Int J Comput Vis
DOI 10.1007/s11263-011-0487-2

Learning Behavioural Context

Jian Li · Shaogang Gong · Tao Xiang

Received: 3 December 2009 / Accepted: 28 July 2011

Abstract We propose a novel framework for automatic discovering and learning of behavioural context for video-based complex behaviour recognition and anomaly detection. Our work differs from most previous efforts on learning visual context in that our model learns multi-scale spatio-temporal rather than static context. Specifically three types of behavioural context are investigated: *behaviour spatial context*, *behaviour correlation context*, and *behaviour temporal context*. To that end, the proposed framework consists of an activity-based semantic scene segmentation model for learning behaviour spatial context, and a cascaded probabilistic topic model for learning both behaviour correlation context and behaviour temporal context at multiple scales. These behaviour context models are deployed for recognising non-exaggerated multi-object interactive and co-existence behaviours in public spaces. In particular, we develop a method for detecting subtle behavioural anomalies against the learned context. The effectiveness of the proposed approach is validated by extensive experiments carried out using data captured from complex and crowded outdoor scenes.

Keywords Visual context · Behavioural context · Video-based behaviour recognition · Activity-based scene segmentation · Cascaded topic models · Anomaly detection

J. Li (✉) · S. Gong · T. Xiang
School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK
e-mail: jianli@eecs.qmul.ac.uk

S. Gong
e-mail: sgg@eecs.qmul.ac.uk

T. Xiang
e-mail: txiang@eecs.qmul.ac.uk

1 Introduction

Visual context is the environment, background, and settings within which objects and associated events are observed visually. Humans employ visual context extensively for both object recognition in a static setting and behaviour recognition in a dynamic environment. For instance, for object recognition we can differentiate and recognise whether a hand-held object is a mobile phone or calculator by its relative position to other body parts (e.g. closeness to the ears), even though they are visually similar and partially occluded by the hand. Similarly for behaviour recognition, the arrival of a bus can be detected/inferred just by looking at the passengers' behaviour at a bus stop. Indeed, extensive cognitive, physiological and psychophysical studies have shown that visual context plays a critical role in human visual perception (Palmer 1975; Biederman et al. 1982; Bar and Ullman 1993; Bar and Aminof 2003; Bar 2004). Motivated by these studies, there is an increasing interest in exploiting contextual information for computer vision tasks such as object detection (Heitz and Koller 2008; Murphy et al. 2003; Kumar and Hebert 2005; Carbonetto et al. 2004; Wolf and Bileschi 2006; Rabinovich et al. 2007; Gupta and Davis 2008; Galleguillos et al. 2008; Zheng et al. 2009), action recognition (Marszalek et al. 2009) and tracking (Yang et al. 2008; Ali and Shah 2008).

Previous studies on visual context are predominantly focused on static visual context particularly regarding the scene background, scene category, and other co-existing objects in a scene. However, for understanding object behaviour in a crowded space, the most relevant visual context is no longer static due to the non-stationary background and non-rigid relationships among co-existing objects in a public space. In particular, a meaningful interpretation of object behaviour depends largely on knowledge of spatial

Published online: 27 August 2011

(a) The two traffic phases

(b) A behavioural anomaly

Fig. 1 The behaviour of an object in a traffic junction needs to be understood by taking into account its spatial context (i.e. where it occurs), temporal context (i.e. when it takes place), and correlation context (i.e. how other correlated objects behave in the same shared space). In (**b**) a fire engine moved horizontally and broke the vertical traffic flow. This is an anomaly because it is incoherent with both the temporal context (it happens during the vertical traffic phase) and correlation context (horizontal and vertical traffic are not expected to occur simultaneously)

and temporal context defining where and when it occurs, and correlation context specifying the expectation inferred from the correlated behaviours of other objects co-existing in the same scene. In this work, we propose a novel framework for unsupervised discovery and learning of object behavioural context for context-aware interactive and group behaviour modelling that can facilitate the detection of global anomalies in a crowded space.

Let us first define what constitutes behavioural context. In this work, we consider three types of behavioural context:

1. *Behaviour Spatial Context* provides situational awareness about *where* a behaviour is likely to take place. A public space serving any public function such as a road junction or a train platform can often be segmented into a number of distinctive zones within which behaviours of certain characteristics are expected in one zone but differ from those observed in other zones. We call these behaviour sensitive zones *semantic regions*. For instance, in a train station behaviours of passengers on the train platform and in front of a ticket machine can be very different. Another example can be seen in Fig. 1 where different traffic zones/lanes play an important role in defining how objects are expected to behave.
2. *Behaviour Correlation Context* specifies *how* the meaning of a behaviour can be affected by those of other objects either nearby in the same semantic region or further away in other regions. Object behaviours in a complex scene are often correlated and need be interpreted together rather than in isolation, e.g. behaviours are contextually correlated when they occur concurrently in a scene. Figure 1 shows some examples of moving vertical traffic flow with standby horizontal flow typically co-occurring, with the exception of emergency vehicles running a red light (see Fig. 1(b)). It is important to note that in a complex dynamic scene composed of multiple semantic regions, there are behaviour correlation context at two scales: local correlation context within each region, and global context that represents behaviour correlations across regions.
3. *Behaviour Temporal Context* provides information regarding *when* different behaviours are expected to happen both inside each semantic region and across regions. This is again illustrated by the examples shown in Fig. 1 where behaviour temporal context is determined by traffic light phases. More specifically, meaningful interpretation of vehicle behaviour needs to take into account temporal phasing of the traffic light, e.g. a vehicle is expected to be moving if the traffic light governing its lane is green and stopped if red. Even if the traffic light is not directly visible in the scene, this translates directly to observing how other vehicles move in the scene, i.e. the expected traffic flow. Similar to behaviour correlation context, the temporal context also has two scales corresponding to within-region and across-region context. Successful identifications of such temporal context in both local regions and globally in a scene are cru-

cial for establishing behaviour constraints at different scales which play a key role in identifying abnormal behaviours.

To model and infer these three types of behavioural context, a novel context learning framework is proposed consisting of two key components:

(a) *A behaviour-based semantic scene segmentation model* for learning automatically behaviour spatial context. Given a crowded public scene, we decompose the scene into a number of disjoint local regions according to how different behaviour patterns are observed spatially over time. To that end, the problem of semantic scene segmentation is treated as an image segmentation problem and solved by employing a spectral clustering algorithm with the number of clusters determined automatically. However, different from conventional image segmentation methods (Shi and Malik 2000; Malik et al. 2001) where each pixel location is represented using static visual appearance features of colour and texture, we represent each pixel location using a behaviour-footprint. To obtain this behaviour-footprint, object behaviours are represented as classes of scene events. Each event class corresponds to the behaviour of a group of objects with a certain size and specific motion directions. The occurrences of different classes of object behaviours accumulated over time are then used to compute the behaviour-footprint. The similarity of footprints between two different pixel locations determines whether they should be grouped into the same semantic region.

(b) *A cascaded probabilistic topic model* for learning both behaviour correlation context and behaviour temporal context. Probabilistic topic models (PTM) (Blei et al. 2003; Teh et al. 2006) such as Latent Dirichlet Allocation (LDA) (Blei et al. 2003) and Probabilistic Latent Semantic Analysis (pLSA) (Hofmann 1999a, 1999b) are Bag of Words (BoW) models traditionally used for capturing co-occurrence of text words in document analysis. In this work, we explore topic models for learning behaviour correlation context with 'words' and 'documents' corresponding to visual events and video clips respectively. We choose topic models because that visual features extracted from a crowded dynamic scene are inevitably noisy, and a BoW model such as a topic model is intrinsically more robust against input noise. However, standard topic models are contextually unscalable, i.e. they are unable to capture and differentiate between local (within region) and global (across region) behavioural context. To address this problem, we propose a Cascaded Latent Dirichlet Allocation (Cas-LDA) model employing two stages of topic modelling in a cascade. The first stage model captures behaviour correlation context within each semantic region via inferred hidden (local) topics. The inferred local context is then used as inputs to a second stage model which aims to capture global correlation context. In addition to correlation context modelling, the proposed cascaded topic model is also exploited for inferring temporal behavioural context. Specifically, topic profiles inferred from both stages of our cascade topic model at any given time are used to represent the temporal characteristics of behaviours and infer any temporal behavioural context both locally (within each region) and globally (across regions).

A practical aim for learning behavioural context is to be able to detect subtle and non-exaggerated abnormal behaviours in public space. A behavioural anomaly captured in video from a crowded public space is only likely to be meaningful when detected and interpreted in a context. We argue that such context is necessarily intricate at different levels for a crowded space and not easily specified, if at all possible, by hard-wired top-down rules, either exhaustively or partially. In particular, an identical behaviour in a public space can be deemed either normal or abnormal depending on when, where, and how other objects behave. For instance, a person running on a platform with train approaching and all other people also running is normal, whilst the same person running on an empty platform with no train in sight is more likely to be abnormal. In other words, a behavioural anomaly can be defined and measured as *contextually incoherent*, that is, behaviour that cannot be predicted nor explained away using the learned or inferred behavioural context. To that end, given learned behavioural context as an intricate part of our model, the model is more adept at detecting subtle and unpredictable (unknown) behavioural anomalies that are otherwise undetectable. The effectiveness of the proposed approach is validated through extensive experiments carried out using complex and crowded outdoor scenes.

The rest of the paper is organised as follows: Sect. 2 reviews related work to highlight the contributions of this work. Section 3 addresses the problem of behaviour representation. A behaviour-based semantic scene segmentation model is described for learning behaviour spatial context in Sect. 4. To that end, we also formulate a spectral clustering algorithm. Section 5 centres on a novel cascaded topic model used for learning both behaviour correlation and temporal context. A context-aware video behaviour anomaly detection method is described in Sect. 6. In Sect. 7, the effectiveness and robustness of our approach is evaluated extensively by a series of experiments using data from three different outdoor public scenes. We draw conclusions in Sect. 8.

2 Related Work

Existing studies on visual context modelling are dominated by the modelling of static scene or object appearance context for object detection in a static image. Objects in a scene captured in an image can be divided into two categories (Heitz

and Koller 2008): monolithic objects, or "things" (e.g. cars and people), and regions with homogeneous or repetitive patterns, or "stuffs" (e.g. roads and sky). Consequently, there are Scene-Thing (Murphy et al. 2003), Stuff-Stuff (Singhal et al. 2003), Thing-Thing (Rabinovich et al. 2007), and Thing-Stuff (Heitz and Koller 2008) context depending on what the target objects are and where the context comes from. In this work, the focus is on dynamic behavioural context in video. Therefore different terminologies and definitions are necessary. Nevertheless it is useful to draw an analogy to the static context for easier understanding. For instance, behaviour correlation context can be seen as an extension of the Thing-Thing context in the space and time domain. The spatial context shares similarity with the Scene-Thing and Thing-Stuff context. The temporal context, on the other hand, is unique to video.

More recently, Marszalek et al. (2009) proposed to exploit static context for dynamic scene understanding. Specifically scene context is represented as scene categories and used for action recognition. However, compared to the behaviour spatial context studied in this work, the scene category information used in Marszalek et al. (2009) carries little information for understanding how objects are expected to behave at different regions of a scene. Moreover, action recognition does not address the problem of behaviour understanding: a person walking in different scenes can be interpreted as very different behaviours. Similar, our previous work (Xiang and Gong 2006a) also considered categories of facial expressions and scene events as visual context. However, as in Marszalek et al. (2009), categories of events and facial actions only convey object-centred and isolated contextual information with little if any measure of spatial, correlation and temporal context that define the changing environment (scene context) within which object behaves. Alternatively, Yang et al. (2008) proposed a method for visual tracking by employing context extracted from the so-called auxiliary objects. This context is dynamic in nature as it is associated with the tracked object and of similar motion characteristics. However, the problem of visual tracking differs significantly from that of behaviour interpretation, and behavioural context is not limited to objects of similar motion directions (e.g. in the example in Fig. 1 stopped vehicles can provide useful contextual information for vehicles that are moving). Another work on tracking with context modelling is presented in Ali and Shah (2008) where floor fields are computed to assist in tracking objects in a very crowded scene. Apart from modelling context for different objectives (tracking vs. behaviour understanding and anomaly detection), our approach differs significantly in that we aim to learn both spatial and temporal dynamic behavioural context at different scales, rather than focusing only on correlations of objects as in Ali and Shah (2008). To our best knowledge, our work is the first attempt to (a) systematically model behavioural context by learning from data of complex scenes without exhaustive top-down hard-wired rules, (b) provide an effective solution for learning different aspects of behavioural context in a principled manner, and (c) demonstrate the usefulness and importance of context learning for behavioural anomaly detection.

There have been some efforts on automatic discovery of the layout of a dynamic scene which is closely related to one aspect of the problem studied here, namely learning behaviour spatial context. To that end, there are techniques focusing on learning two specific types of scene layout, entry and exit points/zones (Breitenstein et al. 2008; Wang et al. 2006, 2008; Makris et al. 2004), and static occlusion zones (Greenhill et al. 2008). There are also techniques attempting at capturing a broader range of scene layout such as routes, junctions and paths (Makris and Ellis 2005; Wang et al. 2010). Most existing techniques rely on object trajectory based scene representations and the motivation for learning scene layout is to achieve more robust tracking between camera views or under occlusion. However, visual tracking is intrinsically limited especially in crowded scenes when object visual appearance is no longer continuous or undergoing smooth change, an underlying assumption for establishing visual tracking. Moreover, due to scene complexity, realistic abnormal behaviours are often not well defined by object trajectories alone. In particular, the context from which behaviour can be interpreted meaningfully is not only object-centred but also spatial, correlation and temporal among objects in shared space. Trajectory based scene model is thus often insufficient for behaviour anomaly detection. To address this problem, the proposed behavioural context learning model does not rely on object tracking nor establishes continuous trajectory, and therefore can be applied to crowded scenes with severe occlusions.

There exist a number of approaches on behaviour correlation modelling, which can potentially be used for correlation context learning and inference. Xiang and Gong (2006b) employ Dynamic Bayesian Networks (DBNs) to learn temporal relationships among scene events. Temporal dynamics of each event type is represented by a Hidden Markov Model (HMM) and the learned topology of a DBN reveals the temporal/causal relationships among different events. However, learning a DBN with multiple temporal processes is computationally very expensive. In particular, learning the topology of a DBN from data rapidly becomes intractable as the number of temporal processes representing event classes increases. To overcome this problem, Wang et al. (2009) adopted hierarchical probabilistic topic models (PTMs), including a Hierarchical Dirichlet Processes (HDP) mixture model and a Dual Hierarchical Dirichlet Processes (Dual-HDP) model, to categorise global visual behaviours using three levels of abstraction from low-level motion patterns, atomic activities to high-level behaviour interactions. Compared to DBNs, topic models are less demanding computationally and also less sensitive to input noise due to their

Bag of Words nature, although it loses sensitivity on coping with scale variations in addition to throwing away any temporal order information. Compared to the hierarchical PTM of Wang et al., our cascaded topic models have two desirable features: (1) We decompose a complex dynamic scene into semantic regions by learning behaviour spatial context. Consequently, behaviour correlation context is modelled at both the local (within region) and global (across region) scales given the learned spatial context. This lead to better behaviour understanding and anomaly detection, as demonstrated in our experiments (see Sect. 7.6). (2) Using a cascade of simple topic models is computationally more efficient than using a single and necessarily complex hierarchical topic model.

