# Discovering Multi-Camera Behaviour Correlations for On-the-Fly Global Activity Prediction and Anomaly Detection *

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

## Abstract

*We propose a unified framework using Latent Dirichlet Allocation (LDA) for discovering behaviour global correlations over a distributed camera network. We explore LDA for categorising object motion patterns as local behaviours in each camera view before correlating these local behaviours globally over different physical locations in multi-camera views. In particular, a Temporal Order Sensitive LDA (TOS-LDA) is formulated to discover behaviour global temporal correlations of different durations among all camera views simultaneously. In addition, a novel online global activity prediction method is proposed based on which global anomalies can be detected on the fly. We validate the effectiveness of our approach using public multi-camera CCTV footages.*

## 1. Introduction

Understanding complex activities and detecting anomalies over a large distributed space, such as a residential building complex or a public infrastructure site, is challenging for computer vision. Due to the nature of an expanded space, such scenarios usually require the installation of distributed multiple CCTV cameras each of which monitors a separate location. To meet this challenge, we consider that solving the following three problems is essential: (1) How to represent reliably object behaviour characteristics in each camera view under difficult and changing viewing conditions due to occlusion, variable lighting and resolution. Visual features and scene complexity can be significantly different in different camera views especially across different physical locations. We need a representational scheme that reflects more about object behaviour characteristics rather than object visual appearances. Such a scheme will be more consistent for all different camera views and less sensitive to variations in viewing conditions. (2) How to correlate local activities observed across all camera views in order to infer a coherent global understanding. This is hard because often only partial visual information is observed in each camera view and meaningful behaviour correlations are visually less well-defined. In typical single camera views, video contents are assumed to be self-contained in the sense that meaningful object behaviour interpretation can be achieved locally within each camera view. In contrast, interpreting local behaviours in a global context across multiple camera views of different locations is inherently more difficult due to greater uncertainties in visual continuity and correlation, e.g. from an object travelling through different views to different objects appearing and behaving across different views introduce significantly different spatial and temporal correlations. (3) How to evaluate on-the-fly global activities and detect anomalies across multi-camera views given partial local observations in each individual camera view. To perform on-the-fly decision making and prediction, a model is required to infer globally temporal correlations among all local behaviours under uncertainty and incompleteness.

To that end, our contributions in this work are three folds. First, we introduce a novel method to informatively and concisely represent visual activities in each camera view through modelling co-occurrences of low-level motion information. Second, we introduce a Temporal Order Sensitive Latent Dirichlet Allocation (TOS-LDA) model to discover any meaningful behaviour global correlations in a multi-camera network. Compared to the conventional Latent Dirichlet Allocation (LDA) model [2] which was designed to extract topics by clustering only co-occurring visual words, the proposed TOS-LDA model is designed to also encode temporal orders among visual words therefore capable of representing both long-scale behaviour co-occurrences and short-scale temporal order dynamics in a single model. Third, we formulate a novel online global anomaly detection method over multi-camera views by continuously evaluating temporal correlations among local behaviours in different camera views. Our experiments demonstrate that our online process is able to achieve comparable performance to that of an offline batch process based model, but with only partial observations running in real-time on-the-fly.

*This work is partially funded by the EPSRC/MOD BEWARE project.

### 1.1. Related Work

Existing work on multi-camera analysis has been focused on two problems: camera topology inference and global activity analysis. Camera topology inference aims to learn connectivities among cameras. Early attempts are based on exhaustive matching of object appearance and estimating a set of constant time delays (plus variances) between entry and exit zones of different camera views through tracking [8, 3]. These methods make strong and often invalid assumptions about object movement characteristics such as speed and trajectory, inter-camera time delay, or object appearance features. More recently, Loy et al. [7] employ Cross Canonical Correlation Analysis (xCCA) to discover inter-camera temporal and causal orders for inferring camera topology. However, xCCA assumes such inter-camera relationships to be single mode and non-variable once discovered. In reality, such relationships can be multi-mode and dynamic. For global activity analysis, Wang et al. [11] employ LDA model to categorise global behaviours through studying co-occurrences of trajectory-based motion patterns in all camera views. However, the trajectory-based representation would limit this method only to scenarios where objects can be reliably tracked. In contrast, our proposed framework does not require any tracking either within or between camera views. Moreover, the LDA model in [11] does not encode any temporal dynamics/delay among activities within the camera network. Although there exists temporal topic models such as [9, 1], they make the first-order Markov assumption for temporal modelling and are unable to model long-term dependency and so cannot detect long-scale temporal anomalies. In our work, both short-scale and long-scale temporal dependences of behaviours within a whole camera network are embedded in the proposed TOS-LDA model and thus is able to detect both types of temporal anomalies. Finally, on-line global activity prediction and anomaly detection are not achievable using the conventional topic models which must wait for the whole video clip to finish before decision making. This problem can be addressed by the proposed TOS-LDA model which enables real-time on-line prediction and anomaly detection with accumulation of partial visual evidence.

