



Detecting and discriminating behavioural anomalies

Chen Change Loy*, Tao Xiang, Shaogang Gong

School of EECS, Queen Mary University of London, London E1 4NS, UK

ARTICLE INFO

Article history:

Received 20 April 2010

Received in revised form

6 July 2010

Accepted 20 July 2010

Keywords:

Anomaly detection

Dynamic Bayesian Networks

Visual surveillance

Behavior decomposition

Duration modelling

ABSTRACT

This paper aims to address the problem of anomaly detection and discrimination in complex behaviours, where anomalies are subtle and difficult to detect owing to the complex temporal dynamics and correlations among multiple objects' behaviours. Specifically, we decompose a complex behaviour pattern according to its temporal characteristics or spatial-temporal visual contexts. The decomposed behaviour is then modelled using a cascade of Dynamic Bayesian Networks (CasDBNs). In contrast to existing standalone models, the proposed behaviour decomposition and cascade modelling offers distinct advantage in simplicity for complex behaviour modelling. Importantly, the decomposition and cascade structure map naturally to the structure of complex behaviour, allowing for a more effective detection of subtle anomalies in surveillance videos. Comparative experiments using both indoor and outdoor data are carried out to demonstrate that, in addition to the novel capability of discriminating different types of anomalies, the proposed framework outperforms existing methods in detecting durational anomalies in complex behaviours and subtle anomalies that are difficult to detect when objects are viewed in isolation.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

The recent large-scale deployments of surveillance cameras have led to a strong demand in systems of automated anomaly detection in visual surveillance [1–3]. Earlier work is mostly focused on anomaly detection from well-defined simple behaviours in an uncrowded scenario [4,5]. More recently, the primary research focus has shifted to complex behaviour scenario in which a behaviour pattern is characterised by hierarchical temporal dynamics and/or complex correlations among multiple objects.

Anomaly detection in complex behaviours is challenging because the differences between real-life true anomalies (rather than exaggerated acts) and normal ones are often rather subtle visually and not well-defined semantically. One way to model such subtle differences is to consider that anomalies are associated with deviations in the expected temporal dynamics embedded in complex behaviours, which in turn can be considered as having layered hierarchical structures. In addition, different ways of deviations from the expected temporal dynamics lead to different types of anomalies, the discrimination of which has never been attempted to date although it is often of practical use in real-world applications. In a crowded multiple object scenario, anomaly detection becomes even more challenging because visual evidences often span across a large spatial and

temporal context, anomaly is thus difficult to detect if an object is viewed in isolation.

To facilitate effective modelling and anomaly detection for complex behaviours, it is natural to decompose the modelling task into a number of sub-tasks. Most existing techniques resort to *object-based decomposition* which employs a standalone model with the model structure being factorised in accordance with the corresponding temporal processes of individual objects [6,7]. However, object-based decomposition relies on object segmentation and tracking and therefore is prone to problems associated with occlusion and trajectory discontinuities when applied to a crowded wide-area scene. In addition, object-based decomposition will lead to very complex model structure making model learning and inference intractable in the presence of large number of objects. Moreover, it offers no mechanism for discriminating different types of anomalies and reducing the effect of noise and error from the observation space.

To address these problems, we propose to perform *behaviour-based decomposition* on a complex behaviour and model the decomposed behaviours with a *cascade of Dynamic Bayesian Networks* (CasDBNs), in which a DBN model at each stage is connected to the model in the next stage via its inferential output. More specifically, behaviour-based decomposition factorises the behaviour space into sub-spaces based on directly exploring the behaviour semantics defined by different temporal characteristics of the behaviour (e.g. co-occurrence, temporal order, and temporal duration) and the spatio-temporal visual context where the behaviour occurs. Behaviours are inherently context-aware, exhibited through constraints imposed by scene layout and the

* Corresponding author. Tel.: +44 20 7882 8019; fax: +44 20 8980 6533.

E-mail addresses: ccloy@dcqs.qmul.ac.uk, ccloy225@gmail.com (C.C. Loy), txiang@dcqs.qmul.ac.uk (T. Xiang), sgg@dcqs.qmul.ac.uk (S. Gong).

temporal nature of activities in a given scene. We believe that better behaviour modelling can be achieved based on behaviour-based decomposition because the important context-awareness nature of complex behaviours is exploited explicitly, which has been largely neglected by previous object-based decomposition based approaches.

