuals behave in isolation over space and time. For example, objects can be tracked across a scene and if a trajectory cannot be matched by a set of known trajectory model templates, it is considered to be abnormal [1,2]. However, due to scene complexity, many types of abnormal behaviour are not well defined by only analysing how individuals behave alone. In other words, many types of anomaly definition are only meaningful when behavioural interactions/correlations among different objects are taken into consideration. In this paper, we present a framework for detecting abnormal behaviour by examining correlations of behaviours from multiple objects. Specifically, we are interested in subtle multiple object abnormality detection that is only possible when behaviours of multiple objects are interpreted in correlation as the behaviour of each object is normal when viewed in isolation. To that end, we formulate a novel approach to representing visual behaviours and modelling behaviour correlations among multiple objects.

In this paper, a type of behaviour is represented as a class of visual events bearing similar features in position, shape and motion information [3]. However, instead of using per frame image events, atomic video events as groups of image events with shared attributes over a temporal window are extracted and utilised

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part IV, LNCS 5305, pp. 383–395, 2008.

as the basic units of representation in our approach. This reduces the sensitivity of events to image noise in crowded scenes. The proposed system relies on both globally and locally classifying atomic video events. 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 decompose semantically the scene into different spatial regions according to the spatial distribution of atomic video events. In each region, events are re-clustered into different groups with ranking on both types of events and their dominating features to represent how objects behave locally within each region. As shown in Section 5, by avoiding any attempt to track individual objects over a prolonged period in space, our representation provides an object-independent representation that aims to capture categories of behaviour regardless contributing objects that are associated with scene location. We demonstrate in our experiments that such an approach is more suitable and effective for discovering unknown and detecting subtle abnormal behaviours attributed by unusual presence of and correlation among multiple objects.

Behavioural correlation has been studied before, although it is relatively new compared to the more established traditional trajectory matching based techniques. Xiang and Gong [3] clustered local events into groups and activities are modelled 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 [4]. 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 [5] attempted modelling scene activities from optical flows 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 anomaly detection is based only on how an individual behaves in isolation. How activities interact in a wider context is not considered. Wang et al [6] proposed hierarchical Bayesian models to learn visual interactions from low-level optical flow features. However, their framework is difficult to be extended to model behaviour correlation across different type of features, in which adding more features would significantly increase complexity of their models.

In our work, we model behaviour correlation by measuring the frequency of co-occurrence of any pairs of commonly occurred behaviours both locally and remotely over spatial locations. An accumulated concurrence matrix is constructed for a given training video set and matched with an instance of this matrix calculated for any testing video clip in order to detect irregular object correlations in the video clip both within the same region and across different regions in the scene. The proposed approach enables behaviour correlation to be modelled beyond a local spatial neighbourhood. Furthermore, representing visual behaviours using different dominant features at different spatial locations makes it possible

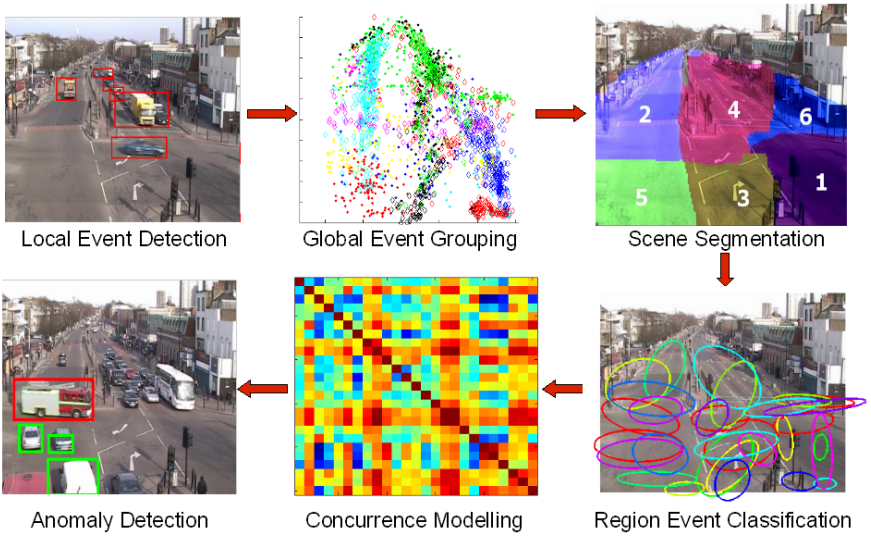



**Fig. 1.** Semantic scene segmentation and behaviour correlation for anomaly detection

to discover subtle unusual object behaviour correlations that either human prior knowledge is unaware of or it is difficult to be defined by human analysis. An overall data flow of the system is shown in Fig. 1.