## 2. Multi-view Behaviour Representation

We wish to represent all camera views by local behaviours in a common framework to reflect global scene semantics and be less sensitive to image feature noise. To this end, each camera view is segmented independently into semantic regions. Local activities in each region are similar to each other whilst being different from those in other regions. Regional visual words are then extracted from each region and indexed in a common global behaviour representational space.

### 2.1. Low Level Feature Space

First, local motions are computed as our low level features. Specifically, optical flows are computed in each frame for those pixels considered to be 'moving', filtered by frame differencing. Second, this low level feature space is quantised both in location and in motion direction using a codebook. For location, the image plane is uniformly divided into $G$ cells of size $l$ by $l$ pixels (in this paper $l$ is set to 20). Pixels within the same cell will have the same location indexed by $g$ with $1 \leq g \leq G$. The motion directions of the flow vectors are also quantised into $P$ cardinal directions ($P = 4$ in this paper). Now after these quantisations, we have a codebook of $G \times P$ possible visual words to represent the pixel level features, and each 'moving pixel' $v_g^p$ is uniquely labelled by its associated cell $g$, where $1 \leq g \leq G$, and its motion direction $p$ where $1 \leq p \leq P$.

Scene segmentation is performed at the cell level. Each cell could be represented using the low-level motion features. However, using such a low-level feature representation would not give us a semantically meaningful segmentation. To that end, each cell is represented using the inferred local behaviour topics using Latent Dirichlet Allocation (LDA). In particular, video clips of 25 frames long are treated as documents, and moving pixels are treated as words for word-document analysis described next.

### 2.2. Inferring Local Behaviour Topics using LDA

Latent Dirichlet Allocation (LDA) [2] has been widely used for text document analysis aiming to discover semantic topics from documents according to co-occurrences of words. In LDA, a document $\mathbf{w}$ is a collection of $N$ words: $\mathbf{w} = \{w_1, \cdots, w_n, \cdots, w_N\}$ and can be modelled as a mixture of $K$ topics $\mathbf{z} = \{z_1, \cdots, z_K\}$. Each topic is modelled as a multinomial distribution over a vocabulary consisting of $V$ words (in our case $V = G \times P$), from where all words in $\mathbf{w}$ are sampled. Given a corpus of documents for training (i.e. a set of equal length video clips), the LDA model learns the following parameters through variational inference:

1. $\alpha$: a $K$ dimension vector governing the Dirichlet distributions of topics in the corpus;

2. $\beta$: a $K \times V$ dimension matrix representing the multinomial distributions of words in a vocabulary for all learned topics where $\beta_{kv} = P(v|z_k)$ and $\sum_v \beta_{kv} = 1$.

Given the learned model parameters, the log-likelihood for a document $\mathbf{w}$, $\log p(\mathbf{w}|\alpha, \beta)$, is written as:

$$\log p(\mathbf{w}|\alpha, \beta) = \log \int p(\theta|\alpha)(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta))d\theta \quad (1)$$

Another important inference task is to compute the topic profile for a new document $P(\mathbf{z}|\mathbf{w})$, that is the likelihood of each topic featuring in the document. $P(\mathbf{z}|\mathbf{w})$ can be learned through iterations between the following two steps:

$$\phi_{nk} \propto \beta_{kw_n} \exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K}\gamma_k\right)\right) \quad (2)$$

$$\gamma_k = \alpha_k + \sum_{n=1}^{N}\phi_{nk} \quad (3)$$

where $\Psi$ is the first order derivative of a log$\Gamma$ function, variational parameter $\{\gamma_k\}$ approximates the topic distribution $P(\mathbf{z}|\mathbf{w})$, and $\phi_{nk}$ represents how likely a word $w_n$ is associated to a topic $z_k$. Computing both $\log p(\mathbf{w}|\alpha,\beta)$ and $P(\mathbf{z}|\mathbf{w})$ is intractable and variational inference need be used for a solution [2].

LDA is essentially a bag-of-words method that clusters co-occurring words into topics. It provides a more concise way of representing the document than using all the words directly, because the number of topics $K$ is in general much smaller than the size of the codebook/vocabulary $V$.