Apart from employing a different decomposition strategy, the proposed framework differs from existing approaches in that it deploys multiple DBN models in a cascade structure. This model structure is motivated by the following key observations:

- (i) It is noted that different DBN models have different levels of sensitivity towards different types of anomalies. It is therefore possible to exploit this characteristic by employing a cascade of DBNs, with each of them being sensitive to one specific type of anomalies. This enables us to integrate the evidences from each DBN models to achieve a more accurate detection, and more importantly behaviour discrimination.
- (ii) It is well known that noise and error in the low-level visual features are inevitable in a real-world scenario. By constructing a cascade structure with each stage being connected using the inferential output of the previous stage, the models in later stages of the cascade will be less affected by the noise and error in the observation space.
- (iii) While a single model generally suffers from the scalability problem given large number of objects, a CasDBNs would benefit greatly from behaviour decomposition in avoiding this problem since the complexity of each individual model in the cascade is well controlled after the decomposition.

We present two instantiations of our framework to address two fundamental and open problems of anomaly detection in complex behaviours. In Section 4, we formulate the framework for *detecting and discriminating anomalies* by their abnormal temporal dynamics (e.g. atypical duration and irregular temporal order) embedded implicitly in the behaviour structure. In Section 5, the framework is used to address the problem of *modelling multi-object correlations* in a crowded wide-area scene and

detecting subtle anomalies that are difficult to detect when objects are viewed in isolation.

1.1. Discriminating different temporal causes of anomalies

It is not only necessary but also critical to both detect and discriminate different types of anomalies based on the temporal characteristics of expected behaviours. In many real-world scenarios, there could be only one type of anomalies that are deemed as critical for triggering an alarm. For instance, in a bank branch, a different order of “entering into the branch” and “using an ATM outside the branch” is of no significance. However, the durational abnormality in front of the ATM may be of more interest. On the other hand, in a convenience store, the temporal order of “paying” and “leaving the shop” is important, whilst variations in the time spent at these atomic actions of the shopper behaviour are less critical.

In order to model and differentiate behaviours by their intrinsic characteristics, we consider a complex behaviour as a spatio-temporal pattern organised naturally in an hierarchical structure. For instance, as can be seen from Fig. 1, a person’s typical behaviour in an office can include a sequence of ordered atomic actions with certain duration such as entering the office, working at a desk, printing, and leaving the office. Each atomic action itself is also composed of multiple constituents having certain duration and temporal order among them (e.g. entering the office can consist of opening the door and then walking toward the desk). A normal behaviour pattern would follow a typical order of atomic actions with certain duration. Deviation from either one or both of these temporal characteristics would cause an anomaly.

In this paper, we show that different DBN models can exhibit different levels of sensitivity given different types of anomalies. Based on this finding, we propose to decompose a complex behaviour based on different temporal characteristics, particularly the temporal order and temporal duration. This is achieved by exploiting different DBN models in a cascade, with each of