## 2 Event Detection and Clustering

### 2.1 Image Events

We define an image event as a group of foreground neighbouring pixels detected using background subtraction. Different background models can be adopted. When only moving objects are of interest, we can use a dynamic Gaussian-Mixture background model [7]. As we also want to extract those long-staying objects, an alternative background model [8] is preferred.

Detected foreground pixels are grouped into blobs using connected components, with each blob corresponding to an image event given by a rectangular bounding box. An image event $\mathbf{v}_f$ is represented by a set of 10 features given the membership of a group as follows:

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

where $(x, y)$ and $(w, h)$ are the centroid position and the width and height of the 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. Clearly, some of these features are more dominant for certain image events depending on their loci in a scene, as

they are triggered by the presence and movement of objects in those areas of the scene. However, at this stage of the computation, we do not have any information about the scene therefore all 10 features are used at this initial step to represent all the detected image events across the entire scene.

Given detected image events, we wish to seek a behavioural grouping of these image events with each group associated with a similar type of behaviour. This shares the spirit with the work of Xiang and Gong [3]. However, direct grouping of these image events is unreliable because they are too noisy due to their spread over a wide-area outdoor scene under variable conditions. It has been shown by Gong and Xiang [9] that precision of feature measurements for events affects strongly the performance of event grouping. When processing video data of crowded outdoor scenes of wide-areas, variable lighting condition and occlusion can inevitably introduce significant noise to the feature measurement. Instead of directly grouping image events, we introduce an intermediate representation of atomic video event which is less susceptible to scene noise.

### 2.2 Atomic Video Events

Derived from image events, an atomic video event is defined as a spatio-temporal group of image events with similar features. To generate atomic video events, a video is cut into short non-overlapping clips and image events within a single clip are clustered into groups using K-means. Each group then corresponds to an atomic video event. In our system, we segment a video into clips of equal frame length $N_f$, where $N_f$ is between 100 to 300 depending on the nature of a scene. For K-means clustering in each clip, the number of clusters is set to the average number of image event across all the frames in this clip. An atomic video event is represented by both the mean feature values of all the membership image events in its cluster, and their corresponding variances, resulting in a 20 components feature vector for each atomic video event, consisting of:

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

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).

### 2.3 Event Grouping

We seek a behavioural grouping of all the atomic video events detected in the scene in a 20 dimensional feature space. Here we assume an atomic video event being a random variable following a Mixture of Gaussian (MoG) distribution. We need to determine both the number of Gaussian components in the mixture (model order selection) and their parameters. To automatically determine the model order, we adopt the Schwarz's Bayesian Information Criterion (BIC) model selection method [10]. Given the number of Gaussians $K$ being determined, the Gaussian parameters and priors are computed using Expectation-Maximisation [11]. Each atomic video event is associated with the $k$th Gaussian representing a behaviour class in the scene, $1 \leq k \leq K$, which gives the maximum posterior probability.

## 3 Scene Segmentation

This behavioural grouping of atomic video events gives a concise and semantically more meaningful representation of a scene (top middle plot in Fig. 1). We consider that each group represents a behaviour type in the scene. 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 impose more contextual constraints, we segment a scene semantically into regions according to event distribution with behaviour labelling, as follows.

We treat the problem similar to an image segmentation problem except that we represent each image position by a multivariate feature vector instead of RGB values. To that end, we introduce a mapping procedure transferring features from event domain to image domain. We assign each image pixel location of the scene a feature vector $\mathbf{p}$ with $K$ components, where $K$ is the number of groups of atomic video events estimated for a given scene, i.e. the number of behaviour types automatically determined by the BIC algorithm (Section 2.3). The value of the $k$th component $p_k$ is given as the count of the $k$th behaviour type occurred at this image position throughout the video. In order to obtain reliable values of $\mathbf{p}$, we use the following procedure. First of all, the behavioural type label for an atomic video event is applied to all image events belonging to this atomic video event. Secondly, given an image event, its label is applied to all pixels within its rectangular bounding box. In other words, each image position is assigned with a histogram of different types of behaviours occurred at that pixel location for a given video. Moreover, because we perform scene segmentation by activities, those locations without or with few activities are to be removed from the segmentation procedure. For doing this, we apply a lower bound threshold $TH_p$ to the number of events happened at each pixel location, i.e. the sum of component values of $\mathbf{p}$. Finally the value of this $K$ dimensional feature vector $\mathbf{p}$ at each pixel location is scaled to $[0, 1]$ for scene segmentation.