### 2.3. Semantic Scene Segmentation

Instead of using low level motion features, each cell can be represented more effectively by semantically meaningful local behaviour topics automatically learned using LDA. Specifically, each cell is represented using the likelihood of observing each possible word given each of the $K$ local behaviour topics. There are $P$ different words for the $g$-th cell, each of which is denoted as $v_g^p$ with $1 \leq p \leq P$. The feature vector representing the $g$-th cell is thus written as:

$$\mathbf{f}_g = [P(v_g^1|z_1), \cdots, P(v_g^p|z_k), \cdots, P(v_g^P|z_K)], \quad (4)$$

where $k = 1, \cdots, K$, $P(v_g^p|z_k)$ is the likelihood of observing $v_g^p$ with the $k$-th local behaviour topic $z_k$ and has been learned as part of model parameter $\beta$ (see Sec. 2.2).

With this behaviour topic based representation, the similarity between each pair of cells in the camera view is measured by examining how similar their topic profile feature vectors $\mathbf{f}_g$ are (Eq. (4)). This similarity measure is then used as the input to a spectral clustering algorithm with the number of clusters automatically determined via model selection [13]. Different clusters then correspond to different semantic regions. Note that our scene segmentation method is similar in spirit to that of Li et al. [6]. However, Li et al. segment a scene through clustering distributions of visual features associated with blobs of foreground pixels, which could be inconsistent and unreliable due to scene complexity in different camera views. Moreover, the proposed method is also intrinsically different from the work in [10] in that we use the LDA profiles for representing local behaviours in a scene instead of the quantised low-level motion features in [10].

### 2.4. Global Behaviour Representation

So far the analysis has been done within each camera view independently. Now the behaviours observed cross camera views are to be represented in a global framework. Assume that $K_R$ semantic regions have been segmented in $C$ camera views. These regions are now referred globally as $\{R_r\}$ where $1 \leq r \leq K_R$. We then represent global behaviours using a codebook of $K_R$ visual words. Each of the visual words represents the regional behaviour associated with the $r$-th region. More specifically, a video is split into non-overlapped sliding windows of 25 frames; we compute the number of 'moving pixels' detected in $r$-th region over a sliding window. If the number is higher than a given threshold, a visual word $v_r$ with $1 \leq r \leq K_R$ is extracted. Note that the sliding window is used here to increase the robustness to noise introduced by low-level features.

Our global behaviour representation method is essentially a dimensionality reduction process. The dimension of the feature space has been reduced from the total number of pixels in multiple camera views (in the order of millions typically) to the total number of cells (thousands), then to the number of semantic regions (dozens). The final result is an extremely concise yet semantically meaningful representation of global behaviours upon which our Temporal Order Sensitive LDA (TOS-LDA) modelling is based.

## 3. Multi-Camera Behaviour Correlations

In a conventional LDA model, each visual word is put into a bag (document) and the temporal order information about occurrences of visual words is therefore lost. However, this information is crucial for behaviour modelling as behaviours are dynamic processes where temporal order matters. To overcome this problem, we formulate a Temporal Order Sensitive LDA (TOS-LDA) for behaviour global correlation modelling so to capture the temporal order information whilst keeping the bag of words model structure.

In our model, a document $\mathbf{w}$ corresponds to a video clip of $T$ continuous sliding windows ($T$ is set to 60 in this paper). Each video frame is composed of $C$ camera views and $K_R$ semantic regions. As described above, our visual words are now defined over all regions and the size of the vocabulary or codebook is $K_R$. Now to make our model temporal order sensitive, each visual word is indexed by both the region label $r$ and the sliding window index $t$. This increases the size of codebook/vocabulary $V$ to $T \times K_R$ and we have a vocabulary $\{v_r^t\}$ where $v_r^t$ is the word extracted from the $r$-th region in the $t$-th sliding window, $1 \leq r \leq K_R$ and $1 \leq t \leq T$. With the introduction of sliding window index $t$, a document is now composed of a set of successive sliding windows: $\mathbf{w} = \{\mathbf{w}_1, \cdots, \mathbf{w}_t, \cdots, \mathbf{w}_T\}$ where $\mathbf{w}_t$ is the $t$-th sliding window in the video clip/document. Each sliding window now contains different types of words as the

sliding window index $t$ is different.

Compared to a standard LDA, our TOS-LDA differs mainly in how the document is represented using visual words. The parameter learning and inference methods are identical to that of LDA (see Sec. 2.2). Yet, this simple extension of LDA brings about the crucial benefit of capturing the dynamic nature of visual behaviours and providing a much more powerful solution to global behaviour modelling. In addition, with the temporal order information encoded in the model, TOS-LDA is effective for modelling both long-scale co-occurrences and short-scale temporal order dynamics of local behaviours. In contrast, the conventional LDA is insensitive to the latter because those instantaneous co-occurrences will be overwhelmed by long-scale thus stronger co-occurrences. Furthermore, an online real-time global behaviour prediction and anomaly method is now made possible.