A PTM is essentially a Bag of Words model that ignores temporal order information for the gain of robustness against noise and input errors. Recently efforts have been taken to introduce dynamics modelling into a topic model in order to make the model sensitive to dynamics of behaviour, whilst keeping the robustness of a topic model. The result is a hybrid model of PTM and DBN, or a dynamic topic model, with a hierarchical model structure. Hospedales et al. (2009) proposed a Markov Clustering Topic Model (MCTM) for modelling video behaviours. Using a MCTM, temporal orders of documents are modelled explicitly. This model was further extended by Kuettel et al. (2010) who developed a Dependent Dirichlet Process Hidden Markov Model (DDP-HMM) to model the temporal order information at both the topic and document levels. With temporal information modelled explicitly, in theory, a dynamic topic model is well suited for detecting behavioural anomalies which could be caused by abnormal temporal orders between behaviours. However, in practice, it is difficult to strike the right balance between model sensitivity and robustness. Importantly, both models are hierarchical models; therefore they suffer from the same problem as the HDP based PTM models in Wang et al. (2009), when applied to abnormal behaviour detection. That is, in a hierarchical model the numbers of inputs in different layers from bottom to top are extremely imbalanced. For example for both DDP-HMM and MCTM, the bottom layer models video words which are in the order of thousands per clip. The number of actions/topics in the layer above are in the order of dozens, whilst the number of temporal phases in the top layer is only a handful. This imbalanced modelling structure may cause problems in detecting abnormities because those occurred at upper layers can be easily overwhelmed by those in the bottom layer, and become undetectable.

The original ideas of semantic scene segmentation (Li et al. 2008a) and anomaly detection using a cascaded topic model (Li et al. 2008b) were introduced by our early work. In this paper, we provide a more coherent and complete treatment on learning behavioural context for anomaly detection in a single framework with extended comparative evaluations against alternative models including the recently proposed hierarchical topic models by Wang et al. (2009) and dynamic topic model by Hospedales et al. (2009). We also analyse the pros and cons of different topic models for different behaviour recognition tasks, with additional implementation details.

3 Behaviour Representation

An event-based behaviour representation is adopted. We consider visual events as significant scene changes occurring over a short temporal window and characterised by the location, shape and motion information associated with each change. These visual scene events are object-independent and location specific and their detection does not rely on object segmentation and tracking. We further consider that visual behaviours are different collections of categorised and labelled scene events. But let us first consider in more details on how we compute scene events.

Suppose a long continuous video of a scene is split into non-overlapping clips. We consider this long video as the training data for learning a model. Within each clip, scene events are first detected and represented in each image frame in isolation. Specifically, this is computed as follows: (1) Foreground pixels are identified by employing a background subtraction method (Russell and Gong 2006). (2) These foreground pixels are then grouped into blobs using connected components and each blob corresponds to a scene event and is assigned a rectangular bounding box. (3) The scene events detected in each frame, or frame-wise events, are represented as a 10-D feature vector:

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

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 median optic flow vector for all foreground pixels in a blob computed using (Lucas and Kanade 1981), $r_u = u/w$ and $r_v = v/h$ are the scaling features between motion information and blob shape. Note that different features have very different value ranges. Before clustering the events to discover groupings, all features are normalised to the range of $[0, 1]$.

However, scene events detected in each frame in isolation are inevitably noisy due to image noise, occlusion between objects and non-stationary background clutter. To minimise error in scene event representation, the detected frame-wise events from each clip are clustered into groups to form clip-wise events. The mean and variance of the 10 event features from each group are employed to represent the corresponding clip-wise event. More specifically, the clustering within

Algorithm 1: Behaviour representation using events.

Input: A continuous video $\mathbf{V}$, divided into T non-overlapping video clips $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_t, \ldots, \mathbf{v}_T\}$. The t-th clip $\mathbf{v}_t$ consisting of F image frames: $\mathbf{v}_t = [\mathbf{I}_{t1}, \ldots, \mathbf{I}_{tf}, \ldots, \mathbf{I}_{tF}]$.

Output: A set of clip-wise events computed from each clip

```
for t = 1 to T do
    for f = 1 to F do
        In the f-th frame of the t-th clip, extract frame-wise events by grouping foreground pixels into blobs;
        Represent each frame-wise event using (1);
    end
    Cluster all frame-wise events from the t-th clip into clip-wise events;
    Represent each clip-wise event using (2);
end
During training, cluster all clip-wise event from V using GMM;
During testing, assign each clip-wise event with a class label using the learned GMM;
```

a clip is by K-means with the number of clusters being automatically determined as the average number of scene events detected per frame over the clip. Given the clustering of all detected frame-wise events in each clip, each clip-wise event, denoted as $\mathbf{e}$, is represented by a 20-D feature vector:

$$\mathbf{e} = [\mathbf{e}_m, \mathbf{e}_v], \tag{2}$$

where $\mathbf{e}_m$ and $\mathbf{e}_v$ correspond to the mean and variance of the 10 features (see (1)) from all the frame-wise events in each group over each clip. With this representation, scene events are no longer solely dependent on local features computed from individual frames. Instead, they are represented by different group average and variance over a short video clip. Consequently, it is much more robust against noise without losing frame locality within each clip.

Clip-wise scene events from all the clips of a set of training videos are further clustered by a Gaussian Mixture Model (GMM) with the number of clusters (V) *automatically determined* using the BIC model selection score (Schwarz 1978). This aims to categorise all the scene events into a finite set of groups, similar to the idea of determining typical words from a given document. Each scene event is then assigned a class label for a particular cluster. A pseudo code for computing events from video clips are provided in Algorithm 1.

The motivations of taking a two-staged clustering approach for event computation are two-fold. First, we aim to remove outliers in the event computed at the frame level so that our representation is more robust against noise. Second, after grouping frame-wise events into clip-wise events, each clip-wise event is represented by both the mean of the corresponding group of frame-wise events, and their variance (see (2)). The variance captures information about how a group of frame-wise events evolve over the duration of a video clip. This implicit temporal information is useful for describing behavioural characteristics.

This representation of behaviour by globally categorised scene events independent from object types has a number of advantages over existing methods. First, compared with the object-centred trajectory based methods (Breitenstein et al. 2008; Wang et al. 2008; Makris et al. 2004), it does not require object segmentation and tracking. This makes it more suitable for crowded scenes where tracking is severely limited intrinsically. Second, compared to other scene event detection methods (Xiang and Gong 2006b; Wang et al. 2009), our scene events are constructed by richer and more reliable features, and smoothed over a temporal window (non-overlapping clip) for more robustness against local noise.

4 Learning Behaviour Spatial Context

We consider that behaviour spatial context is given as disjoint semantic regions where behaviour patterns observed within each region are similar to each other whilst being dissimilar to those occurring in other regions. Moreover, we wish to discover such semantic regions automatically and unsupervised from data. We address this problem using a two-steps approach as follows: (1) Each pixel location in the scene is labelled by a behaviour-footprint measuring how different behaviour patterns occur over time. This pixel-wise location behaviour-footprint is necessarily a distribution measurement, e.g. a histogram of scene event classes occurring at each location over time (Fig. 2). (2) Given behaviour-footprints computed at all pixel locations of a scene, a spectral clustering algorithm is employed to segment all pixel locations by their behaviour-footprints into different non-overlapping regions with the optimal number of regions determined automatically. Let us describe in more details as follows.

Behaviour-Footprint First, a behaviour-footprint is computed for each pixel location in a scene. As described in Sect. 3, behaviours are represented by object-independent scene events categorised in space and over time. Suppose there are V classes of scene events, to measure how different behaviours have taken place at a pixel location in a video, its behaviour-footprint is computed as a histogram of V bins:

$$\mathbf{p} = [p_1, \ldots, p_v, \ldots, p_V] \tag{3}$$

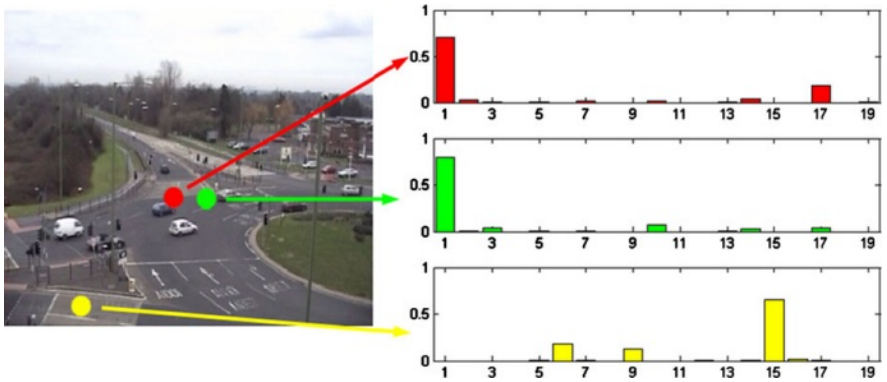


Fig. 2 (Color online) Examples of behaviour-footprint. Three pixel locations and their corresponding behaviour-footprints are shown colour-coded. It is evident that the *red* and *green* locations have similar behaviour-footprints that differ from the *yellow* location

where p_v counts for the number of occurrence of the v-th event class at this pixel location. Figure 2 shows examples of computed behaviour-footprints at different locations of a scene.

Scene Segmentation Second, learning behaviour spatial context is treated as a segmentation problem where the image space is segmented by pixel-wise feature vectors given by the V dimensional behaviour-footprints (see (3)). For segmentation, we consider to employ the spectral clustering model of Zelnik-Manor and Perona (2004). Given a scene with N pixel locations, an $N \times N$ affinity matrix $\mathbf{A}$ is constructed and the similarity between the behaviour-footprints at the i-th and j-th locations is computed as:

$$\mathbf{A}(i,j) = \begin{cases} \exp(-\frac{(d(\mathbf{p}_i,\mathbf{p}_j))^2}{\sigma_i\sigma_j})\exp(-\frac{(d(\mathbf{x}_i,\mathbf{x}_j))^2}{\sigma_x^2}), \\ \quad \text{if } \|\mathbf{x}_i - \mathbf{x}_j\| \le r, \\ 0, \quad \text{otherwise} \end{cases} \tag{4}$$

where $\mathbf{p}_i$ and $\mathbf{p}_j$ are the behaviour-footprints at the i-th and the j-th pixel loci, d represents Euclidean distance, σ_i and σ_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 σ_x is the spatial scaling factor. r is the radius indicating a circle only within which, similarity is computed.[1] Using this model, two pixel locations will have strong similarity and thus be grouped into one semantic region if they have similar behaviour-footprints and are also close to each other spatially, e.g. the red and green dots in Fig. 2.

For spectral clustering, choosing correct scaling factors is critical for obtaining meaningful segmentation. The Zelnik-Perona model computes σ_i using a pre-defined constant distance between the behaviour-footprint at the i-th pixel location and its neighbours, which can be rather arbitrary and sensitive outliers. This results in mostly under-fitting for our problem (see Sect. 7.2). To address this problem, we revise the model as follows. We compute σ_i as the standard deviation of behaviour-footprint distances between the i-th pixel location and all locations within a given radius r. Similarly, the spatial scaling factor σ_x is computed as the mean of the spatial distances between all locations within radius r and the centre.[2] The affinity matrix is then normalised according to:

$$\bar{\mathbf{A}} = \mathbf{L}^{-\frac{1}{2}}\mathbf{A}\mathbf{L}^{-\frac{1}{2}} \tag{5}$$

where $\mathbf{L}$ is a diagonal matrix with:

$$\mathbf{L}(i,i) = \sum_{j=1}^{N}(\mathbf{A}(i,j)). \tag{6}$$

$\bar{\mathbf{A}}$ is then used as the input to the Zelnik-Perona's algorithm which automatically determines the number of clusters and performs segmentation. This procedure groups pixel locations into Q optimal regions by their behaviour-footprints for a given scene. Note that because we perform scene segmentation by behaviours, those locations without or with few objects detected (via background subtraction) are absent from the segmentation process, i.e. the N pixel locations to be clustered using the above algorithm do not include those where no or few foreground objects have been detected.

[1] This hard cut-off constraint is introduced to further prevent pixels that are far apart from being grouped together. It is also for reducing the computational cost of computing the affinity matrix $\mathbf{A}$. Without this constraint, the affinity between each pair of behaviour-footprints must be evaluated exhaustively which can be very expensive. For example, with a moderate-sized behaviour-footprint image of 200 by 200, the number of affinity values to compute without the constraint is 1,600,000,000. This number is reduced to around 12,000,000 with the constraint when r is set to 10.

[2] Note that both σ_i and σ_x aim to measure the distribution of distances within a neighbourhood. σ_i measures how similar the behaviour-footprints are within the neighbourhood. Since the similarity/distance values can vary greatly, using standard deviation is more robust against outliers than using mean. σ_x, on the other hand, measures the distribution of image coordinate distances of pixel locations within the neighbourhood. This distribution is constant across neighbourhood and using mean is sufficient.

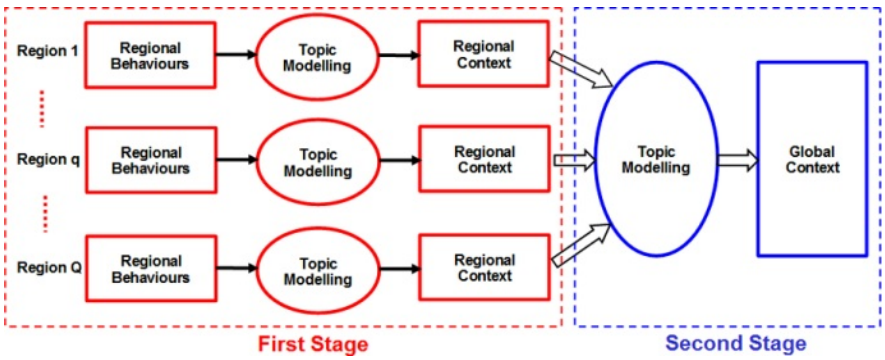


Fig. 3 Structure of the cascade topic modelling

5 Learning Behaviour Correlation and Temporal Context

We aim to learn behaviour correlation and temporal context which constrains how behaviour of an object is correlated to and thus affected by those of other objects both nearby in the same semantic region and further away in other regions of a scene. Given the learned spatial behavioural context (Sect. 4), we consider both behaviour correlation and temporal context at two scales: regional context and global context. In order to discover the context at both scales, a two-stage cascaded topic model is formulated in this section. Figure 3 shows the structure of the proposed model. The inputs to the first stage of the model are regional behaviours represented as labels of regional events (Sect. 5.1). The inferred topics from the first stage modelling correspond to regional behaviour correlation context and are utilised for computing regional (local) temporal context, corresponding to temporal phases of behaviours in different regions. The learned regional temporal context is further used as input to a second stage topic modelling. We define global behaviour correlation context as inferred topics from this model, which are also used to compute global temporal context. More details are as follows.

5.1 Regional Behaviour Representation

Let us first describe how the input to the proposed model, the regional behaviours, are represented. 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 scene events (see (2)). These settings are not necessarily optimal for accurately describing behaviours once the scene has been decomposed semantically into regions. In particular since most behaviour patterns are likely to be similar within each region but differ more across regions, it is sensible to select different features for event representation in different regions. We call them regional events (as compared to scene events). This needs to be determined automatically and to be scalable. To this end, we follow the same procedure as described in Sect. 3 but perform an additional refinement step on event grouping in each region. Specifically, given a decomposed scene, we determine the most representative features in each region by computing entropy values for the 10 features (see (1)) in that region and select the top 5 features (i.e. half) with the highest entropy values. This results in a smaller and more selective set of different features representing events tuned to different regions. After feature selection, these regional events are represented by clustering using GMM within each region. We yield V^q regional event classes in each region q, where $1 \leq q \leq Q$. Note that each region may have different number of event classes and those numbers are determined by automatic model order selection using BIC as in Sect. 3. The regional event class labels are inputs to the cascaded topic model described below.