With this normalised behavioural histogram representation in the image domain, we employ a spectral clustering technique modified from the method proposed by Zelnik-Manor and Perona [12]. Given a scene in which $N$ locations with activities, an $N \times N$ affinity matrix $\mathbf{A}$ is constructed and the similarity between the features at the $i$th position and the $j$th position is computed according to Eqn. (3),

$$\mathbf{A}(i,j) = \begin{cases} exp\left(-\frac{(d(\mathbf{p}_i,\mathbf{p}_j))^2}{\sigma_i \sigma_j}\right) exp\left(-\frac{(d(\mathbf{x}_i,\mathbf{x}_j))^2}{\sigma_x^2}\right), & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\| \leq r \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $\mathbf{p}_i$ and $\mathbf{p}_j$ are feature vectors at the $i$th and the $j$th locations, $d$ represents Euclidean distance, $\sigma_i$ and $\sigma_j$ correspond to the scaling factors for the feature vectors at the $i$th and the $j$th positions, $\mathbf{x}_i$ and $\mathbf{x}_j$ are the coordinates and $\sigma_x$ is the spatial scaling factor. $r$ is the radius indicating a circle only within which, similarity is computed.

Proper computation of the scaling factors is a key for reliable spectral clustering. The original Zelnik-Perona's method computes $\sigma_i$ using the distance between the current feature and the feature for a specific neighbour. This setting is very arbitrary and we will show that it suffers from under-fitting in our experiment. In order to capture more accurate statistics of local feature similarities, we compute $\sigma_i$ as the standard deviation of feature distances between the current location and all locations within a given radius $r$. The scaling factor $\sigma_x$ is computed as the mean of the distances between all positions and the circle center within the radius $r$. The affinity matrix is then normalised according to:

$$\bar{\mathbf{A}} = \mathbf{L}^{-\frac{1}{2}}\mathbf{A}\mathbf{L}^{-\frac{1}{2}} \tag{4}$$

where $\mathbf{L}$ is a diagonal matrix with $\mathbf{L}(s,s) = \sum_{t=1}^{N}(\mathbf{A}(s,t))$. $\bar{\mathbf{A}}$ is then used as the input to the Zelnik-Perona's algorithm which automatically determines the number of segments and performs segmentation. This procedure groups those pixel locations with activities into $M$ regions for a given scene.

## 4 Behaviour Concurrence Modelling

### 4.1 Regional Event Classification

Recall that due to the lack of any prior information at the initial behavioural grouping stage for scene segmentation, all 10 features together with their corresponding variances were used to represent atomic video events. These settings are not necessarily optimal for accurately describing behaviours once the scene has been segmented semantically into regions. To address the problem, we re-classify behaviours in each region. Essentially, we follow the same procedure described in Section 2 but perform an additional computation to refine the grouping of atomic video events in each individual region as follows.

Given a segmented scene, we determine the most representative features in each region by computing entropy values for the features in $\mathbf{v}_f$ in each region and select the top five features with high entropy values. The selected features are then used for grouping image events in each video clip into atomic video events. When representing atomic video events, their corresponding 5 variances are also considered. This results in different and smaller set of features being selected for representing events in different regions. After atomic video event clustering, we obtain $K_m$ regional event classes in each region $m$, where $1 \leq m \leq M$.