## 4. Online Prediction and Anomaly Detection

Whilst some early attempts have been made for performing online activity evaluation on individual objects in isolation from single camera views [5, 4], it is less obvious how to evaluate global activities observed over different camera views because multi-camera behaviour global correlations are much less structured. In this section, a method is introduced for performing multi-camera online activity evaluation using our TOS-LDA. More precisely, given a trained model, TOS-LDA is used to predict likely global correlations of local behaviours based on partial observations and detect anomalies on-the-fly.

### 4.1. Prediction

Recall that our document $\mathbf{w}$ is a video clip of $T$ successive sliding windows: $\mathbf{w} = \{\mathbf{w}_1, \cdots, \mathbf{w}_t, \cdots, \mathbf{w}_T\}$. Let us first introduce the definition of an *accumulative temporal document*, which is denoted as $\mathbf{w}_{1:t} = \{\mathbf{w}_1, \cdots, \mathbf{w}_t\}$ and composed of all the sliding windows up to index $t$. Clearly, two successive accumulative temporal documents $\mathbf{w}_{1:t}$ and $\mathbf{w}_{1:t+1}$ have the following relationship: $\mathbf{w}_{1:t+1} = \{\mathbf{w}_{1:t}, \mathbf{w}_{t+1}\}$; we also have $\mathbf{w}_{1:T} = \mathbf{w}$.

Given an accumulative temporal document $\mathbf{w}_{1:t}$, we wish to make prediction for the next sliding window $\mathbf{w}_{t+1}$ by evaluating how likely a local behaviour will be observed in each of the $K_R$ regions cross all camera views. This is expressed as $P(\widehat{w}^r_{t+1}|\mathbf{w}_{1:t})$ where $\widehat{w}^r_{t+1}$ is the visual word corresponding to the occurrence of a local behaviour in the $r$-th region in frame $t+1$. Its value is computed as:

$$P(\widehat{w}^r_{t+1}|\mathbf{w}_{1:t}) = \frac{P(\widehat{w}^r_{t+1}, \mathbf{w}_{1:t})}{P(\mathbf{w}_{1:t})}, \tag{5}$$

where $P(\widehat{w}^r_{t+1}, \mathbf{w}_{1:t})$ is the joint probability of $\widehat{w}^r_{t+1}$ and $\mathbf{w}_{1:t}$. The profile of $K$ topics inferred from $\mathbf{w}_{1:t}$, referring as a $K$-component vector $\gamma_{1:t}$ (see Eqs. (2) and (3)), is used to compute $P(\widehat{w}^r_{t+1}, \mathbf{w}_{1:t})$. More precisely, Eq. (5) is rewritten as:

$$P(\widehat{w}^r_{t+1}|\mathbf{w}_{1:t}) = \frac{P(\widehat{w}^r_{t+1}, \mathbf{w}_{1:t}|\alpha, \beta, \gamma_{1:t})}{P(\mathbf{w}_{1:t}|\alpha, \beta, \gamma_{1:t})}. \tag{6}$$

This results in an approximation of the log-likelihood of $P(\widehat{w}^r_{t+1}|\mathbf{w}_{1:t})$ as:

$$\begin{aligned}\log P(\widehat{w}^r_{t+1}|\mathbf{w}_{1:t}) \approx\ & L\left(\gamma_{1:t}, \phi(\widehat{w}^r_{t+1}, \mathbf{w}_{1:t}); \alpha, \beta\right) \\ & - L\left(\gamma_{1:t}, \phi(\mathbf{w}_{1:t}); \alpha, \beta\right), \end{aligned} \tag{7}$$

where $L(*)$ represents the lower bound of $\log P(*)$. In computing $L(\gamma_{1:t}, \phi(\mathbf{w}_{1:t}); \alpha, \beta)$, we follow the standard procedure of variational inference [2] in which $\gamma_{1:t}$ and $\phi(\mathbf{w}_{1:t})$ are inferred through iterative update using Eq. (2) and Eq. (3). For computing $L(\gamma_{1:t}, \phi(\widehat{w}^r_{t+1}, \mathbf{w}_{1:t}); \alpha, \beta)$, we set $\gamma_{1:t}$ as constant and only update $\phi$ using $\widehat{w}^r_{t+1}$ and $\mathbf{w}_{1:t}$ according to Eq. (2). Following the same procedure, the likelihoods of occurrences of local behaviours in all regions in the next sliding window can be computed.