5.2 Multi-Scale Context Learning

For learning and modelling multi-scale context, we explore a topic model based on the concept of Latent Dirichlet Allocation (Blei et al. 2003). LDA has been widely used for text document analysis aiming to discover semantic topics from text documents according to concurrent correlation of words. In LDA, a document $\mathbf{w}$ is a collection of N_w words: $\mathbf{w} = \{w_1, \ldots, w_n, \ldots, w_{N_w}\}$ and can be modelled as a mixture of K topics $\mathbf{z} = \{z_1, \ldots, z_K\}$. Each topic is modelled as a multinomial distribution over a vocabulary consisting of V words, from which all words in $\mathbf{w}$ are sampled. Given the vocabulary, a document is represented as a V-dimensional feature vector, each element of which corresponds to the count of how many times a specific word occurs in the document. LDA is essentially a Bag of Words model that clusters co-occurring words into topics. It provides a more concise (and arguably more semantic) representation of a document than using all the words directly. The number of topics K is in general much smaller than the size of the codebook/vocabulary V. In order to learn behavioural context both locally within each semantic region and globally across different regions, we formulate a Cascaded Latent Dirichlet Allocation (Cas-LDA) model with two stages.

Stage I—Learning Regional Context For modelling regional behaviour correlation and temporal context, we consider that regional events correspond to words, all events detected from a single region over a clip form a document, correlations of regional events (regional correlation context) correspond to topics, and the inferred topic profile for each document is used for categorising each document into temporal phases (regional temporal context).

More specifically, a training video (or a set of videos) of a scene is segmented temporally into equal-length and non-overlapping short video clips. These short video clips are treated as documents for training a cascaded topic model, given Q semantic regions of a spatially segmented scene (Sect. 4). In the first stage of our cascaded topic model learning, each region is modelled using a LDA. Regional events detected in the q-th region are visual words that form a document denoted as d_t^q; the documents corresponding to all T clips in the q-th region form the regional corpus $\mathcal{D}^q = \{d_t^q\}$, where $t = 1, \ldots, T$ is the clip index and subsequently omitted for conciseness. Assuming that there are V^q classes of regional events in the q-th region, the size of the codebook/vocabulary is thus V^q. Each document is modelled as a mixture of K^q topics, which correspond to K^q different types of regional behaviour correlations and represent our regional correlation context. Note, different from the conventional LDA formulation where a document is represented as the counts of different visual words, our document is represented as a binary V^q dimensional feature vector with each element being a binary value indicating whether a certain regional event class is present in that clip. This is because (1) we are interested in how different behaviours correlate by co-occurrence in each document/clip, rather than how often they occur. (2) This binary vector representation is more robust to noise/error from event detection. Our extensive experiments demonstrate that this modified document representation is more advantageous than the standard representation (see Sect. 7.6).

A LDA model for the q-th region has the following parameters:

1. α^q: a K^q dimension vector governing the Dirichlet distributions of topics in the corpus, i.e. all clips for the q-th region in the training video;
2. β^q: a $K^q \times V^q$ dimension matrix representing the multinomial distributions of words in the vocabulary for all learned topics where $\beta_{k,n}^q = P(w_n^q | z_k^q)$ and $\sum_{n=1}^{V^q} \beta_{k,n}^q = 1$.

Given these model parameters, visual words in a local document can be repeatedly sampled to generate a document of N_w^q words, $d^q = \{w_n^q\}$, as follows:

1. Sample a K^q dimensional vector θ^q from the Dirichlet distribution governed by parameter α^q: $\theta^q \sim \text{Dir}(\alpha^q)$. Vector θ^q contains the information about how the K^q topics are to be mixed in the document.
2. Sample words from topics:
   (a) Choose a topic for w_n^q: $z_n^q \sim \text{Multinomial}(\theta^q)$.
   (b) Choose a word w_n^q from the vocabulary of V^q words according to $P(w_n^q | z_n^q, \beta^q)$.

Following the conditional dependency of the components in the generative process, we can compute the log-likelihood of a document d^q given the model parameters as:

$$\begin{aligned} &\log p(d^q | \alpha^q, \beta^q) \\ &= \log \int p(\theta^q | \alpha^q) \\ &\quad \times \left(\prod_{n=1}^{N_w^q} \sum_{z_n^q} p(z_n^q | \theta^q) p(w_n^q | z_n^q, \beta^q) \right) d\theta^q, \end{aligned} \tag{7}$$

and the log-likelihood of a whole corpus $\mathcal{D}^q = \{d_t^q\}$ of T clips for the q-th region as:

$$\log p(\mathcal{D}^q) = \sum_{t=1}^{T} \log p(d_t^q | \alpha^q, \beta^q), \tag{8}$$

where t is the clip index and T is the total number of documents in the corpus (i.e. all the clips in the training video).

The model parameters α^q and β^q are estimated by maximising the log-likelihood function $\log p(\mathcal{D}^q)$ in (8). However, there is no close-form analytical solution to the problem. A variational EM algorithm can be employed (Blei et al. 2003).

In the E-step of variational inference, the posterior distribution of the hidden variables $p(\theta^q, \{z_n^q\} | d^q, \alpha^q, \beta^q)$ in a specific document d^q is approximated by a variational distribution $p(\theta^q, \{z_n^q\} | \gamma^q, \phi^q)$ where γ^q and ϕ^q are document-specific variational parameters (note that the clip index t is omitted here). As shown by Blei et al. (2003), maximising the log-likelihood $\log p(d^q | \alpha^q, \beta^q)$ corresponds to minimising the Kullback-Leibler (KL) divergence between $q(\theta^q, \{z_n^q\} | \gamma^q, \phi^q)$ and $p(\theta^q, \{z_n^q\} | d^q, \alpha^q, \beta^q)$, resulting in $\log p(d^q | \alpha^q, \beta^q)$ being approximated by its maximised lower bound $L(\gamma^q, \phi^q; \alpha^q, \beta^q)$. By setting α^q and β^q as constants, the variational parameters for d_t^q are estimated according to the following pair of updating equations:

$$\phi_{n,k}^q \propto \beta_{k,v}^q \exp\left(\Psi(\gamma_k^q) - \Psi\left(\sum_{k=1}^{K^q} \gamma_k^q \right) \right), \tag{9}$$

$$\gamma_k^q = \alpha_k^q + \sum_{n=1}^{N_w^q} \phi_{n,k}^q, \tag{10}$$

where $n = 1, \ldots, N_w^q$ indicates the n-th word in d^q; $k = 1, \ldots, K^q$ indicates the k-th regional topic; $v = 1, \ldots, V^q$ indicates the v-th word in the regional vocabulary; and Ψ is the first order derivative of a $\log \Gamma$ function.

In the M-step, the learned variational parameters $\{\gamma^q\}$ and $\{\phi^q\}$ are set as constant and the model parameters α^q and β^q are learned by maximising the lower bound of the log-likelihood of the whole corpus:

$$L(\{\gamma_t^q\}, \{\phi_t^q\}; \alpha^q, \beta^q) = \sum_{t=1}^{T} L(\gamma_t^q, \phi_t^q; \alpha^q, \beta^q). \tag{11}$$

The learned model parameter β^q specifies the probability of each words/regional events given each topic. It thus captures how different types of events are correlated within each region and represents the regional behaviour correlation context.

To compute regional temporal context which corresponds to the temporal phase of behaviours occurring in each region, we perform document clustering (i.e. clip clustering) in the training video as follows. The Dirichlet parameter γ_t^q represents a document in the topic simplex and thus can be viewed as the topic profile in the q-th region and t-th clip, i.e. how likely different topics are combined. Given that the set of topic profiles $\{\gamma_t^q\}$ for a regional corpus $\mathcal{D}^q$ consist of T documents/clips in the q-th region, documents can therefore be grouped into C^q categories $\mathbf{h}^q = \{h_c^q\}$, where $c = 1, \ldots, C^q$. Such clustering can be readily performed by K-means. Consequently, regional behaviours occurring within each document (clip) are uniquely assigned a temporal phase using the class label of that document.

Stage II—Learning Global Context A single second-stage LDA, termed as global context LDA, is employed for learning global behaviour correlation and temporal context. For this LDA, a document is a collection of words corresponding to temporal phases of different semantic regions in the scene. Given a total number of $C = \sum_{q=1}^{Q} C^q$ regional temporal phases classified from the first-stage LDA in all Q regions in a scene, the vocabulary of words for generating documents in the second-stage LDA can then be represented as:

$$\mathcal{H} = [h_1^1, \ldots, h_{C^1}^1, \ldots, h_1^q, \ldots, h_{C^q}^q, \ldots, h_1^Q, \ldots, h_{C^Q}^Q], \tag{12}$$

and each element of $\mathcal{H}$ corresponds to the index of a regional temporal phase. A document $d = \{w_n\}$, where $n = 1, \ldots, N_w$, represents the occurrences of temporal phases in different scene regions at the same time. Thus any word w_n is sampled from $\mathcal{H}$ in (12).

The learning and inference processes of this global context LDA are identical to those for regional context LDA. Suppose K types of global correlation context are discovered from the corpus $\mathcal{D} = \{d_t\}$, where $t = 1, \ldots, T$, the learned parameter β is then a $K \times C$ matrix representing the probabilities of occurrence of each of the C categories of regional temporal phases in K topics of global correlations, corresponding to global behavioural context. Meanwhile, the second-stage LDA also infers the topic profiles $\{\gamma_t\}$, i.e. how global topics are mixed in each of the documents d. The topic profiles can then be employed to classify documents (clips) into a number of temporal phases corresponding to global behaviour temporal context.

After training the cascaded topic model, the model can be applied to interpret behaviours captured in unseen videos. In particular, the model can be employed to infer a topic profile using either the regional LDA or the global LDA. The former reveals what types of regional behaviour correlation exist. The latter informs existence of any global behaviour correlations. Moreover, the inferred regional and global temporal phases assigned to clips can be used for temporal segmentation of unseen videos by topics.

It is worth pointing out that although LDA is explored for topic modelling in our formulation of a cascaded topic model, any other alternative topic model can also be used in our model. For instance, a widely adopted alternative topic model is Probabilistic Latent Semantic Analysis (pLSA) (Hofmann 1999a, 1999b). In our experiments, the effectiveness of LDA and pLSA for context learning are compared (Sect. 7). It should also be noted that in our implementation of LDA, although the model input is a binary vector rather than a vector of counts of occurrences of different words, we do not change the way words are sampled from topics. In other words, it is not enforced that only binary documents can be sampled from the learned model, even though the model is learned with only binary documents. It is found empirically that the change of model input alone is sufficient for improving the model robustness against input noise (see Sect. 7.6). Nevertheless, one could enforce a stricter binary word sampling process by modifying how words are sampled from the model, that is, each word is sampled depending on not only the topic-word distribution, but also the previously sampled words in the same document.

6 Context-Aware Abnormal Behaviour Detection

Given the learned behavioural context, behavioural anomalies are detected as *contextually incoherent* behaviours that cannot be explained away or predicted using the learned behaviour context. Importantly, we also formulate in this section a novel method for not only detecting an anomaly in a video document (clip), but also locating which semantic region and what regional events have caused the anomaly. In a wide area scene featured with multiple objects appearing simultaneously and constantly (e.g. a public road junction or a train platform), it is critical to know both when and where an anomaly occurs. Existing topic models by nature sacrifice specificity of location information about topics (and words)

within each document, therefore is incapable of performing such a task.

First, abnormal clips are detected. Given a test video consisting of non-overlapping clips as documents, we use the global context LDA to examine whether each clip (document) contains abnormal behaviours. This is achieved by computing the normalised log-likelihood of a clip given the trained LDA model. Specifically, the log-likelihood of observing an unseen clip d^* is approximated by its maximised lower bound:

$$\log p(d^*) \approx L(\gamma^*, \phi^*; \alpha, \beta) \tag{13}$$

in which γ^* and ϕ^* are Dirichlet parameters representing d^* in topic simplex and the posterior probabilities of topics of words occurring in d^*. α and β are learned model parameters from the second stage of LDA modelling. $L(\gamma^*, \phi^*; \alpha, \beta)$ is computed by setting α and β as constant and updating γ^* and ϕ^* until $L(\gamma^*, \phi^*; \alpha, \beta)$ is maximised, using (9) and (10). An anomaly score for the clip is then computed as:

$$S(d^*) = \frac{L(\gamma^*, \phi^*; \alpha, \beta)}{N_w^*}, \tag{14}$$

where N_w^* is the number of words in d^*, which in our case corresponds to the number of regional temporal phases identified in the clip d^*. Any clip with anomaly score lower than a threshold TH_c is then detected as containing abnormal behaviours. The value of TH_c is set according to application requirements on the acceptable balance between detection rate and false alarm rate.

Second, once a clip d^* is detected as being abnormal, the regions and subsequently the regional events that triggered the anomaly are identified. This is achieved through a top-down manner in the cascade structure of our model. More specifically, the anomalous region is firstly detected using the global context LDA. The regional events within that region are then examined using the corresponding regional LDA for locating the contributing events. More details are as follows.

Recall that a clip/document $d^* = \{w_n^*\}$, $n = 1, \ldots, N_w^*$, is represented using the inferred regional temporal phases in $\mathcal{H}$ (see (12)). To localise regions containing abnormal behaviours, regional temporal phases are evaluated against the learned global behaviour correlation context. This is carried out by computing $\log p(w_n^*|d^*, \alpha, \beta)$. In this work, we approximate $\log p(w_n^*|d^*, \alpha, \beta)$ by computing $\log p(w_n^*|d_{-n}^*, \alpha, \beta)$, where d_{-n}^* represents a document in which the word w_n^* in d^* is removed. $\log p(w_n^*|d_{-n}^*, \alpha, \beta)$ is the log-likelihood of w_n^* being co-occurring with all other distinct words in d^*, and can be obtained by computing the difference between log-likelihoods of the original document d^* and the document d_{-n}^* as:

$$\begin{aligned}&\log p(w_n^*|d_{-n}^*, \alpha, \beta)\\ &\quad = \log p(w_n^*, d_{-n}^*|\alpha, \beta) - \log p(d_{-n}^*|\alpha, \beta)\\ &\quad = \log p(d^*|\alpha, \beta) - \log p(d_{-n}^*|\alpha, \beta).\end{aligned} \tag{15}$$

Because d_{-n}^* is derived from d^* by removing a single word, it is reasonable to assume that they have the same topic profile γ^*. This implies that for computing the log-likelihood for d_{-n}^*, we can use γ^* learned from d^*. To compute the log-likelihoods of the original document d^* and the document d_{-n}^* (i.e. $\log p(d^*|\alpha, \beta)$ and $\log p(d_{-n}^*|\alpha, \beta)$ in (15)), the following two steps are taken:

- Step 1: Given α and β, compute $L(\gamma^*, \phi^*; \alpha, \beta))$ to maximise the lower bound of $\log p(d^*|\alpha, \beta)$ through iteratively updating (9) and (10), and storing the value γ^*;
- Step 2: Given the document d_{-n}^*, set γ^* in (10) as constant and only update ϕ in (9) resulting the maximised lower bound of $\log p(d_{-n}^*|\alpha, \beta)$ being computed as $L(\gamma^*, \phi_{-n}^*; \alpha, \beta)$ where each element in ϕ_{-n}^* represents how likely that assigning each global topic to a word in d_{-n}^*.