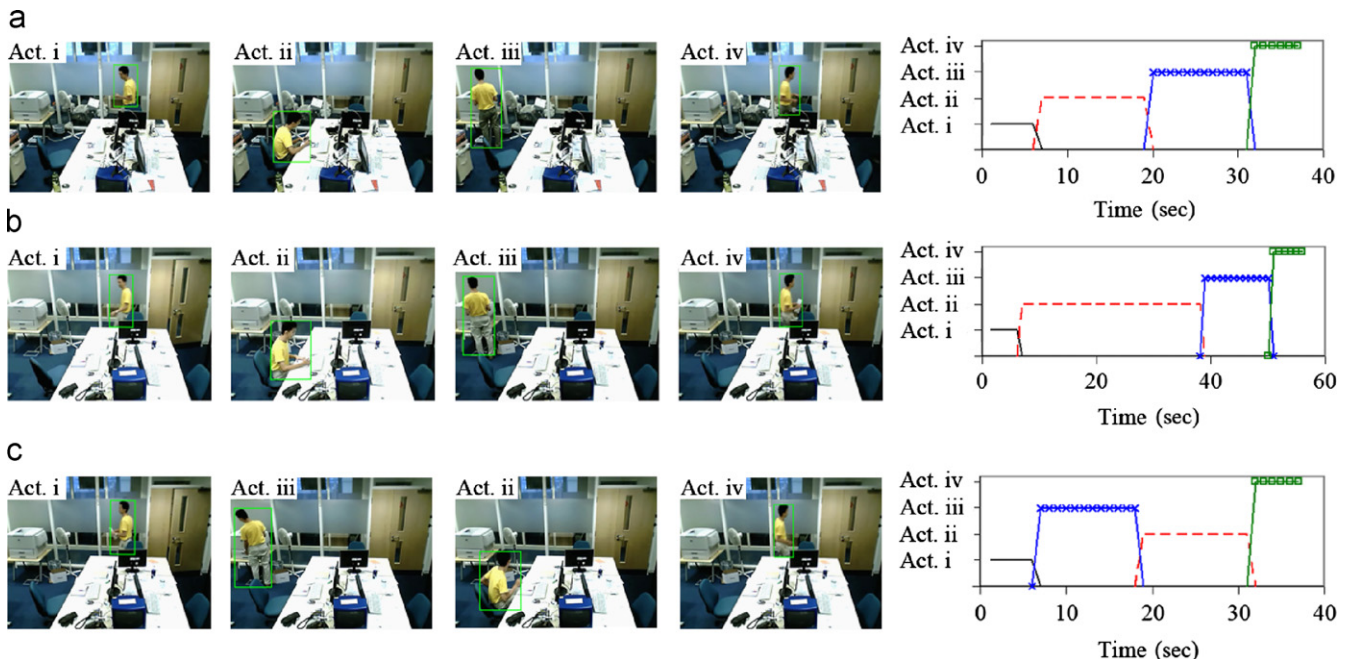


Fig. 1. Example frames of three behaviour sequences in an office environment and the associated ground truth of action occurrences. Although the behaviour sequences share the same set of atomic actions ([Act. i] entering, [Act. ii] working at a desk, [Act. iii] printing and [Act. iv] leaving), sequence (b) and sequence (c) exhibit abnormal temporal dynamics. (a) Normal behaviour sequence; (b) Behaviour sequence with atypical temporal duration; (c) Behaviour sequence with irregular temporal order.

them being employed to model one of the temporal characteristics. The resultant framework can then be deployed to detect and distinguish different types of anomalies, whilst existing techniques fail.

1.2. Detecting abnormal correlations

Behaviours involving multiple objects are inherently constrained by the visual contexts of a scene [8]. Specifically, a behaviour correlation can be either *local* or *global* depending on whether it takes place within a local or global context. The former corresponds to correlated objects in proximity in terms of space and time, whilst the latter corresponds to objects which are further apart in both space and time. Considering a public wide-area scene as shown in Fig. 2, anomaly detection in this case is challenging as visual evidences critical for detection often span across large spatial and temporal visual context. Importantly, potential anomalies are usually difficult to detect if objects are viewed in isolation.

Given the potentially different natures of abnormal correlations and the need to quantify their differences, we can put anomalies into three categories based on their visual distinctiveness and their frequency of occurrence in a training set. Anomalies in *Category-A* are often signalled by behaviour patterns that are visually very different from what have been observed from the training set. An example is given in Fig. 2(a) which shows a fire engine in an emergency causing interruption to the vertical traffic flow at a junction. *Category-B* corresponds to anomalies that are ambiguous due to their rare occurrence in the training set. Anomalies in *Category-C* are supported only by very weak visual evidence, i.e., featured with very subtle deviation from the normal temporal order/durations of different correlated temporal processes. An example is given in Fig. 2(b) showing a white van running the red light in the horizontal direction which has caused no interruption to the traffic flow. Anomalies in both Categories A and C refer to those that have never occurred in the training set, whilst anomalies in Category-B are those that appear in the training set but are statistically under-represented. From the perspective of a human observer, anomalies in Category-A are visually obvious thus easy to detect. In contrast, anomalies in both Categories-B and C are likely to be missed.

For all three categories of behavioural anomalies, a large number of objects are influencing each other either explicitly or

implicitly in a complex visual context. Anomalies thus can only be detected effectively and robustly by modelling the local and global context both spatially and temporally. To that end, we propose to decompose the complex behaviour semantically in accordance with the spatial contexts, and employ CasDBNs to model the temporal aspect of the decomposed behaviours. Specifically the decomposed behaviours occurring in a local context are modelled using DBNs in the first stage of a cascade. The global correlations of them are then modelled in the second stage. Based on this novel behaviour decomposition and model structure, the proposed approach is more sensitive to subtle and ambiguous anomalies (i.e. those in Categories-B and C) as compared to existing DBN-based approaches. Moreover, it is computationally more tractable and more robust to noise and errors in the behaviour representation, as will be shown in our experiments in Section 6.2.