### 4.2. Anomaly Detection

With the online prediction described above, global behaviour anomalies can be detected on-the-fly as follows:

1. At sliding window $t$, using the TOS-LDA model parameters $\alpha, \beta$ to infer the topic profile $\gamma_{1:t}$ using $\mathbf{w}_{1:t}$ (Eqs. (2) and (3));
2. Compute the likelihoods of local behaviour occurrences for all $K_R$ regions in the next sliding window $t+1$ using $\log P(\widehat{w}^r_{t+1}|\mathbf{w}_{1:t})$ (Eq. (7));
3. Given the real observations at time $t+1$, $\mathbf{w}_{t+1}$, compute an anomaly score $A_s = \sum_r \log P(\widehat{w}^r_{t+1}|\mathbf{w}_{1:t})$ for the regions where local behaviours have been observed. This sliding window is deemed as being abnormal if $A_s < Th_f$ where $Th_f$ is a global behaviour anomaly threshold;
4. If frame $t+1$ is abnormal, locate the contributing local behaviours by examining all new observations in $\mathbf{w}_{t+1}$ with corresponding $\log P(\widehat{w}^r_{t+1}|\mathbf{w}_{1:t})$. If $\log P(\widehat{w}^r_{t+1}|\mathbf{w}_{1:t}) < Th_b$, then the corresponding local behaviour is identified as one of the causes of the global anomaly; $Th_b$ is the local anomaly identification threshold.

## 5. Experiments

### 5.1. Dataset and Settings

The proposed approach was evaluated using real-world surveillance videos from 5 cameras monitoring both the inside and outside of a residential building. The recording

lasted for three hours with a frame rate of 25Hz and frame size of 720 × 576 pixels. This gives a total of 15 hours of videos or 1350000 frames. Fig. 1 shows examples of the views with paths of typical and diverse local behaviours in each view and the topology of the camera network. In particular, camera 1, 5 and 4 monitored the front entrance, the lift and back exit connecting stairs of the building. These cameras were connected through camera 2 and 3 monitoring the lobby and lift lobby respectively. Typical global behaviour would be people walking through camera 1 and 2 and either waiting for the lift in 3 or using the staircase through 4. Although the paths of regular behaviours in Fig. 1 seem to suggest that these behaviours are relatively simple, the behaviours in this network can be quite complex and uncertain. For example, people can either wait for the lift in camera 3 or use the stairs in camera 4 depending on which floor the lift was (cannot be detected visually using the 5 views) and which floor they wanted to go. For waiting in the lift lobby, some people preferred to walk around (green arrow in camera 3) whilst other stood still, depending on personal preference. As shown in Fig. 1, most views are non-overlapping and with low resolution. The whole building is poorly lit with unstable lighting, especially in camera 1 and 2. All these conditions make this scenario challenging for modelling behaviour global correlations.

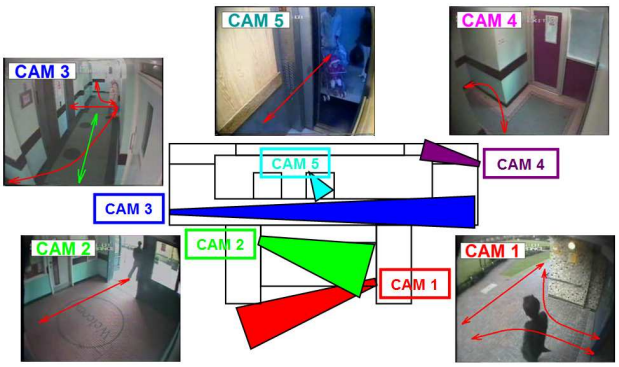



Figure 1. Camera configuration and views in the scenario.

The three-hour long footage of 5 camera views was split into video clips of one-minute long, each containing 60 sliding windows. Our dataset thus consisted of 159 clips in total. After careful human examination, 40 clips were labelled to contain abnormal behaviour correlations and the remaining 119 normal. We randomly selected 79 normal clips for training and the rest 40 normal and 40 abnormal clips for testing. The number of local behaviour topics for each view was set to 10. For global behaviour modelling using TOS-LDA, the number of topics were also set to 10.

## 5.2. Semantic Scene Segmentation

The scene segmentation results are illustrated in Fig. 2(a). For comparison, we also implemented the method proposed in [6] and the results are shown in Fig. 2(b). It is evident from Fig. 2 that our method produced more meaningful segmentation in all views. For instance, the behaviours in camera 3 were far more complex than those in cameras 2 and 5 and the segmentation results using our method reflected such behaviour complexity differences. In contrast, the method in [6] produced more regions in cameras 2 and 5 than camera 3 and thus failed to precisely capture the behaviour complexities in different camera views. Moreover, the method in [6] requires the removal of non-activity pixels beforehand to ensure reasonable performance (camera 4, for example). In contrast, our method does not require any threshold to remove non-activity pixels and is able to produce meaningful segments for both busy and quiet scenes.


(a) Our method


(b) The method in [6]

Figure 2. Comparison of semantic scene segmentations.