### 4.2 Behaviour Correlation Modelling

Suppose we have now obtained in total $K_o$ clusters of atomic video events in all regions, i.e. $K_o = \sum_{m=1}^{M} K_m$, we wish to examine the frequency of concurrence among all pairs of behaviours happened in the scene throughout a video. Given a training video $\mathbf{F}$ which is segmented into $N_c$ non-overlapping clips $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_{N_c}]$, each atomic video event in a single clip $\mathbf{f}_n$, $1 \leq n \leq N_c$, has been clustered to a specific regional event class $\mathbf{b}_i$, where $1 \leq i \leq K_o$. To

indicate the concurrence of a pair of regional event classes $\mathbf{b}_i$ and $\mathbf{b}_j$ occurred in clip $n$, we construct a $K_o \times K_o$ dimension binary matrix $\mathbf{C}_n$ so that

$$\mathbf{C}_n(i,j) = \begin{cases} 1, & \text{if } \mathbf{b}_i = TRUE \text{ and } \mathbf{b}_j = TRUE \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

An accumulated concurrence matrix $\mathbf{C}$ over all the clips in the video is then computed as:

$$\mathbf{C} = \sum_{n=1}^{N_c} \mathbf{C}_n \tag{6}$$

It is clear that the diagonal components of $\mathbf{C}$ indicate the number of occurrence of event class $\mathbf{b}_i$ throughout the video and each other component $\mathbf{C}(i,j)$ corresponds to the total number of concurrence of event classes $\mathbf{b}_i$ and $\mathbf{b}_j$. To normalise the accumulated concurrence matrix $\mathbf{C}$, components in each row of $\mathbf{C}$ is divided by the diagonal component in this row. This results in a non-symmetric normalised matrix $\mathbf{C}_e$. The final symmetric concurrence matrix is computed as:

$$\mathbf{C}_f = \frac{1}{2}(\mathbf{C}_e + \mathbf{C}_e^T), \tag{7}$$

where $T$ is transpose. After re-scaling the values in $\mathbf{C}_f$ to $[0,1]$, $\mathbf{C}_f$ is then used as the model to recognise irregular behaviour labelled atomic video event concurrence. It is worth pointing out that in practice, a measurement of concurrent frequency between a pair of atomic video event classes $\mathbf{b}_i$ and $\mathbf{b}_j$ is meaningful only when $\mathbf{b}_i$ and $\mathbf{b}_j$ individually occur sufficiently frequently. In order to remove those rarely occurred regional event classes from the concurrence matrix during training, we set a lower bound threshold $TH_b$ to the diagonal components of accumulated concurrence matrix $\mathbf{C}$. If $\mathbf{C}(i,i) < TH_b$, the $i$th row and the $i$th column are removed from $\mathbf{C}$. The rectified matrix $\mathbf{C}$ is then used for generating the concurrence matrix $\mathbf{C}_f$.

### 4.3 Anomaly Detection

A test video is segmented into clips in the same way as the training video set. Image events are grouped into atomic video events using K-means. Each atomic video event is then assigned to a regional event class. In order to detect anomaly due to unexpected multi-object behaviour concurrence, we identify abnormal video clips as those with unexpected pairs of concurrences of regional event classes when compared with the concurrence matrix constructed from the training video set. More precisely, for a test video $\mathbf{Q}$ with $N_q$ clips: $\mathbf{Q} = [\mathbf{q}_1, \cdots, \mathbf{q}_{N_q}]$, we generate a binary concurrence matrix $\mathbf{C}_t$ for each clip $\mathbf{q}_t$ by Eqn. (5). We then generate a matrix $\mathbf{CT}_t$ according to Eqn. (8).

$$\mathbf{CT}_t(i,j) = \begin{cases} 1 - \mathbf{C}_f(i,j), & \text{if } \mathbf{C}_t(i,j) = 1 \text{ and } \mathbf{C}_f(i,j) \leq TH_c \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $TH_c$ is a threshold. Given matrix $\mathbf{CT}_t$ for clip $\mathbf{q}_t$, a score $S_t$ is computed as the mean of all the non-zero values in $\mathbf{CT}_t$. Based on the values of $S_t$, $t =$

$1, \cdots, N_q$, those clips with unexpected behavioural concurrence can be identified if the corresponding $S_t$ values are higher than a threshold $TH_s$. In the identified irregular video clips, pairs of unexpected concurrent regional event classes can be further detected as the pairs whose values in $\mathbf{C}_f$ are lower than $TH_c$.