$\log p(w_n^*|d_{-n}^*, \alpha, \beta)$ can finally be computed as:

$$\log p(w_n^*|d_{-n}^*, \alpha, \beta) \approx L(\gamma^*, \phi^*; \alpha, \beta)) - L(\gamma^*, \phi_{-n}^*; \alpha, \beta). \tag{16}$$

The regions with the lowest values of $\log p(w_n^*|d_{-n}^*, \alpha, \beta)$ are considered to contain abnormal behaviours. In those regions, the above procedure can be further deployed to localise any abnormal regional events by computing:

$$\begin{aligned}\log p(w_n^{q*}|d_{-n}^{q*}, \alpha^q, \beta^q) \approx\ & L(\gamma^{q*}, \phi^{q*}; \alpha^q, \beta^q)\\ & - L(\gamma^{q*}, \phi_{-n}^{q*}; \alpha^q, \beta^q)\end{aligned} \tag{17}$$

for all regional events w_n^{q*} using the regional LDAs. The abnormal regional events are then detected as those giving lowest log-likelihoods of co-occurring with all other events occurring in the same region of the same clip.

7 Experiments

7.1 Datasets

Three datasets were employed in our experiments for evaluating the effectiveness of the proposed framework for learning behavioural context.[3] All datasets were collected from real-world surveillance scenes featured with large numbers of objects exhibiting complex behaviours.

[3]The Junction and Roundabout datasets can be downloaded at http://www.eecs.qmul.ac.uk/~jianli/Dataset_List.html, where the anomaly annotation and evaluation protocol can also be found.

(a) Vertical traffic (b) Horizontal traffic

Fig. 4 Example frames from the Junction dataset

(a) Vertical traffic (b) Horizontal traffic

Fig. 5 Example frames from the Roundabout dataset

Junction Dataset This dataset contains videos of a busy urban road junction recorded at 25 Hz with a frame size of 360×288 pixels. The dataset consists of 34000 frames, among which 22000 frames were used for training and the rest 12000 frames were used for testing. Example frames are shown in Fig. 4. This scene contains different types of objects including vehicles and pedestrians moving at different regions of the scene. The behaviour of an object is governed by both the traffic lights and the behaviours of other objects co-existing in the scene. Specifically, as illustrated in Fig. 4, there are two traffic phases: the vertical traffic phase and the horizontal traffic phase. During the horizontal phase, there are also two sub-phases corresponding to the leftward and rightward horizontal traffic respectively. To make things even more complicated, the vehicles waiting in the central waiting zone for horizontal turning (see Fig. 4(a)) can do so whenever there is a gap in the vertical flow.

Roundabout Dataset This dataset contains videos of a traffic roundabout recorded at 25 Hz with a frame size of 360×288 pixels. The training and test sets contain 45000 frames and 18000 frames respectively. Example frames are shown in Fig. 5. Vehicles in this scene usually enter the scene from the entrances near the left boundary and right bottom corner. They move towards the exits located on the top, at left bottom corner and near the right boundary. Similar to the junction scene, the roundabout traffic was controlled by multiple sets of traffic lights and can be roughly divided into two phases temporally. However, compared to a junction, behaviours of vehicles in a roundabout are less regulated by traffic lights. Consequently, the two traffic phases are less distinctive visually (e.g. vehicles can leave the scene using the exits in the top during both traffic phases).

Fig. 6 Example frames from the MIT Traffic dataset

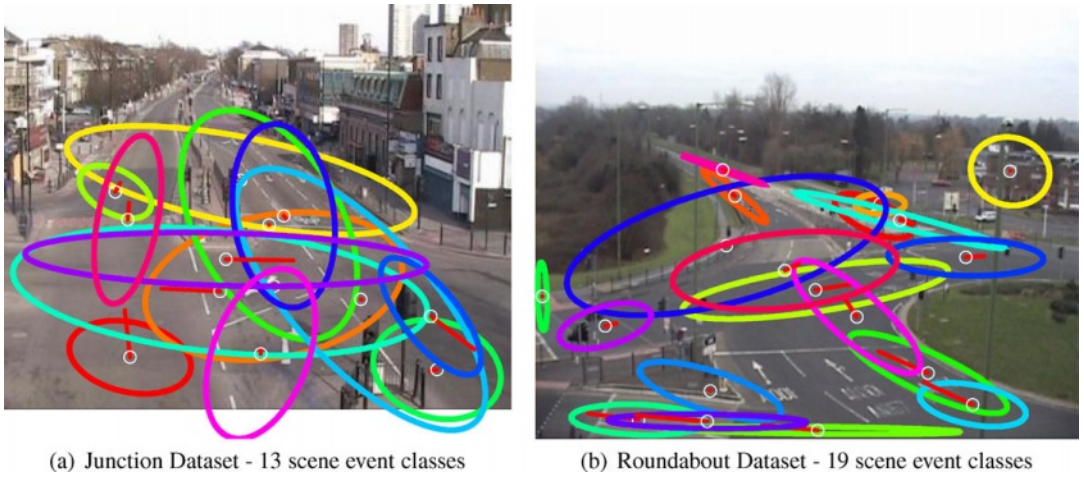
MIT Traffic Dataset (Wang et al. 2009) This dataset contains 1.5 hours of video with an image size of 720 × 480 pixels and frame rate of 30 Hz. Figure 6 shows example frames. The video was cut into 540 non-overlap clips of 10 second long each. This dataset was used to compare the performance of our model to that of Wang et al. (2009) (see Sect. 7.6).

7.2 Learning Behaviour Spatial Context

Scene Event Detection In the Junction Dataset, 121583 clip-wise scene events were detected. After being smoothed within each clip (see Sect. 3), they were automatically grouped into 13 clusters, each of which represents one scene event class. In the Roundabout Dataset, 440607 scene events were detected which led to 19 scene event classes. The detected scene event classes in both datasets are illustrated in Fig. 7. It can be seen that different scene events correspond semantically to objects at different regions of the scene performing different behaviours (e.g. moving towards certain directions at certain speeds).

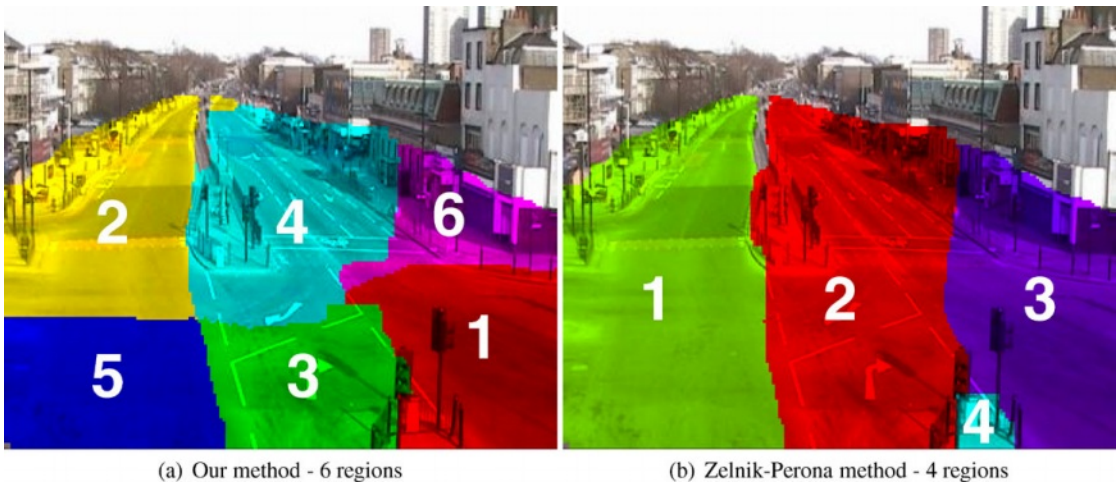
Semantic Scene Segmentation Following the procedure described in Sect. 4, each scene was decomposed into semantic regions for learning behaviour spatial context. Specifically, the behaviour-footprint images for all three scenes were resized to a similar size.[4] The radius in (4) for computing scaling factors of the affinity matrices was set to $r = 10$. Our spectral clustering algorithm automatically decomposed the Junction Scene into 6 regions and the Roundabout Scene into 9 regions. Figures 8(a) and 9(a) show that semantically meaningful regions are obtained using our method. In particular, the decomposed regions corresponds to various traffic lanes and waiting zones where object behaviours tend to follow a similar pattern. Our proposed spectral clustering algorithm was compared with the original Zelnik-Perona (ZP) method. The results in Figs. 8 and 9

[4] After resizing, the sizes of the Junction and Roundabout scenes are 180×144 pixels and the size for the MIT scene is 180×120 pixels.



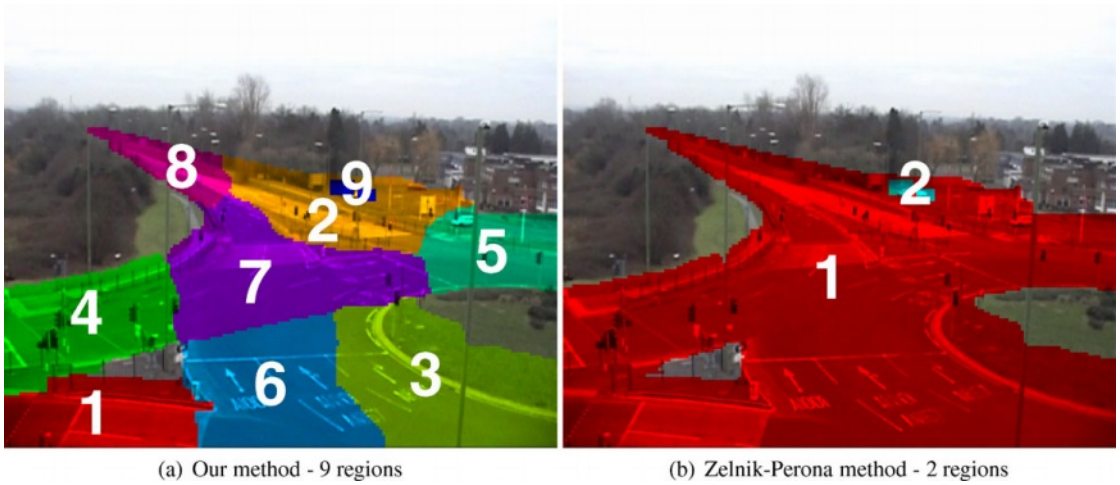
(a) Junction Dataset - 13 scene event classes (b) Roundabout Dataset - 19 scene event classes

Fig. 7 (Color online) The detected scene event classes. Each class is represented using an ellipse. The centre of each ellipse correspond to the mean positions of all scene events belonging to that class. The orientation of an ellipse is determined by the angle between the major and minor axes of spatial distributions of events. The mean speed and orientation for the events in each class is depicted using a *red arrow* originating from the centre of the ellipse



(a) Our method - 6 regions (b) Zelnik-Perona method - 4 regions

Fig. 8 Semantic scene segmentation for the Junction Dataset



(a) Our method - 9 regions (b) Zelnik-Perona method - 2 regions

Fig. 9 Semantic scene segmentation for the Roundabout Dataset

indicate clearly that for both scenes, the original Zelnik-Perona (ZP) method suffers from under-fitting severely and was not able to decompose both scenes correctly according to the expected traffic behaviour patterns. These results suggest that setting the scaling factors (see (4)) properly is crucial for meaningful scene segmentation using spectral clustering.

7.3 Learning Behaviour Correlation Context

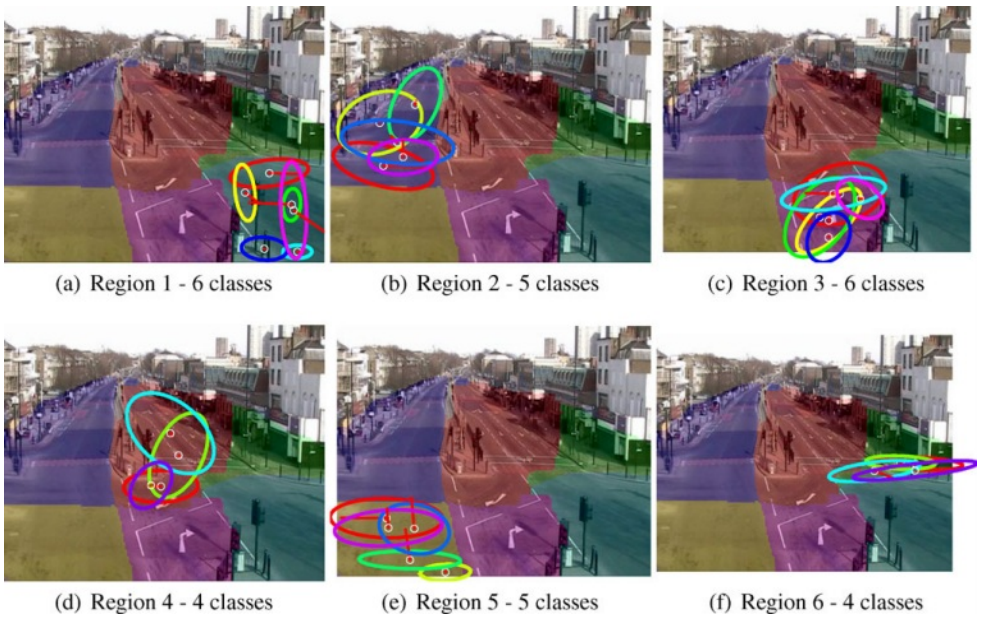
Regional Event Detection With the learned behaviour spatial context, behaviours occurring within each semantic region are represented using regional events (Sect. 5.1). Automatically selected features for each region of both scenes are shown in Tables 1 and 2. It is evident from the tables that different features were selected for different regions. For example, Table 2 shows that motion features were selected as informative features for Region 7 but not for Region 4. This reflects the fact that in the Roundabout Dataset (see Fig. 9), Region 4 contains predominantly static or slow moving objects, whilst Region 7 contains mostly objects moving rapidly towards different directions. Using the selected features, 30 classes of regional events were detected automatically for the Junction Dataset and 52 for the Roundabout Dataset. These regional event classes for the two datasets are shown in Figs. 10 and 11 respectively. Compared with the detected scene events (see Fig. 7), these regional events detected using the learned behaviour spatial context are more fine-grained and also more meaningful. For example, in Region 1 of the Junction scene, two classes of regional events were detected at the bottom of this region (see the blue and cyan ellipses in Fig. 10(a)). They correspond to (1) pedestrian waiting in the crossroad island and (2) pedestrian crossing the road respectively. These two classes of events are important for understanding the behaviours of objects in the region (e.g. when pedestrians are crossing, vehicles are not supposed to be present in the region) and of those in the whole scene (e.g. pedestrians waiting in the crossroad island indicates the vertical traffic flow phase). Both events are missed from the detected scene event classes (see Fig. 7(a)). This is because that behaviour spatial context is not utilised in detecting scene events.