## 5.3. Behaviour Global Correlation Modelling

The global topics learned using our TOS-LDA model correspond to typical global behaviour correlations represented as local behaviours occurring according to certain temporal order. An example of the learned global behaviour topics is illustrated in Fig. 3(a) by highlighting the top 2 local behaviours (words) that are most likely to happen in the corresponding sliding window in the topic given by the learned model parameter $\beta$ (see Sec. 2.2). Note that each word is associated with a sliding window index $t$. It can be seen clearly from Fig. 3(a) that this topic corresponds to the global behaviour of people reaching the ground floor via the staircase (camera 4, regions 29, 30), walking pass the lift lobby (camera 3, regions 21, 23) and the front lobby (camera 3, regions 14, 16), and appearing outside the building (camera 1, regions 5, 10). The limited space allows us only show one example of the discovered topics. According to our observation, all other topics are also meaningful and informative.

For comparison, we learned a conventional LDA without introducing the sliding window index $t$ in the visual words. Although similar global topics can be learned, these LDA topics contain no information about the temporal order of the local behaviours. For instance, a topic learned using the LDA is depicted in Fig. 3(b), which corresponds to the same global behaviour as in Fig. 3(a). However, this topic


(a) An example of topic learned using TOS-LDA


(b) An example of topic learned using LDA

Figure 3. Comparing topics learned using TOS-LDA and LDA.


Figure 4. Example activities of people moving in (top row) and moving out (bottom row) of the building using stairs in camera 4.

only suggests that those local behaviours are expected to take place in the same video clip; it says nothing about by what order they are supposed to take place. Fig. 4 shows two different global behaviours. They have different topic profiles (Eq. (2) and Eq. (3)) therefore separable using our TOS-LDA model, whilst having the same profile thus indistinguishable using the LDA (both have the topic in Fig. 3(b) as the dominant topic).

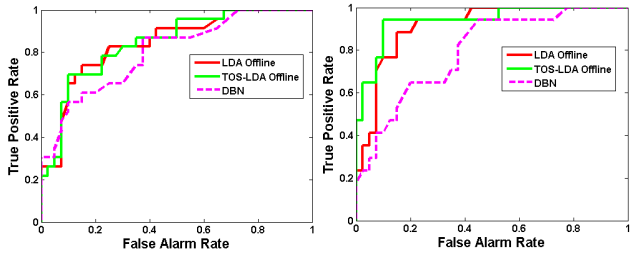
### 5.4. Global Activity Prediction / Anomaly Detection

Three experiments were conducted to compare the performance for global anomaly detection using (1) our TOS-LDA model and a conventional LDA model; (2) our TOS-LDA offline with complete observation (i.e. using $\log p(\mathbf{w}|\alpha, \beta)$ in Eq. (1) as the abnormality measure), and online with only partial observations (using the procedure described in Sec. 4.2); (3) our TOS-LDA model and an alternative Dynamic Bayesian Network based model proposed in [12]. To gain some insight into how different methods perform given different types of global anomalies, we classified the 40 abnormal videos into two categories in which anomalies are mainly caused by: (1) long scale abnormal co-occurrences (23 videos); (2) short scale abnormal temporal order of local behaviours (17 videos). Some examples are shown in Fig. 5. In (a), a group of people moved out of the lift but went to the back exit in camera 4 instead of using the front entrance in camera 1. This results in abnormal co-occurrences of local behaviours over the whole video clip. Fig. 5 (b) shows a rare short scale temporal order anomaly where people loitered in the whole scenario and caused significant unexpected temporal orders between local behaviours across camera views. The results are shown using ROC curves in Figs. 6 and 8.


(a) Long scale abnormal co-occurrences


(b) Short scale temporal order anomaly

Figure 5. Examples of abnormal global behaviours in the multi-camera scenario.




(a) Long scale anomalies (b) Short scale anomalies

Figure 6. Detection performance comparison between TOS-LDA, LDA, and a DBN based method in [12]

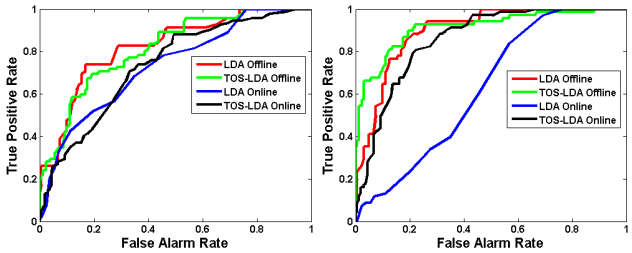
**TOS-LDA vs. LDA** – The results in Fig. 6(a) shows that both the TOS-LDA and the LDA model yielded similar accuracy for detecting long-scale temporal abnormal co-occurrences, whereas our TOS-LDA outperformed the LDA for detecting abnormal temporal order of local behaviours, especially when the false alarm rate was low (see Fig. 6(b)). To examine the cause of such difference, we set the false alarm rate to 5% in Fig. 6 (b) and the corresponding true positive rates are 41% and 64% for the LDA and the TOS-LDA. This results in the correct detection of 7 and 11 abnormal videos respectively out of the 17 anomalies. The 11 detections from the TOS-LDA model included all 7 videos detected by the LDA model. In Fig. 7, we show the synchronised frames extracted from the 4 anomalies that were missed by the LDA model. In these videos, a group of

people walked around randomly across the camera network without clear intention. Consequently the temporal orders structure of behaviour global correlations were significantly different from those of normal global behaviours where people had clear goals of movement. This temporal order structure cannot be captured by a standard LDA resulting in miss-detection.