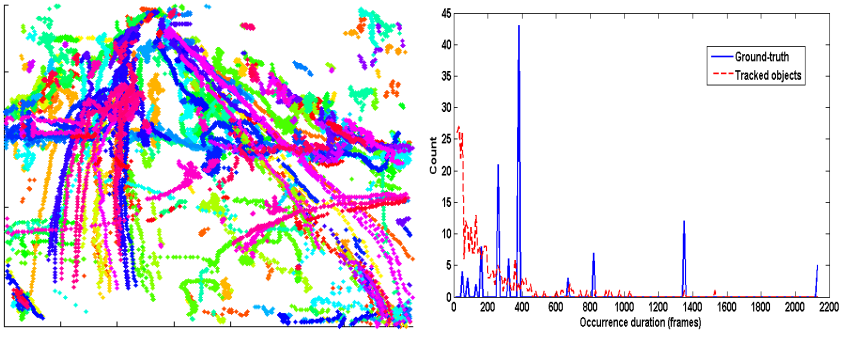
## 5 Experiments

We evaluated the performance of the proposed system using video data captured from two different public road junctions (Scene-1 and Scene-2). Example frames are shown in Fig. 2. Scene-1 is dominated by three types of traffic patterns: the vertical traffic, the leftward horizontal traffic and the rightward traffic, from multiple entry and exit points. In addition, vehicles are allowed to stop between the vertical traffic lanes waiting for turning right or left. In Scene-2, vehicles usually move in from the entrances near the left boundary and near the right bottom corner. They move towards the exits located on the top, at left bottom corner and near the right boundary. Both videos were recorded at 25Hz and have a frame size of $360 \times 288$ pixels.

**Failure Mode For Tracking:** We first highlight the inadequacy of tracking based representation for behaviour modelling in a crowded scene such as Scene-1. Fig. 3 (a) shows the trajectories extracted from a two-minute video clip. In (b), we plot a histogram of the durations of all the tracked object trajectories (red), 331 in total and compare it to that of the ground-truth (blue), which was exhaustively labelled manually for all the objects appeared in the scene (in total 114 objects). It is evident that inevitable and significant fragmentation of object trajectories makes a purely trajectory based representation unsuitable for accurate behaviour analysis in this type of scenes. Moreover, it is equally important to point out that monitoring object in isolation even over a prolonged period of time through tracking does not necessarily facilitate the detection and discovery of unexpected and previously unknown anomaly in a complex scene.

(a) Scene-1 (b) Scene-2

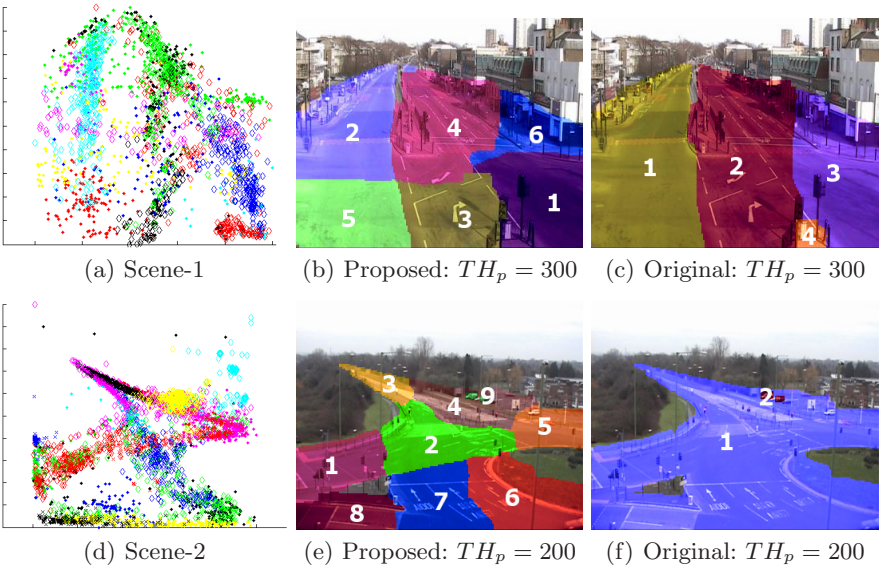
**Fig. 2.** Two public road scenarios for experiment




(a) Trajectories (b) Histogram of duration

**Fig. 3.** Trajectory analysis

**Event Clustering and Scene Segmentation:** In this section, we show the performance of semantic event clustering and scene segmentation. In Scene-1, 22000 frames were used for training, in which 121583 image events were detected and grouped into 2117 atomic video events using K-means. In Scene-2, 415637 image events were detected from 45000 frames and grouped into 4182 atomic video events. The global atomic video events were automatically grouped into 13 and 19 clusters using the EM algorithm where the number of clusters in each scene was automatically determined by the BIC model selection method. The clustering results are shown in Fig. 4 (a) and (d) where clusters are