Table 1 Representative features selected for the Junction Dataset

	x	y	w	h	r_s	r_p	u	v	r_u	r_v
Region 1	√	√	√			√	√			
Region 2	√	√	√			√		√		
Region 3	√	√	√	√		√				
Region 4	√	√	√	√		√				
Region 5	√	√		√		√		√		
Region 6	√				√	√	√		√	

Table 2 Representative features selected for the Roundabout Dataset

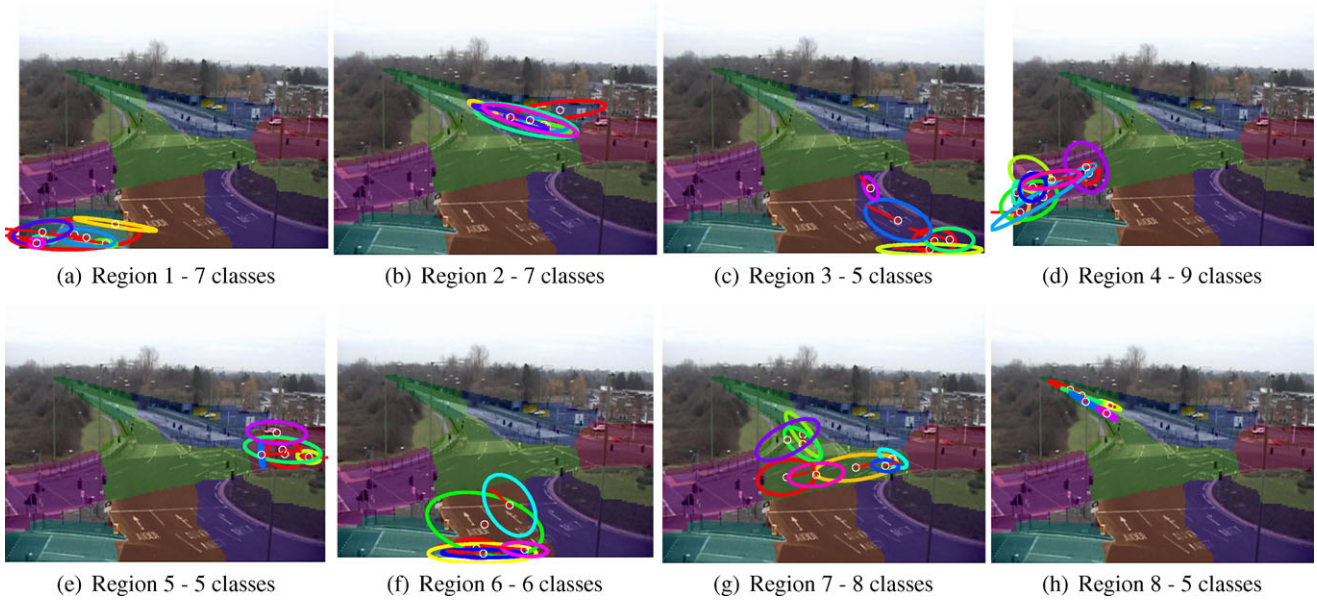
	x	y	w	h	r_s	r_p	u	v	r_u	r_v
Region 1	√	√		√		√	√			
Region 2	√	√	√	√		√				
Region 3	√	√		√		√	√			
Region 4	√	√	√	√		√				
Region 5	√	√		√		√	√			
Region 6	√	√				√	√	√		
Region 7	√	√				√	√	√		
Region 8	√	√				√	√	√		



(a) Region 1 - 6 classes (b) Region 2 - 5 classes (c) Region 3 - 6 classes

(d) Region 4 - 4 classes (e) Region 5 - 5 classes (f) Region 6 - 4 classes

Fig. 10 The detected 30 regional event classes in each region of the Junction Scene



(a) Region 1 - 7 classes (b) Region 2 - 7 classes (c) Region 3 - 5 classes (d) Region 4 - 9 classes

(e) Region 5 - 5 classes (f) Region 6 - 6 classes (g) Region 7 - 8 classes (h) Region 8 - 5 classes

Fig. 11 The detected 52 regional event classes in each region of the Roundabout Scene

Multi-scale Behaviour Correlation Learning A cascaded LDA model (denoted as Cas-LDA) was trained for each dataset for learning multi-scale behaviour correlation context. Specifically, the detected regional events from each region were used for training a first-stage LDA for regional behaviour correlation modelling, with the learned hidden topics representing the local behaviour correlation context. The regional LDAs then provide inputs for a second-stage LDA for modelling global behaviour correlation. The learned topics using this LDA then correspond to global behaviour correlation context.

Let us first look at the results obtained from the Junction Dataset. The training data was first temporally segmented into non-overlapping clips with equal length of 300 frames, resulting in 73 clips/documents for the first stage LDA training. It was found that varying the numbers of topics for the local LDAs has little effect on the global correlation learning in the second stage. In this experiment, we set the numbers of topics for different regions to be equal and search for the optimal number that gives the best global temporal phase inference performance using cross validation (see Sects. 5.2 and 7.4). Subsequently the number of topics was automatically set to 4 for all 6 regions in the scene. The learned topics for each region, which correspond to different types of regional behaviour correlations, are illustrated in Fig. 12. As can be seen in Fig. 12, each learned topic or regional behaviour correlation context captures one type of commonly observed concurrent object behaviours under one specific traffic phase. For instance, in Region 1, both Topic 1 and 4 correspond to the vertical traffic flow in that region, whilst Topic 2 represent leftward horizontal flow with moving pedestrians and Topic 3 captures the rightward horizontal flow with pedestrians waiting for crossing. For the second stage LDA, the number of topics was set to 2, corresponding to the two global traffic phases. The learned topics are shown in Fig. 13. It is evident that Topic 1 (Fig. 13(a)) represents the concurrent behaviours of various objects in different regions during the vertical traffic phase, whilst Topic 2 (Fig. 13(b)) corresponds to those during the horizontal phase.

For the Roundabout Dataset, there were 146 clips obtained after segmenting the training data into non-overlapping clips. The number of topics for the first stage LDAs was determined as 2 and the learned topics using the first stage and second stage LDAs are illustrated in Figs. 14 and 15 respectively. It can be seen in Fig. 14 that the learned topics using the first stage LDAs reveal clearly the typical concurrent object behaviours in each region. In particular, it is noted that in regions where the traffic phase has a significant influence on the behaviour, the two topics correspond to the two traffic phases. For instance, in Region 6, the two topics capture vehicles waiting during horizontal traffic phase and vehicles moving upwards during vertical phase respectively. In Region 7, the regional LDA reveals how vehicles typically move in different directions during the two traffic phases. In contrast, in Regions 2 and 8, the two topics are alike since these regions are less effected by the traffic phases. Similar to the Junction Dataset, the learned 2 topics using the second stage LDA discovers correlations of object behaviours across different regions during the two traffic phases (see Fig. 15).

Fig. 12 The learned regional topics for each region in the Junction Scene using the Cas-LDA model. Each topic is illustrated using the top two dominant/likely words/regional event classes

Cas-LDA vs. LDA To evaluate the effectiveness of the proposed two-stage cascaded model structure, our Cas-LDA model was compared with a standard LDA using detected scene events as input. The latter has a single stage learning and does not rely on scene segmentation for computing model input. The inferred topics using the single-stage LDA for the two datasets are shown in Figs. 13 and 15 respectively. It is evident that the learned topics using a standard single-stage LDA capture less meaningful correlations of object behaviours compared to our Cas-LDA. It can also be observed from the comparative results that without exploiting behaviour spatial context, the visual words (scene events) alone failed to capture the subtle difference between object behaviours in different regions. Consequently, the correlations learned using the standard LDA are less meaningful and more difficult to interpret, indicating weakened capability for modelling complex behaviour correlations.

Cas-LDA vs. Markov Clustering Topic Model (MCTM) We further compare Cas-LDA with the Markov Clustering Topic Model (MCTM) by Hospedales et al. (2009) which has a single hierarchical model structure as opposed to our cascaded structure, and is a hybrid of PTM and DBN. In Hospedales et al. (2009), the mode input to MCTM are local motion events/words computed from optical flow vectors at each regular grid location. The computational cost of learning a MCTM is extremely high given the complex model structure and sheer number of words computed from each clip. Here, for fair comparison, our implementation of MCTM directly uses detected scene events as input, i.e. the same input as our Cas-LDA. A MCTM groups events/words into actions/topics, and documents into behaviours, which correspond to the temporal phases learned by our temporal context learning model. We set the number of topics to eight and learned two behaviours each of which corresponds to a traffic phase. The learned behaviours using MCTM for the

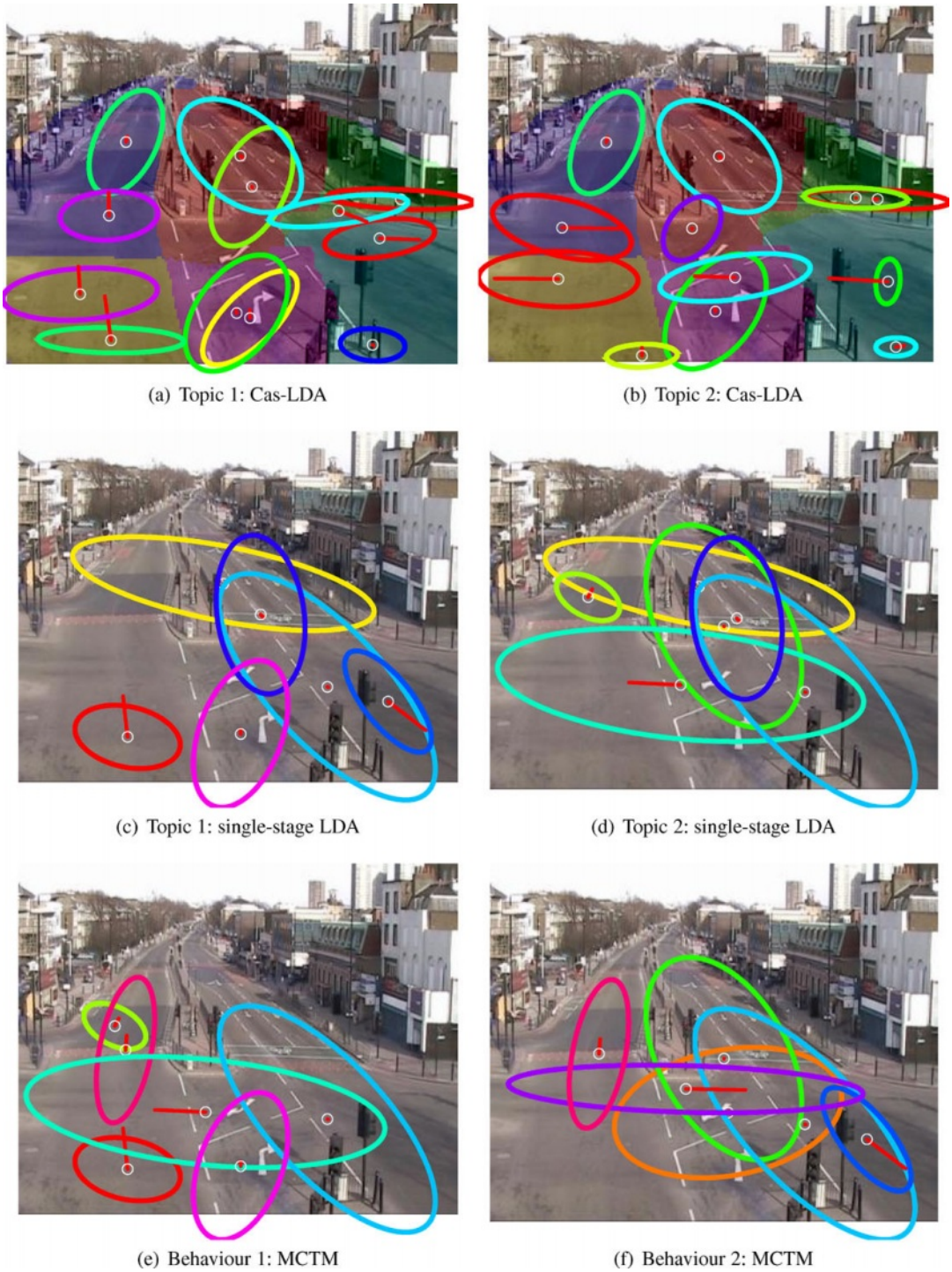


Fig. 13 Learning global behaviour correlation context in the Junction Scene. (**a**)–(**b**): The learned topics using the second-stage LDA of the cascade, which correspond to the object behaviour correlations during the two traffic phase. Each topic is illustrated using the top two most dominant/likely words/regional event classes in each region. (**c**)–(**d**): The topics learned using a single-stage LDA without the cascade structure. The top 6 most dominant/likely scene event classes are shown for each topic. (**e**)–(**f**): The behaviour temporal phases learned using a Markov Clustering Topic Model (MCTM)

Junction scene and Roundabout scene are shown in Figs. 13 and 15, respectively. The results show that with a hierarchical model structure, the MCTM is not able to learn a set of interpretable behaviour correlations for different temporal phases, even though it models the temporal ordering of behaviour temporal phases explicitly.

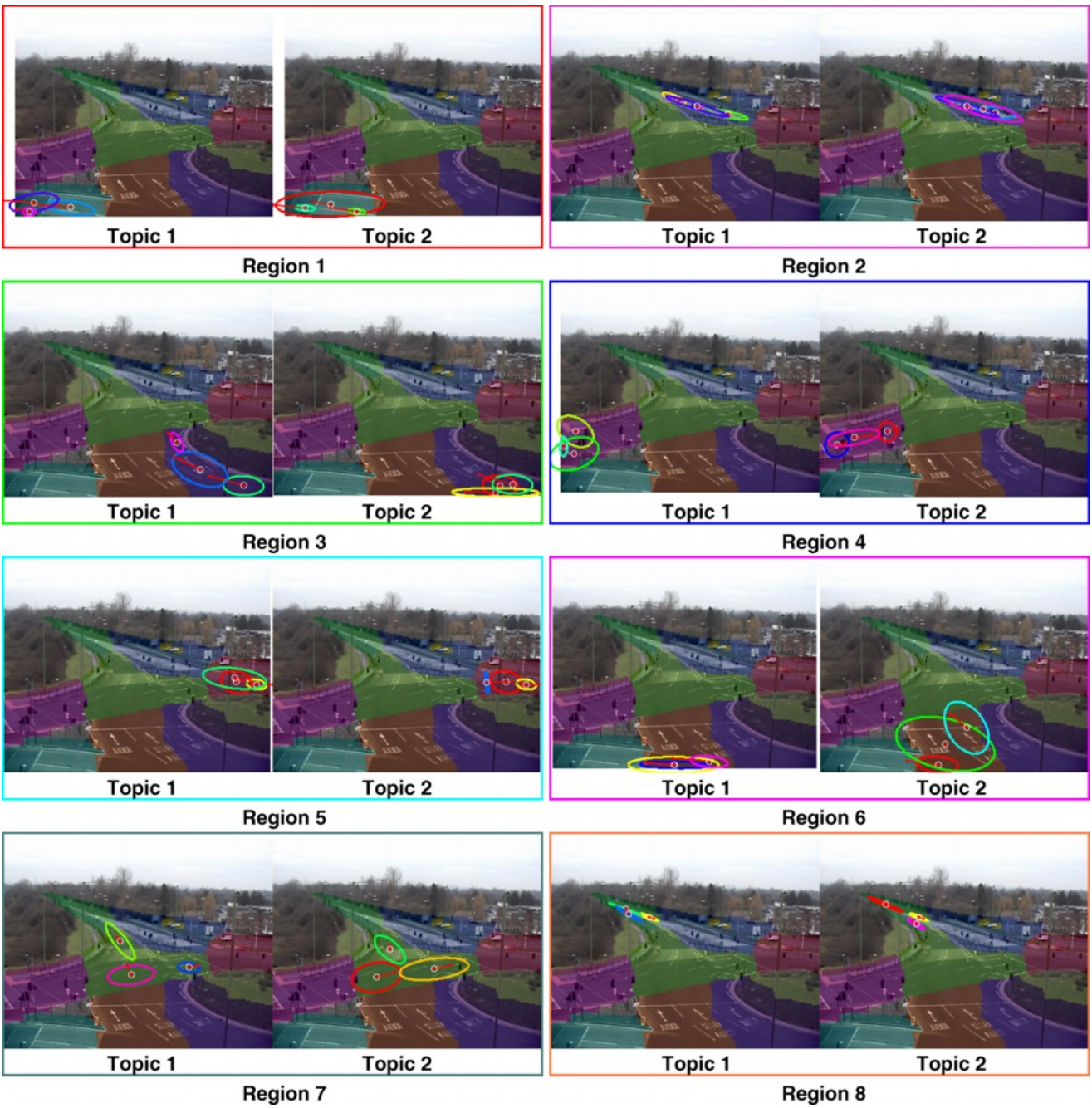


Fig. 14 The learned regional topics for each region in the Roundabout Scene using the Cas-LDA model. Each topic is illustrated using the top three dominant (most likely) words/regional event classes