Figure 7. Four examples of short scale temporal order anomalies missed by the LDA model.

**TOS-LDA vs. DBN** — We tested the performance of TOS-LDA against a Multi-Observation Hidden Markov Model (MOHMM) based Dynamic Bayesian Network (DBN) described in [12]. We used the same representation but extended it to a multi-camera scenario where our aim was to learn the temporal dynamics of topic profiles associated with multiple behaviours in all camera views. Given a training video, we associated each of the local behaviours with a dominant topic learned from the LDA model. The dominant topics in all sliding windows in the video were used as the inputs of the DBN. The results are shown in Fig. 6 and compared with our TOS-LDA model. It is evident that our TOS-LDA significantly outperforms the DBN for both anomaly categories. This is not surprising given the complex and uncertain nature of multi-camera scenarios. Specifically, multi-camera scenarios usually contain significantly uncertainties on the spatial and temporal characteristics. A DBN tends to be over-fitting given sparse data and is more sensitive to noise in behaviour representation. On the other hand, our TOS-LDA is a bag-of-words model which is much less sensitive to noise and more likely to perform well given sparse data.

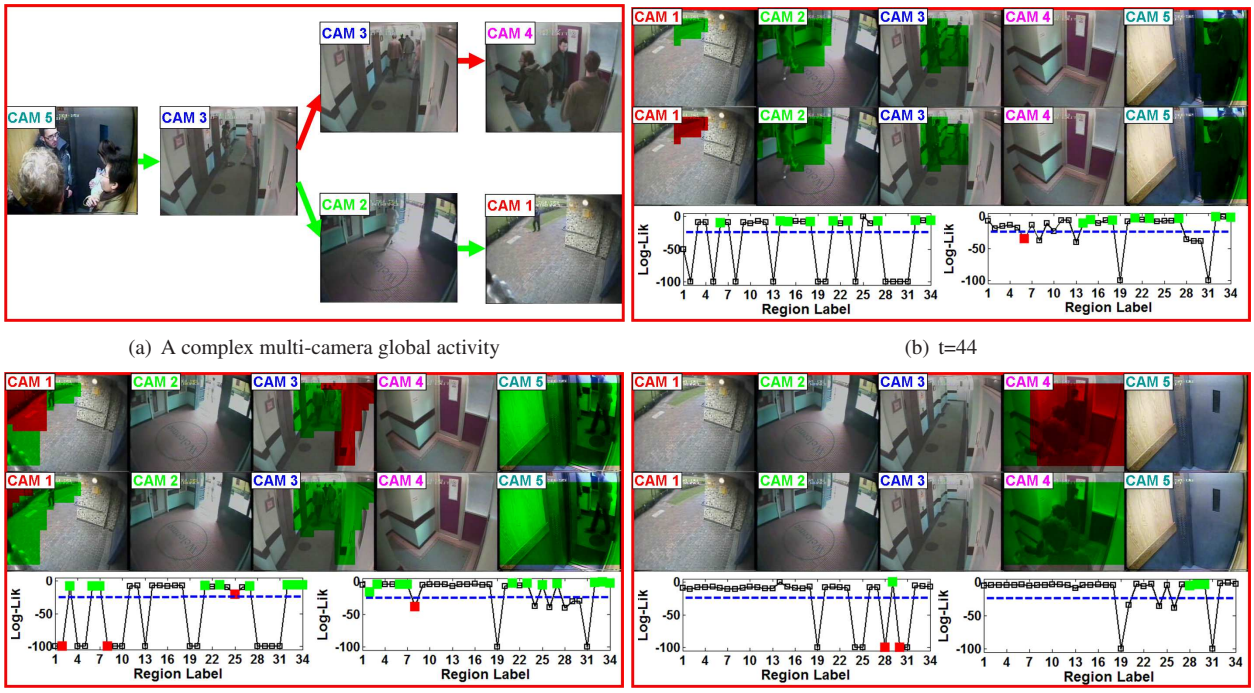
**Online vs. Offline** — Fig. 8 shows that using our online detection procedure, the detection accuracy was degraded. This is expected as the online procedure was only based on partial visual evidences. Nevertheless, the online TOS-LDA detection still gave good accuracy for detecting both long-scale co-occurrence anomalies and short-scale temporal order anomalies whereas the online detection accuracy using the LDA model was significantly worsen, especially for short-scale temporal order anomalies.