(a) Scene-1 (b) Proposed: $TH_p = 300$ (c) Original: $TH_p = 300$

(d) Scene-2 (e) Proposed: $TH_p = 200$ (f) Original: $TH_p = 200$

**Fig. 4.** Atomic video event classification and semantic scene segmentation

**Table 1.** Regional feature selection

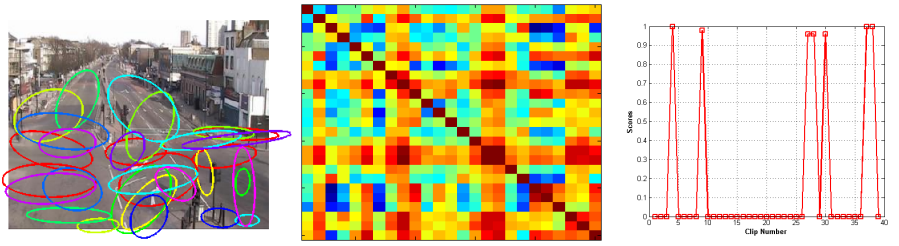
| | $x$ | $y$ | $w$ | $h$ | $r_s$ | $r_p$ | $u$ | $v$ | $r_u$ | $r_v$ |
|---|---|---|---|---|---|---|---|---|---|---|
| R1 | √ | √ | √ | | | √ | √ | | | |
| R2 | √ | √ | √ | | | √ | | √ | | |
| R3 | √ | √ | √ | √ | | √ | | | | |
| R4 | √ | √ | √ | √ | | √ | | | | |
| R5 | √ | √ | | √ | | √ | | √ | | |
| R6 | √ | | | | √ | √ | √ | | √ | |

distinguished by colour and labels. After mapping from feature domain to image domain, the modified Zelnik-Manor and Perona's image segmentation algorithm was then used to segment Scene-1 and Scene-2 into 6 regions and 9 regions, respectively, as shown in Fig. 4 (b) and (e). For comparison, we also segmented the scenes using Zelnik-Manor and Perona's original algorithm (ZP) which resulted in 4 segments for Scene-1 and 2 segments for Scene-2 (Fig. 4 (c) and (f)). It is evident that Zelnik-Manor and Perona's original algorithm suffered from under-fitting severely and was not able to segment those scenes correctly according to expected traffic behaviours. In contrast, our approach provides a more meaningful semantic segmentation of both scenes.

**Anomaly Detection:** We tested the performance of anomaly detection using Scene-1. Comparing to Scene-2, Scene-1 contains more complex behaviour correlations that also subject to frequent deviations from normal correlations. Given the labelled scene segmentation shown in Fig. 4 (b), we re-classified atomic video events in each region. We performed a feature selection procedure which selected the 5 dominant features in each region with largest entropy values. The selected features in each region are shown in Table 1.

Atomic video events were then clustered in each region. From region 1 to region 6, the BIC determined 6, 5, 6, 4, 5 and 4 classes of events (behaviours) respectively. The clustering resulted in 30 local clusters of atomic video events in total (see Fig. 5 (a)). The number of concurrence for each pair of atomic event classes was then accumulated using the 73 clips in the training data to construct a $30 \times 30$ dimension accumulating concurrence matrix $\mathbf{C}$. By removing those behaviour which occurred less than 10 times (i.e. $TH_b = 10$), the dimension of the matrix $\mathbf{C}$ was reduced to $25 \times 25$. The concurrence matrix $\mathbf{C}_f$ was then computed by normalising and re-scaling $\mathbf{C}$ which is shown in Fig. 5 (b).

According to the scores shown in Fig. 5 (c), 7 clips had been picked out of a testing video consisting of 12000 frames (39 clips) as being abnormal with irregular concurrences shown in Fig. 6, in which objects with irregular concurrence are bounded by red and green boxes and the corresponding segments are highlighted using colour. Clip 4 detected a situation when a fire engine suddenly appeared and the surrounding moving vehicles had to stop unexpectedly. In Clip 28, another fire engine appeared. Although the fire engine did not significantly interrupt the normal traffic, it did caused a white van to stop in Region 3 which was not expected to be concurrent with horizontal traffic. A typical example was




(a) Distributions of Local Behaviours (b) Concurrence Matrix (c) Anomaly Scores: $TH_c = 0.12$, $TH_s = 0.9$

**Fig. 5.** Local events classification and anomaly detection. In (a), the mean and covariance of the location of different classes of regional events are illustrated using ellipses in different colour.