7.4 Learning Behaviour Temporal Context

Temporal Context Learning for Video Segmentation As described in Sect. 5, given a learned Cas-LDA model, each video document (clip) was represented by the inferred topic profile indicating how likely each topic is present in that clip. The topic profiles were then used for classifying documents (clips) into different types of temporal context. Learning temporal context also provides a mechanism for segmenting an unseen video into different temporal phases. In particular, for both the Junction and Roundabout datasets, there are two types of temporal visual context corresponding the vertical and horizontal traffic flow phases respectively. To evaluate the temporal context learning, we generated ground truth by manually labelling exhaustively all the clips in the testing video frames from both datasets into two traffic phases. Table 3 shows that an accuracy of 87.2% and 74.6% were obtained on the two datasets respectively

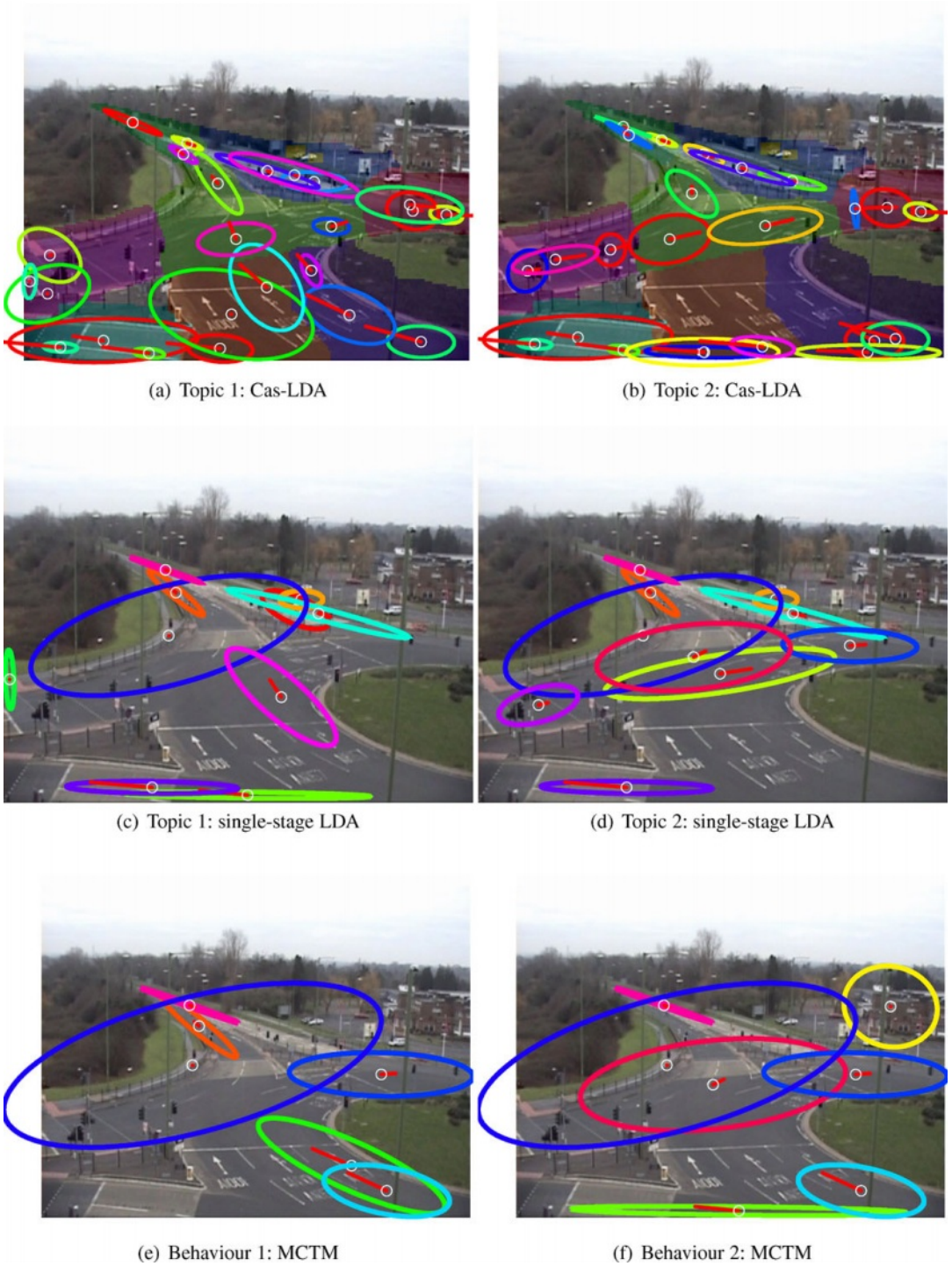


Fig. 15 Learning global behaviour correlation context in the Roundabout Scene. (**a**)–(**b**): The learned topics using the second-stage LDA of the cascade, which correspond to the object behaviour correlations during the two traffic phase. Each topic is illustrated using the top three most dominant/likely words/regional event classes in each region. (**c**)–(**d**): The topics learned using a single-stage LDA without the cascade structure. The top 10 most dominant/likely scene event classes are shown for each topic. (**e**)–(**f**): The behaviour temporal phases learned using a Markov Clustering Topic Model (MCTM)

from applying our model for inferring and predicting traffic phase. These results show that our Cas-LDA model is able to learn the temporal context in a meaningful way. It is noted that lower segmentation accuracy was obtained for the Roundabout Scene due to the more complex and less regulated object behaviours controlled by the traffic lights.

(a) A fire engine interrupting normal traffic flow

(b) A white car jumping a red-light and getting dangerously close to a black car

Fig. 16 Examples of abnormal behaviours

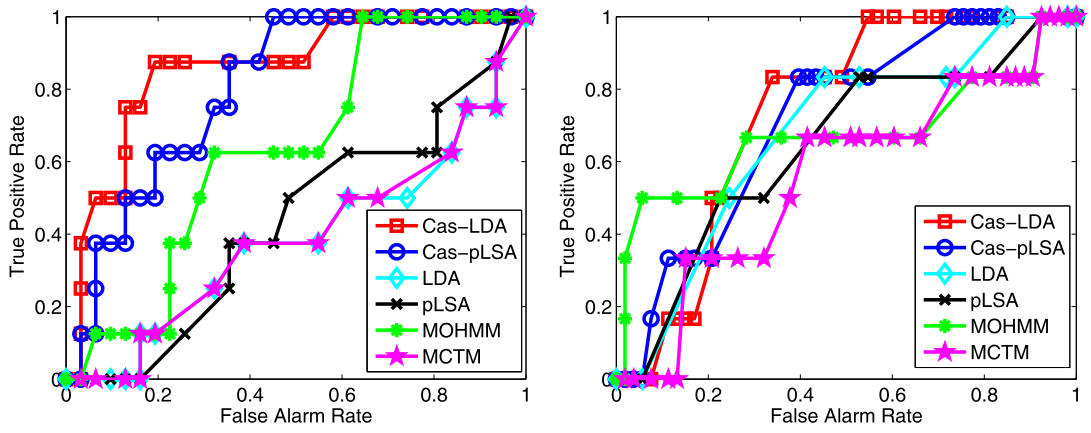
Table 3 Temporal video segmentation accuracy using different models

	MOHMM	MCTM	pLSA	Cas-pLSA	LDA	Cas-LDA
Jun.	56.4%	51.79%	56.4%	**89.7**%	61.5%	87.2%
Rou.	66.1%	68.36%	52.5%	**76.2**%	55.9%	74.6%

Comparing with Alternative Models The proposed Cas-LDA model was compared with a host of alternative topic models for learning behavioural context, including a single-stage LDA (Blei et al. 2003), single-stage pLSA (Hofmann 1999a, 1999b), Cas-pLSA (Li et al. 2008b), a Dynamic Bayesian Network (DBN) in the form of a Multi-Observation Hidden Markov Model (MOHMM) (Xiang and Gong 2008) and a Markov Clustering Topic Model (MCTM) (Hospedales et al. 2009). The single-stage pLSA and Cas-pLSA use the same model input and similar learning and inference algorithms, but differ from the single-stage LDA and Cas-LDA in topic model formulation. For the MOHMM, detected regional events were used as observations and inferred hidden states were used for video segmentation. This means that the inputs to the MOHMM are obtained based on the learned spatial context information. The results in Table 3 show that the cascaded model structure makes a huge difference for both LDA and pLSA with a relative increase of around 50% in performance compared with the corresponding single stage topic models. Table 3 also shows that the MOHMM model generated poor results despite that it has already benefited from the learned spatial context by using regional events as input. MCTM is capable of grouping documents/clips into temporal phases and modelling the temporal ordering of different phases explicitly. It is thus in theory well-suited for learning temporal context. However, our results suggest that the temporal phases learned use a MCTM are less meaningful compared with a Cas-LDA (see Figs. 13 and 15). Consequently, less accurate temporal segmentation of video is achieved. Furthermore, we observed that using the two different topic models in our cascaded structure seems to make little difference with Cas-pLSA achieving slightly better performance than Cas-LDA.

7.5 Context-Aware Anomaly Detection

Anomalies as Contextually Incoherent Behaviours To evaluate the proposed approach for abnormal behaviour detection (Sect. 6), every clip in the testing datasets is labelled as either normal or abnormal by careful manual examination. This led to 8 out of 39 clips and 6 out of 59 clips being identified as being abnormal for the Junction and Roundabout Datasets respectively. The anomalies in the testing part of the Junction Dataset fall into two categories: emergency vehicles such as fire engine and ambulance interrupting normal traffic flow, or dangerous leftward or rightward turning during the vertical traffic flow phase. For the Roundabout Dataset, all abnormal clips contain vehicles jumping a red-light. It is important to observe that in all anomalies, the behaviours of individual objects alone exhibit visually little if any information for being abnormal. They all look normal in isolation. The abnormalities were caused by behaviours taking place in a wrong place at the wrong time resulting in abnormal correlations with other objects in the scene, i.e. being contextually incoherent. Some examples of anomalies are shown in Fig. 16.



(a) Junction Dataset (b) Roundabout Dataset

Fig. 17 (Color online) Comparing anomaly detection performance of different models using ROC curves

Table 4 Comparing different models on anomaly detection using Area under ROC (AUROC) values

	MOHMM	MCTM	pLSA	Cas-pLSA	LDA	Cas-LDA
Jun.	0.6351	0.3911	0.4355	0.8024	0.3871	**0.8589**
Rou.	0.6730	0.5579	0.6431	0.7154	0.6761	**0.7374**

Comparing with Alternative Models We compared the performance of our Cas-LDA model with the single-stage LDA and pLSA, Cas-pLSA, MOHMM and MCTM. Specifically, we varied the threshold TH_c (see Sect. 6) for detection and obtained Receiver Operating Characteristic (ROC) curves, from which the Area Under ROC (AUROC) values were computed. The results are shown in Fig. 17 and Table 4. Similar to the traffic phasing prediction and segmentation results early, both cascaded topic models achieved significantly better performance compared to their single stage counterparts. This again highlights the importance of utilising behavioural context and a cascaded model structure. Our cascaded topic models also outperform the MOHMM based DBN model and MCTM. Again, the performances of Cas-pLSA and Cas-LDA are similar but this time with Cas-LDA yielding better detection result.

Locating Contributing Events in an Anomaly Using the method formulated in Sect. 6 (see (16) and (17)), we were able to locate regional events that caused those clips being detected as abnormal. Figure 18 shows examples when the proposed method identified and located accurately, among dozens of objects present in the scene, those specific object behaviours that were most abnormal. Specifically, the concurrent correlation of these objects are at odds with the expected behaviour correlation and temporal context. For example, this could mean that the objects are moving on a collision course with other objects (see Figs. 18(a) and (d)). This is a very useful feature with which an automated system can be utilised to pinpoint the abnormality both in space and over time in a complex scene involving multiple objects concurrently, aiding rapid human response.

7.6 Further Comparisons with Alternative Approaches

Comparing with Tracking-Based Approach Our approach is based on a discrete event-based representation for behaviour modelling. This is different from most existing approaches which rely on tracking. To highlight the inadequacy of tracking based representation for behaviour modelling in crowded scenes, in Fig. 19(a), we show the trajectories extracted from a two-minute video clip from the Junction Dataset. In Fig. 19(b), we plot a histogram of the durations of all the tracked objects 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 clip (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 representation and subsequent analysis in this type of scenes. Moreover, it is equally important to point out that monitoring objects in isolation even over a prolonged period of time through tracking does not necessarily facilitate temporal segmentation such as traffic phase inference and prediction, and anomaly detection in the context of regulated traffic flow.

Comparing with Hierarchical Topic Models For correlation modelling in a complex scene, Wang et al. (2009) proposed to use Dual Hierarchical Dirichlet Processes (Dual-HDP). In the bottom layer of a Dual-HDP model, concurrent quantised motion information (visual words) are grouped into atomic activities, whilst in the top layer, these activities are grouped into interactions if they co-occur. Compared to

Fig. 18 (Color online) The contributing events are identified and highlighted using *red* bounding boxes. (**a**)–(**b**): events identified for the fire engine anomaly in the Junction Dataset. (**c**)–(**d**): events identified for the running traffic light anomaly in the Roundabout Dataset

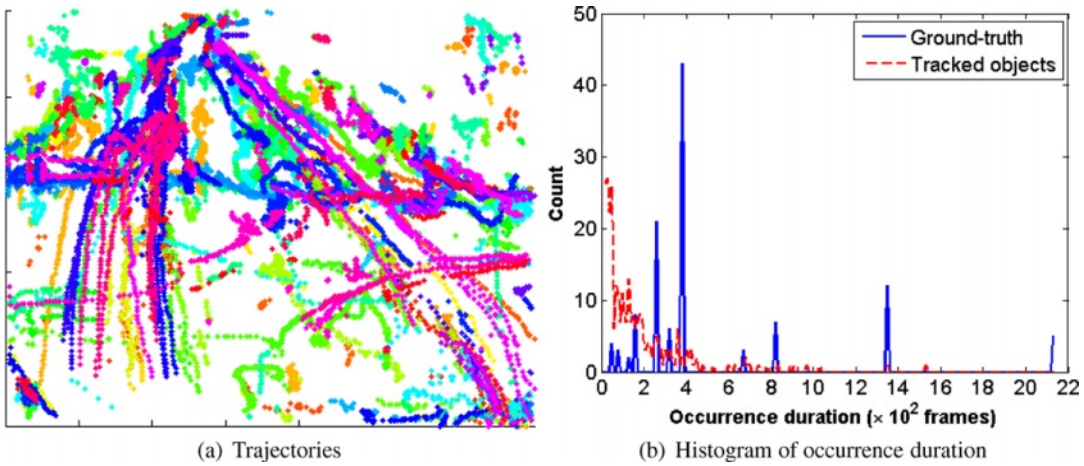


Fig. 19 (Color online) Failure of tracking in the Junction scenario

our cascaded topic model, the discovered interactions using the Dual-HDP correspond to the global correlation and temporal context discovered using the second stage LDA. We carried out experiments to compare the Dual-HDP model with our Cas-LDA model for learning behavioural context of complex scenes. Since the implementation of the Dual-HDP model is non-trivial and unavailable to us, we were not able to evaluate it on the Junction and Roundabout datasets. Instead, we apply our Cas-LDA to the MIT Traffic Dataset used by Wang et al. (2009) (see Fig. 6) and compared results with those reported in their paper. To ensure that the difference in performance is only caused by the model used rather than behaviour representation, the same low-level motion feature based representation as adopted by Wang et al. (2009) was used.