(a) Long scale anomalies (b) Short scale anomalies

Figure 8. Detection accuracy comparison between online and offline processes.

Fig. 9 illustrates the online anomaly detection process for a complex behaviour (Fig. 9 (a)): people went down using the lift (camera 5) to the ground floor and then split into two groups: (1) one person walked out of the building following a normal path (cameras 3-2-1) and (2) a group of persons moved to the staircase area in camera 4 following an abnormal path (cameras 3-4). The results for selected frames are shown in Fig. 9 (b)-(d), each of which shows the detected behaviours being classified to normal (green) and abnormal (red) using the TOS-LDA model (top row) and the LDA model (middle row). Clearly, the proposed TOS-LDA was able to produce more accurate detection of the triggered abnormal behaviours in such complex situation involving multiple objects in a distributed camera network. To further investigate the reason, in each figure of (b)-(d), we plotted the predicted log-likelihoods of occurrences of regional behaviours in the corresponding sliding window by using TOS-LDA (bottom left) and LDA (bottom right) where the log-likelihoods above the threshold (blue dash line) indicate the likely occurrences of corresponding regional behaviours and the log-likelihoods corresponding to truly observed regional behaviours are highlighted by green (normal) and red (abnormal). It can be seen that with the increase of available observations, the LDA tended to predict that all regional behaviours were likely to occur except those corresponding to regions without activities (for example region 19 in camera 3). In fact, Eq. (7) evaluated how likely a possible regional behaviour would occur given the available observations. In the LDA model, adding a new word in the temporal document just slightly increased the counts of the words. It thus became less sensitive to subtle changes of profiles of counts as the number of available observations increased. In contrast, topics learned from TOS-LDA model encoded temporal structure of words. Introducing an unexpected new word to a temporal document would therefore significantly affect the likelihood of the new observation. This led to more accurate prediction of how likely regional behaviours would or would not occur in the next sliding window.




(a) A complex multi-camera global activity

(b) t=44

(c) t=47

(d) t=58

Figure 9. Online behaviour prediction and anomaly detection. (a) Illustration of normal behaviours (following the green arrows) and abnormal behaviours (following the red arrows). (b)-(d): Detected behaviours in the sliding windows. Top row and middle row: the detection results using TOS-LDA and LDA. Bottom left and right: predicted log-likelihoods from TOS-LDA and LDA. Behaviours with log-likelihoods above the threshold (blue dash line) are likely to occur and log-likelihoods corresponding to truly occurred behaviours were highlighted by green (normal) and red (abnormal).

## 6. Conclusions

In this paper, we proposed a unified framework using Latent Dirichlet Allocation (LDA) for representing and modelling behaviour global correlations within a distributed camera network. The proposed Temporal Order Sensitive LDA (TOS-LDA) produced superior overall accuracy than that of both the LDA model and a Dynamic Bayesian Network based model for detecting both long-scale co-occurring anomalies and short-scale temporal order anomalies under significant correlation uncertainties. Furthermore, we proposed a novel online behaviour prediction and anomaly detection procedure. Experiments using real-world multi-camera CCTV footages demonstrated its comparable accuracy to off-line batch-mode processing but with a significant advantage of on-the-fly processing for anomaly detection.

## References


[1] D. Blei and J. Lafferty. Dynamic topic models. In *ICML*, Pittsburgh, 2006. 2

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. 1, 2, 3, 4

[3] A. Gilbert and R. Bowden. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In *ECCV*, pages 125–136, Graz, 2006. 2

[4] 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. 4

[5] N. Johnson. *Behaviour Model and Analysis*. PhD thesis, University of Leeds, 1999. 4

[6] J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. In *ECCV*, pages 383–395, Marseille, 2008. 3, 5

[7] C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *CVPR*, Miami, 2009. 2

[8] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *CVPR*, pages 205–210, Washington, 2004. 2

[9] L. Ren, D. Dunson, and L. Carin. The dynamic hierarchical Dirichlet process. In *ICML*, Helsinki, 2008. 2

[10] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception by hierarchical bayesian models. In *CVPR*, pages 1–8, Minneapolis, June 2007. 3

[11] X. Wang, K. Tieu, and E. Grimson. Correspondence-free activity analysis and scene modeling in multiple camera views. *PAMI*, 1(1):1–17, 2009. 2

[12] T. Xiang and S. Gong. Activity based surveillance video modelling. *Pattern Recognition*, 41(7):2309–2326, 2008. 6, 7

[13] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, 2004. 3