(a) Clip 4 (b) Clip 9 (c) Clip 27

(d) Clip 28 (e) Clip 30 (f) Clip 37

(g) Clip 38 (h) Clip 38

**Fig. 6.** Detected irregular concurrences


(a) Anomaly scores (b) False Alarm 1 (c) False Alarm 2

**Fig. 7.** False detections without scene segmentation

detected in Clip 30. Moreover, the second fire engine also caused strange driving behaviour for another car labelled in Clip 28 which strongly conflicted with the normal traffic. In Clip 9 and 37, two right-turn vehicles were detected in Region 2 and Region 5 respectively showing that they were quite close to each other which were not observed in the training data. Clip 27 indicates a false alarm mainly due to the imperfect blob detection which resulted in regional events being classified into wrong classes. In Clip 38, the irregular atomic events were detected in the same clip without frame overlapping (Fig. 6 (g) and (h)). This is an example that when the size of objects are large enough to cover two regions, error could also be introduced as most of vehicles in the training data have smaller size.

For comparison, we performed irregular concurrence detection without scene segmentation, i.e. only using globally clustered behaviours. The results are shown in Fig. 7. Compared with the proposed scheme, the scheme without scene segmentation gave much more false alarms (comparing (a) of Fig. 7 with (c) of Fig. 5). From the examples of false detections in Fig. 7 (b) and (c), it can be seen that using global behaviours without scene decomposition cannot accurately represent how objects behave locally. In other words, each of the global behaviour categories for the vehicles and pedestrians may not truly reflect the local behaviours of the objects and this would introduce more errors in detecting such abnormal correlations of subtle and short-duration behaviours. On the other hand, true irregular incidents were missed, e.g. the interruption from the fire engine was ignored. To summarise, when only using global classification, contextual constraints on local behaviour is not described accurately enough and general global correlation is too arbitrary. This demonstrates the advantage in behaviour correlation based on contextual constraint from semantic scene segmentation.

## 6 Conclusion

This paper presented a novel framework for detecting abnormal pedestrian and vehicle behaviour by modelling cross-correlation among different co-occurring objects both locally and globally in a given scene. Without tracking objects, the system was built based on local image events and atomic video events, which

made the system more suitable for crowded scenes. Based on globally classified atomic video events, a scene was semantically segmented into regions and in each region, more detailed local events were re-classified. Local and global events correlations were learned by modelling event concurrence within the same region and across different regions. The correlation model was then used for detecting anomaly.

The experiments with public traffic data have shown the effectiveness of the proposed system on scene segmentation and anomaly detection. Compared with the scheme which identified irregularities only using atomic video events classified globally, the proposed system provided more detailed description of local behaviour, and showed more accurate anomaly detection and less false alarms. Furthermore, the proposed system is entirely unsupervised which ensures its generalisation ability and flexibility on processing video data with different scene content and complexity.

## References

1. Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., Maybank, S.: A system for learning statistical motion patterns. PAMI 28 (9), 1450–1464 (2006)
2. Johnson, N., Hogg, D.: Learning the distribution of object trajectories for event recognition. BMVC 2, 583–592 (1995)
3. Xiang, T., Gong, S.: Beyond tracking: Modelling activity and understanding behaviour. IJCV 67 (1), 21–51 (2006)
4. Xiang, T., Gong, S.: Video behavior profiling for anomaly detection. PAMI 30(5), 893–908 (2008)
5. Brand, M., Kettnaker, V.: Discovery and segmentation of activities in video. PAMI 22(8), 844–851 (2000)
6. Wang, X., Ma, X., Grimson, W.E.L.: Unsupervised activity perception by hierarchical bayesian models. In: CVPR, Minneapolis, USA, June 18-23, pp. 1–8 (2007)
7. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: CVPR, vol. 2, pp. 246–252 (1999)
8. Russell, D., Gong, S.: Minimum cuts of a time-varying background. In: BMVC, Edinburgh, UK, 1–10 (September 2006)
9. Gong, S., Xiang, T.: Scene event recognition without tracking. Special issue on visual surveillance, Acta Automatica Sinica 29(3), 321–331 (2003)
10. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics 6(2), 461–464 (1978)
11. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, series B 39(1), 1–38 (1977)
12. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: NIPS (2004)