Figure 20 shows the 9 semantic regions discovered by our spatial context learning. Five types of interactions, i.e. temporal phases, learned using the second-stage LDA are illustrated using concurrent traffic flows in Figs. 21(a) and (b). For direct comparison, the discovered global correlations using Dual-HDP is shown in Fig. 21(c). These results were reproduced from Fig. 11(b) in Wang et al. (2009).

It is evident from Fig. 21 that more meaningful behaviour correlations are discovered using our Cas-LDA model. In

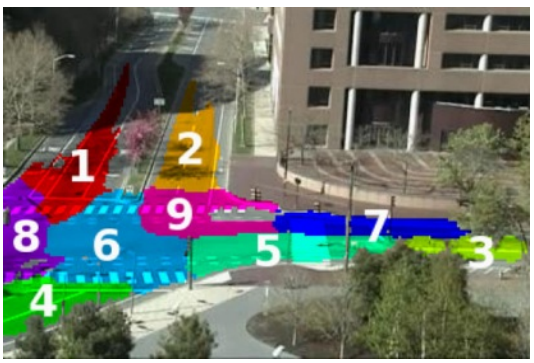
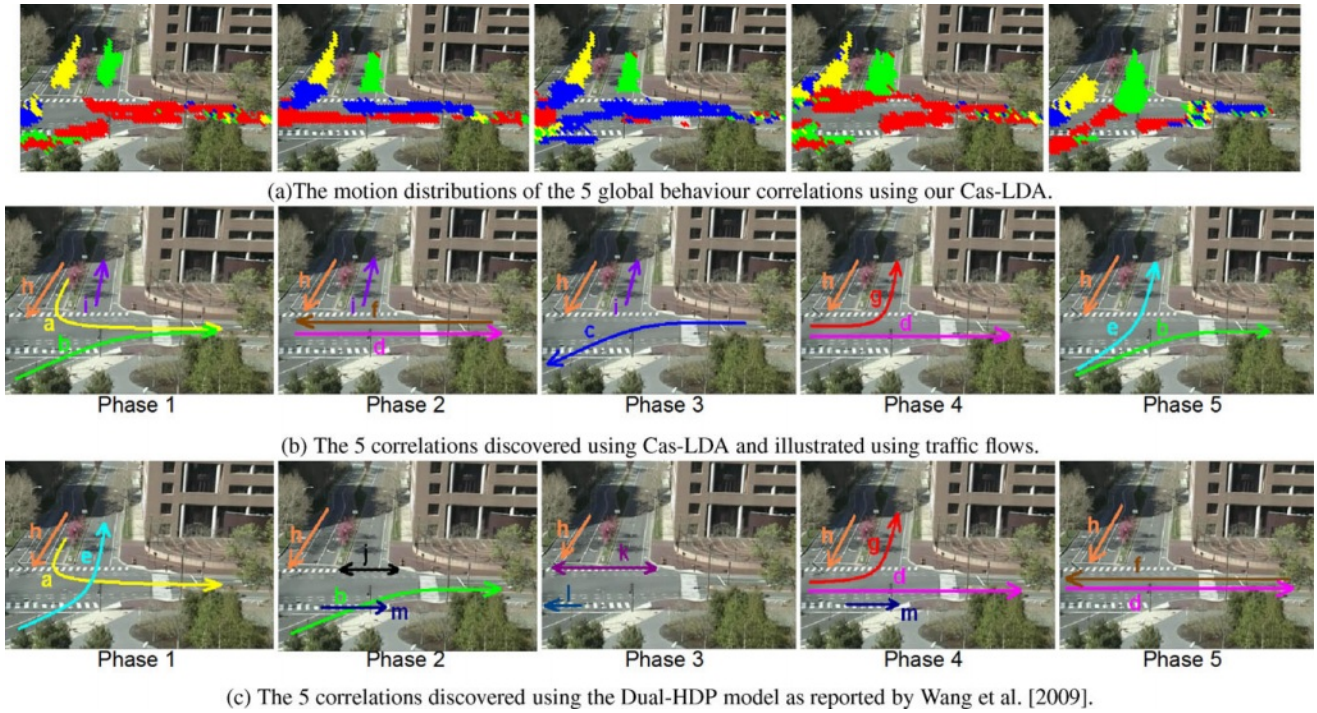


Fig. 20 Semantic scene segmentation for the traffic scene used by Wang et al. (2009)

particular, it can be seen from Figs. 21(b) and (c) that Phase 2 and 4 of Cas-LDA are almost identical to Phase 4 and 5 of Dual-HDP respectively. But the other 3 phases are very different. Specifically, Phase 1 of Dual-HDP suggests that flow *a* and *e* occur simultaneously. In reality, since the flow are on collision course, they can only co-exist in a same clip if there are gaps in either *a* or *e* which is rare in the video used in the experiment. What was taking place much more frequently in the video is the co-occurrence of flow *a* and



(a)The motion distributions of the 5 global behaviour correlations using our Cas-LDA.

(b) The 5 correlations discovered using Cas-LDA and illustrated using traffic flows.

(c) The 5 correlations discovered using the Dual-HDP model as reported by Wang et al. [2009].

Fig. 21 (Color online) Comparing Cas-LDA with Dual-HDP for global correlation modelling. In (**a**), different colours represent different motion directions: *red* →, *green* ↑, *blue* ←, *yellow* ↓. In (**b**) and (**c**) each flow is labelled and has different colours with the arrow indicating the flow direction

Table 5 Accuracy of temporal segmentation using Cas-LDA

Ground truth	Cas-LDA labels				
	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
Phase 1	78	0	1	15	7
Phase 2	1	89	3	2	1
Phase 3	0	0	61	4	0
Phase 4	0	8	2	109	0
Phase 5	12	0	3	11	133

b, and *e* and *b*, which have been captured by Phase 1 and 5 respectively using our Cas-LDA. Phase 2 of Dual-HDP indicates that typically flow *b* co-occurs with pedestrians crossing horizontally (flow *j* and *m*) but not with other traffic. Again this happens very rarely in the video because when flow *b* is taking place, pedestrians cannot cross in the bottom of the image (flow *m*). Phase 3 of Dual-HDP corresponds to pedestrians crossing horizontally (flow *k* and *l*) with not much vehicle traffic. However, it is perfectly normal to have horizontal vehicle traffics co-existing with pedestrian crossing as indicated by Phases 4 and 5 of Dual-HDP. Compared with Dual-HDP, our Cas-LDA discovered an important correlation missed by Dual-HDP (Phase 3 in Fig. 21(b)), that is, horizontal right-to-left traffic performs left-turn (flow *c*) without the co-existence of left-to-right-moving traffic (flow *d*).

The 540 clips in the MIT Traffic dataset were manually labelled exhaustively into 5 temporal phases according to the discovered 5 correlations. The confusion matrix between the segmentation result and the ground truth is shown in Table 5. The average segmentation accuracy is 87.04%. This is similar to the 85.74% result obtained from Dual-HDP reported by Wang et al. (2009), although as explained above the discovered phases have different meanings. On computational cost, Dual-HDP model training takes 12 hours to process the 1.5 hour video data, as reported by Wang et al. (2009). In comparison, our Cas-LDA is computationally much more efficient, requiring only 25 minutes on a 2.5 GHz PC platform.

We computed the value of anomaly score using (14) and the top 5 most abnormal clips are shown in Fig. 22. Among the top 5 video clips, we detected one illegal vehicle U-turn (Fig. 22(c)), one pedestrian crossing against the traffic light (Fig. 22(b)), two vehicles left turning with pedestrian crossing in close proximity (Figs. 22(a) and (e)), which are legal but dangerous, and one pedestrian crossing outside the cross

(a) 1st (b) 2nd (c) 3rd (d) 4th (e) 5th

Fig. 22 (Color online) The top 5 abnormal clips detected using our Cas-LDA model. The contributing words (object motions) are highlighted in *red*

Table 6 Comparing our LDA formulation with the standard LDA formulation for video segmentation

	Cas-LDA with binary rep.	Cas-LDA with count-based rep.
Jun.	**87.2**%	**87.2**%
Rou.	**74.6**%	71.2%

Table 7 Area under ROC (AUROC) values for anomaly detection using our LDA formulation and standard LDA formulation

	Cas-LDA with binary rep.	Cas-LDA with count-based rep.
Jun.	**0.8589**	0.4919
Rou.	**0.7374**	0.4796

walk (Fig. 22(d)). Compared to the top 5 detected abnormal clips (Fig. 15 from Wang et al. 2009), our model seems to be more sensitive to abnormal behaviour caused by abnormal correlations of multiple objects rather than the abnormal motion patterns of each single object. For instance our model picked up a number of instances of vehicles and pedestrians moving on a collision course whilst the Dual-HDP detected a few clips where bicycles or pedestrians crossing outside the lanes/crosswalks, for which correlation modelling is not essential. In other words, these experiments demonstrate that our Cas-LDA model is more sensitive to those anomalies where each individual object behaves normally but collectively an object's behaviour is deemed abnormal due to our model's ability to detect contextually incoherent behaviours.

Comparing with Standard Topic Model Formulation As described in Sect. 5.2, the LDA models used in both stages of our Cas-LDA have an important difference from the standard topic model formulation (Hofmann 1999a; Blei et al. 2003). That is, each video document (clip) is represented by whether each event class occurs in that clip (a binary value), rather than by the counts of their occurrence. Tables 6 and 7 compare the effectiveness of these two different representations for video temporal segmentation and anomaly detection respectively. It is evident from these results that our binary document (clip) representation is superior to the standard counts-based representation especially for the task of anomaly detection.

7.7 Discussions

The Importance of Utilising Context for Behaviour Understanding Our extensive experimental results have demonstrated compellingly that learning behavioural context is hugely beneficial and can be crucial for understanding behaviours in complex dynamic scenes. This is because that the proposed behavioural context learning framework provides an effective means of decomposing a complex multi-object dynamic scene according to spatial, correlation and temporal distributions of object behaviours. Rather than building a single model based on a single representation, the discovered spatial, correlation and temporal context enables us to adapt different representations for different behaviours, and facilitates not only grouping, at multiple scales, different correlated behaviours for analysis, but also embedding temporal constraints for interpreting behaviours. As a result, better understanding of multi-object behaviour can be achieved, leading to high sensitivity particularly to subtle anomalies with improved robustness to noise at the same time.

Probabilistic Topic Models (PTM) vs. Dynamic Bayesian Networks (DBN) Our results suggest that PTMs outperform DBNs for both video segmentation and anomaly detection. One would have thought that a DBN is more suitable for dynamic scene understanding as the temporal order information is utilised, which should be particularly useful for temporal video segmentation. However, one clear drawback of DBNs against the Bag-of-Words based PTM is that it is more sensitive to noise, which explains the inferior performance of DBNs in our experiments.

Cascade PTM vs. Hierarchical PTM The results in Sect. 7.6 show that a cascaded PTM outperforms a hierarchical PTM such as the Dual-HDP model and MCTM for behaviour context modelling. Theoretically a hierarchical model may be more attractive as it allows for the clustering of words or atomic activities simultaneously in order to help the discovering of correlations among activities (Wang

et al. 2009). However, in practice, such a model is not always better and can suffer from a number of drawbacks: (1) A single hierarchical model needs more parameter to describe, and thus has poorer scalability and tractability and higher computational cost. As a result it is often less applicable to larger and more complex problems (e.g. the multi-camera behaviour modelling problem). This drawback is evident from our experiments reported in Sect. 7.6, where our model is compared with the Dual-HDP model in Wang et al. (2009). (2) The model inputs computed from real-world video data inevitably contain noise and errors. These errors can be propagated from the bottom to different layers of a hierarchical model with unpredictable effects. (3) In a hierarchical model the numbers of inputs in different layers from bottom to top are extremely imbalanced. For example for both the DDP-HMM in Kuettel et al. (2010) and MCTM in Hospedales et al. (2009), the bottom layer models video words which are local motion events and in the order of thousands per clip. The number of actions/topics in the layer above are in the order of dozens, whilst the number of temporal phases in the top layer is only a handful. This imbalanced modelling structure causes problems in detecting abnormities because those occurred at upper layers can be easily overwhelmed by those in the bottom layer, and become undetectable. This is reflected by the results reported in Hospedales et al. (2009). For fair comparison, in our experiments we used the same inputs for MCTM as for our cascaded LDA model, i.e. scene events. Even with this much reduced words number, the results obtained using MCTM are inferior to those of cascaded LDA (see Sects. 7.4 and 7.5). (4) A cascaded model enables us to incorporate easily behaviour spatial context information which has shown to be critical for understanding complex behaviour. On the contrary, integration of spatial context for the already complex hierarchical PTM would be very difficult both for designing inference and learning algorithms, and for maintaining tractability.

Cascaded LDA vs. Cascaded pLSA As pointed out by Blei et al. (2003), the advantage of LDA over pLSA is that LDA models the prior distribution of topics over documents and thus would fare better in generalisation to unseen documents and also overcome the over-fitting problem of pLSA. Our results indicate that LDA models indeed have better generative power which is beneficial for anomaly detection, but not for temporal context learning and video temporal segmentation. Similar observations were obtained on action recognition by Niebles et al. (2008). This suggests that for selecting different topic models in the proposed cascaded model structure, one needs to balance generative and discriminative power of a model according to the nature of the visual tasks. For instance, for robust anomaly detection, the model must be able to generalise well as normal behaviours can be executed in many variations which should not be confused with anomalies.

Binary Document Representation vs. Accumulative Count-Based Representation Our results show that a binary document representation is advantageous over count-based representation for topic model based visual behaviour understanding. This is due to two reasons: (1) for visual behaviour understanding, whether one type of behaviour has happened is more important semantically than how many times it has taken place. (2) More importantly, differing from text analysis where document representation is mostly free from noise and errors, 'clean' visual data for representation is mostly implausible due to image noise and visual ambiguities. Word counts in a document, is far more susceptible to noise compared to a binary representation. This explains why the performance of anomaly detection is improved drastically when the binary representation is adopted (see Table 7). Note that, more recently, the Indian Buffet Process (IBP) (Griffiths and Ghahramani 2005) has been introduced to model binary representation of documents. Extending our existing LDA modelling using IBP can be considered.

Relationships Between Different Behaviour Context The three types of behaviour contex studied in this work are closely related. Specifically, the learning of spatial context enables the efficient and effective learning of correlation and temporal context at different scales. Behaviour correlation context and temporal context are related in the sense that with different temporal context, i.e. in different temporal phases of a global behaviour, local behaviours are correlated in a different way leading to different correlation context. These two types of context also have important differences. More specifically, correlation context specifies how different local behaviours in a visual scene are correlated with each other. This is discovered in our framework by the learning of co-occurrence of scene events. Temporal context, on the other hand, corresponds to different temporal phases of a global behaviour. The inferred correlation context can thus be used as input to learn temporal context. This is precisely what we propose to do in our framework, that is, the topic profile inferred using a LDA is used as input to a clustering algorithm to discover temporal context.

Temporal Ordering Information Modelling Although the proposed cascaded topic model does not model temporal order of words, topics, or documents explicitly, the correlation context learned using the model can be utilised to detect abnormal behaviours caused by abnormal temporal order. In particular, we have demonstrated that the model can detect anomalies such as emergency vehicle interrupting normal traffic flow (see Sect. 7.5). These anomalies are

caused by abnormal temporal order of events (e.g. horizontal traffic should follow vertical traffic). However, this abnormal temporal order also results in abnormal co-occurrences of events (co-occurrence of horizontal and vertical traffic). Therefore, these anomalies can also be detected using our model. One can consider that temporal ordering information is captured implicitly by our model. Importantly, this implicit modelling of temporal ordering information makes the model more robust against noise in mode inputs. This is validated by our experiments on comparing our model with models that explicitly model temporal ordering information, including MOHMM and MCTM (see Sect. 7.5).

Computational Cost The computational cost of a PTM has two components, one in the learning stage for parameter estimation and the other in the testing stage for inference of latent variables. Compared to a more complex PTM such as a Dual-HDP or a HDP mixture model of Wang et al. (2009) and MCTM of Hospedales et al. (2009), the learning cost of the model proposed here is much lower. Note that this computational cost also depends on the learning algorithm so it is difficult to give an analytic form. In particular, there are two categories of learning algorithms, variational Bayesian, which is used in our work, and Gibbs sampling. The former is in general considered to be more efficient than the latter, although this also depends on the settings (e.g. how many samples and sweeps to use in Gibbs sampling). The testing stage is a different matter. As shown in Hospedales et al. (2009), one can develop an efficient Gibbs sampling method based on an approximation to online Bayesian inference. This can make a more complicated model such as MCTM run in real time during testing.

The Flexibility of Our Approach It is worth pointing out that the proposed approach is very flexible in many ways as follows: (1) Behaviour representation—although a discrete event based represented is adopted in this work, any bag-of-words representation can be used (e.g. for the MIT Traffic Dataset, a low-level motion feature based representation was used in Sect. 7.6). (2) Behaviour-footprint—in the current work, a behaviour-footprint is computed as an event histogram for each pixel location. Different ways of computing behaviour-footprint can be considered. Note that this histogram-based representation ignores temporal order of events; it is thus limited in describing behaviour spatial context. However, it is found in Sect. 7.2 that this representation is sufficient in segmenting the scenes experimented in this paper. This is because different semantic regions in those scenes are featured with different events or events of different frequency of occurrence. In a more complicated dynamic scene there could be regions that differ from each other only in the temporal order of event occurrence. One could then compute behaviour-footprint as a time series of event occurrences, at an extra computational cost, and to measure the similarity between footprints based on temporal order. (3) Topic model selection—two topic models, LDA and pLSA, are considered in this paper. However any topic model can be adopted. One could also employ different topic models at different stages of the cascade. (4) The number of stages in a cascade—in this work, a two stage cascaded topic model is formulated. However one could readily include more stages. For instance, for learning correlation context for multiple camera views decomposed into semantic regions across all views globally, a third stage can be added for capturing camera-view-level correlations. Moreover, with more complex behaviour patterns involving multiple objects from multiple views requiring more stages, the advantage of a cascaded model over a hierarchical model in terms of computational cost and scalability become more apparent. (5) Finally, although all datasets used in our experiments are featured with traffic scene, our approach is equally suitable for other public scenes of crowded spaces. For example, recently we have shown that the learning of behaviour spatial context facilitates the understanding of a busy underground station scene (Loy et al. 2009).

8 Conclusions

This paper defines comprehensively behavioural context and presents a novel framework for learning three different types of behavioural context, including behaviour spatial context, correlation context, and temporal context. For learning spatial context a semantic scene segmentation method is formulated. The learned spatial context is then exploited to compute a cascaded topic model for learning correlation and temporal context at multiple scales. The learned cascaded model is employed to address the problems of video temporal segmentation and context-aware anomaly detection. Extensive experiments are carried out using three different busy public space scenes to demonstrate the effectiveness of the proposed behavioural context learning framework and its usefulness for understanding complex multi-object behaviours in a crowded space.

The presented work has a number of limitations which need be addressed in the future work.

– Although dynamic behavioural context is discovered from video data, once learned the current framework does not accommodate the change of context, which is common when video data of long duration needs be analysed. For instance, the semantic layout of the scene and the correlations between objects can differ at different times of a day, different days of the year. One solution to this problem is via incremental learning. For example an incremental spectral clustering algorithm by Chi et al. (2007) can be adopted to update the scene segmentation

on-the-fly. An incremental topic model learning method by Canini et al. (2009) can also be employed for incremental learning of the topic models used in this work.
– The model complexity corresponding to the number of topics in each cascade stage is currently determined either a priori by domain knowledge or through cross validation. One could use a Bayesian model selection approach to determine the model complexity automatically (Wallach et al. 2009).
– The current method is only applicable to unseen data from the same scene of the same viewpoint. It would be desirable if a model learned from one scene can be used to facilitate the understanding of behaviour in a different scene (or the same scene captured from a different viewpoint). Solving that problem is particularly useful for abnormal behaviour detection because abnormal behaviours are typically rare; therefore there is a big incentive to transfer knowledge from one scene to another. One solution to the problem is by transfer learning. In the past 5 years, there have been extensive efforts on applying transfer learning techniques for object recognition and action recognition (Fei-Fei et al. 2006; Duan et al. 2009). However, it is not straightforward to apply those techniques for our problem here because: (1) Most transfer learning techniques are designed for discriminative models, with few exception such as the one-shoot-learning work by Fei-Fei et al. (2006). It is not clear which part of our topic models can be transferred. (2) Transferring target object/action/behaviour has been attempted before. But we are not aware of any previous work on transferring context, which poses additional challenges because behavioural context is defined in conjunction with behaviour and one cannot simply transfer context alone.
– The correlations modelled in this work is limited to co-occurrence. This is mainly due to the use of topic model which is based on a Bag of Words representation and unable to capture temporal order information. In other words, model robustness to noise is gained at the price of being unable to model more complex correlations. As a simple extension of the current framework, instead of clustering the topic profile of the second stage LDA for global temporal context modelling, one could adopt an HMM to model the temporal order of different phases explicitly. Recently there have been a number of attempts at re-introducing temporal ordering information to topic models (Griffiths et al. 2007; Wallach 2006), which may provide a solution for this problem. However, it remains an open problem on how to strike the right balance between model sensitivity and robustness when dynamics are modelled in a topic model.
– A spectral clustering based segmentation method is adopted for semantic scene segmentation in this work. However, there are a large variety of other image segmentation methods can be considered here. In particular, recently topic models have been employed for simultaneous image segmentation and object categorisation (Cao and Fei-Fei 2007; Blei and Lafferty 2007). In particular, when a video is split into multiple sub-sequences, a topic model based segmentation method, such as those introduced in Cao and Fei-Fei (2007), Blei and Lafferty (2007), can be applied for scene segmentation. Furthermore, since both scene segmentation and behaviour modelling can be done using topic models, it is possible to combine them into a unified topic model for doing both tasks simultaneously. We note that in a recent effort (Haines and Xiang 2009), a regional LDA model is formulated which attempts to encode spatial awareness into a LDA model for behaviour modelling. However, this model tends to over-segment and has a much higher computational cost compared to a standard LDA model, even when behaviour is modelled only at a single scale. Therefore further investigations are necessary for developing a unified topic model which is tractable and capable of performing multi-scale behaviour context modelling.

Acknowledgements We shall thank Xiaogang Wang for providing us with the MIT Traffic Dataset for our comparative experiments and evaluation. This work was partially supported by the UK Engineering and Physical Sciences Research Council (grant number EP/E028594/1).

References

Ali, S., & Shah, M. (2008). Floor fields for tracking in high density crowd scenes. In *European conference on computer vision* (Vol. II), Marseille, France, October 2008.
Bar, M. (2004). Visual objects in context. *Nature Reviews. Neuroscience*, *5*, 617–629.
Bar, M., & Aminof, E. (2003). Cortical analysis of visual context. *Neuron*, *38*, 347–358.
Bar, M., & Ullman, S. (1993). Spatial context in recognition. *Perception*, *25*, 343–352.
Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*, 143–177.
Blei, D. M., & Lafferty, J. D. (2007). Spatial latent Dirichlet allocation. In *Advances in neural information processing systems*.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.
Breitenstein, M. D., Sommerlade, E., Leibe, B., Van Gool, L., & Reid, I. (2008). Probabilistic parameter selection for learning scene structure from video. In *British machine vision conference*, Leeds, UK, September 2008.
Canini, K. R., Shi, L., & Griffiths, T. L. (2009). Online inference of topics with latent Dirichlet allocation. In *International conference on artificial intelligence and statistics*.
Cao, L., & Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent object segmentation and classification. In *International conference on computer vision*.
Carbonetto, P., de Freitas, N., & Barnard, K. (2004). A statistical model for general contextual object recognition. In *European conference on computer vision*, Prague, Czech Republic, May 2004.

Chi, Y., Song, X., Zhou, D., Hino, K., & Tseng, B. (2007). Evolutionary spectral clustering by incorporating temporal smoothness. In *ACM SIGKDD international conference on Knowledge discovery and data mining*.

Duan, L., Tsang, I. W., Xu, D., & Maybank, S. J. (2009). Domain transfer svm for video concept detection. In *IEEE conference on computer vision and pattern recognition*.

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 594–611.

Galleguillos, C., Rabinovich, A., & Belongie, S. (2008). Object categorization using co-occurrence, location and appearance. In *IEEE conference on computer vision and pattern recognition*, Alaska, USA, June 2008.

Greenhill, D., Renno, J., Orwell, J., & Jones, G. A. (2008). Occlusion analysis: learning and utilising depth maps in object tracking. *Image and Vision Computing*, *26*(3), 430–441.

Griffiths, T., & Ghahramani, Z. (2005). *Infinite latent feature models and the Indian buffet process* (Technical report). University College London.

Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2007). Integrating topics and syntax. In *Neural information processing systems*.

Gupta, A., & Davis, L. S. (2008). Beyond nouns: exploiting prepositions and comparative adjectives for learning visual classifier. In *European conference on computer vision*, Marseille, France, October 2008.

Haines, T., & Xiang, T. (2009). Video topic modelling with behavioural segmentation. In *IACM workshop on multimodal pervasive video analysis*.

Heitz, G., & Koller, D. (2008). Learning spatial context: using stuff to find things. In *European conference on computer vision*, Marseille, France, October 2008.

Hofmann, T. (1999a). Probabilistic latent semantic analysis. In *Uncertainty in artificial intelligence*.

Hofmann, T. (1999b). Probabilistic latent semantic indexing. In *SIGIR*.

Hospedales, T. M., Gong, S., & Xiang, T. (2009). A Markov clustering topic model for mining behaviours in video. In *International conference on computer vision*.

Kuettel, D., Breitenstein, M. D., Van Gool, L., & Ferrari, V. (2010). What's going on? Discovering spatio-temporal dependencies in dynamic scenes. In *IEEE conference on computer vision and pattern recognition*, San Francisco, USA, June 2010 (pp. 1951–1958).

Kumar, S., & Hebert, M. (2005). A hierarchical field framework for unified context-based classification. In *International conference on computer vision*, Beijing, China, October 2005.

Li, J., Gong, S., & Xiang, T. (2008a). Scene segmentation for behaviour correlation. In *European conference on computer vision*, Marseille, France, October 2008 (Vol. IV, pp. 383–395).

Li, J., Gong, S., & Xiang, T. (2008b). Global behaviour inference using probabilistic latent semantic analysis. In *British machine vision conference*, Leeds, UK, September 2008 (pp. 193–202).

Loy, C. C., Xiang, T., & Gong, S. (2009). Multi-camera activity correlation analysis. In *IEEE conference on computer vision and pattern recognition*, Miami, USA, June 2009 (pp. 1988–1995).

Lucas, B. D., & Kanade, T. (1981). An interative image registration technique with an application to stereo vision. In *Proceedings of imaging understanding workshop* (pp. 121–130).

Makris, D., & Ellis, T. (2005). Learning semantic scene models from observing activity in visual surveillance. *IEEE Transactions on Systems, Man and Cybernetics. Part B. Cybernetics*, *35*(3), 397–408.

Makris, D., Ellis, T., & Black, J. (2004). Bridging the gaps between cameras. In *IEEE conference on computer vision and pattern recognition*, Washington (pp. 205–210).

Malik, J., Belongie, S., Leung, T., & Shi, J. (2001). Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, *43*(1), 7–27.

Marszalek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In *IEEE conference on computer vision and pattern recognition*, Miami, USA, June 2009.

Murphy, K., Torralba, A., & Freeman, W. (2003). Using the forest to see the tree: a graphical model relating features, objects and the scenes. In *Neural information processing systems*.

Niebles, J., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, *79*(3), 299–318.

Palmer, S. (1975). The effects of contextual scenes on the identification of objects. *Memory and Cognition*, *3*, 519–526.

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). Objects in context. In *International conference on computer vision*, Rio de Janeiro, Brazil, October 2007.

Russell, D., & Gong, S. (2006). Minimum cuts of a time-varying background. In *British machine vision conference*, Edinburgh, UK, September 2006.

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

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.

Singhal, A., Luo, J., & Zhu, W. (2003). Probabilistic spatial context models for scene content understanding. In *IEEE international conference on computer vision and pattern recognition*, Madison, USA, June 2003.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.

Wallach, H. (2006). Topic modeling: beyond bag-of-words. In *International conference on machine learning*, Pittsburgh, USA, June 2006.

Wallach, H., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *International conference on machine learning*, Montreal, Canada, June 2009.

Wang, X., Tieu, K., & Grimson, E. (2006). Learning semantic scene models by trajectory analysis. In *European conference on computer vision* (Vol. 3, pp. 110–123).

Wang, X., Ma, K. T., Ng, G., & Grimson, E. (2008). Trajectory analysis and semantic region modeling using a nonparametric Bayesian model. In *IEEE conference on computer vision and pattern recognition*, Alaska, USA, June 2008.

Wang, X., Ma, X., & Grimson, W. E. L. (2009). Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(3), 539–555.

Wang, X., Tieu, K., & Grimson, E. (2010). Correspondence-free activity analysis and scene modeling in multiple camera views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 56–71.

Wolf, L., & Bileschi, S. (2006). A critical view of context. *International Journal of Computer Vision*, *69*(2), 251–261.

Xiang, T., & Gong, S. (2006a). Model selection for unsupervised learning of visual context. *International Journal of Computer Vision*, *69*(2), 181–201.

Xiang, T., & Gong, S. (2006b). Beyond tracking: modelling activity and understanding behaviour. *International Journal of Computer Vision*, *67*(1), 21–51.

Xiang, T., & Gong, S. (2008). Video behavior profiling for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(5), 893–908.

Yang, M., Wu, Y., & Hua, G. (2008). Context-aware visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(7), 1195–1209.

Zelnik-Manor, L., & Perona, P. (2004). Self-tuning spectral clustering. In *Neural information processing systems*.

Zheng, W., Gong, S., & Xiang, T. (2009). Quantifying contextual information for object detection. In *International conference on computer vision*, Kyoto, Japan, September 2009.