2. Related work

A number of approaches have been proposed for behaviour modelling and recognition, including probabilistic graphical models (e.g. Dynamic Bayesian Networks (DBNs) [9–13], propagation net [14]), petri nets [15], syntactic approaches (e.g. context-free grammars [16], stochastic context-free grammars [17]) and logic based approaches [18]. Among these approaches, graphical models especially DBNs are the most popular method [19,20].

Various DBN topologies have been developed, which perform object-based decomposition and factorise the state space and/or observation space by introducing multiple hidden state variables and observation state variables, e.g. multi-observation HMM (MOHMM) [21], parallel HMM (PaHMM) [7] and coupled HMM (CHMM) [6]. In the case of single object behaviour modelling, there are also several attempts to embed hierarchical behaviour structure in the model topology. Examples include hierarchical HMM (HHMM) [22] and switching hidden semi-Markov model (S-HSMM) [10], in which the state space is decomposed into multiple levels of states according to the hierarchical structure of behaviour. In spite of these efforts, existing DBNs suffer from the following shortcomings and therefore are inadequate for either discriminating different types of anomalies or detecting subtle and ambiguous multi-object behavioural anomalies.

Ineffective and inefficient temporal duration modelling: A fundamental requirement in improving the fidelity of, therefore



Fig. 2. (a) Key frames of an abnormal traffic sequence caused by a fire engine that interrupted the normal traffic flow from vertical directions (sequence 1). (b) Key frames showing a white van running the red light. It can be seen that the vertical traffic has stopped and the horizontal traffic was expected; therefore no traffic interruption was caused and the sequence appears to be normal. However a careful examination can reveal that the white van crossed the junction slightly sooner than normal after the previous traffic flow has finished, which gives away the fact that the red light for the horizontal traffic (invisible from the scene) was still on. Note that the behaviour of each individual object in these two sequences was normal (e.g. no illegal U-turn or driving on the pavement) when viewed in isolation (sequence 2). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

reducing false alarm from, an algorithm for detecting complex behavioural anomalies that exhibits long-term temporal dependency is the ability to explicitly and flexibly model the duration of the behaviour [23,24]. Hidden semi-Markov model (HSMM), a.k.a. variable duration HMM (VDHMM) [25,26] uses continuous probabilistic density function (pdf) to achieve more accurate behaviour duration modelling than a first-order HMM [27]. However, the introduction of an underlying semi-Markov process results in a significant increase in computational and numerical complexity [28]. As compared to HSMM, expanded state HMMs (ESHMMs) [28], such as multinomial HSMM (Mult-HSMM) [23] and Coxian HSMM (Cox-HSMM) [24], are more widely used for behaviour duration modelling [24,10,23,12,29]. ESHMMs avoid the use of semi-Markov model thus their computational costs are lower than that of HSMM. Nevertheless the performance and efficiency of ESHMMs are still not satisfactory when dealing with long behaviour sequences as these methods require large number of sub-phases to achieve accurate approximation of the true duration distribution. As an hierarchical extension of ESHMMs, S-HSMM [10] models both the temporal order of behaviour and the associated durational characteristics. However, in order to both capture the hierarchical behaviour structure and approximate the temporal duration in a single model, the S-HSMM inevitably has a complex structure with a large number of model parameters. Importantly, among the existing studies, although the one on S-HSMM [10] recognised the importance of both temporal order and temporal duration in detecting anomalies, there has been no attempt to differentiate them.

Poor scalability for complex multi-object behaviour modelling: In addition to the problem of dealing with occlusions and trajectory discontinuities in a busy multi-object scene, the existing object decomposition based DBNs suffer from the lack of scalability when presented with large number of objects. For instance, exact inference on a CHMM [6,30] beyond two chains (each chain corresponds to one object) is likely to be computationally intractable [30]. The same problem should surface for the dynamically multi-linked HMM (DML-HMM) [9]. It is also noted that previous studies are concerned with object correlations in small local context, and there is no investigation on detecting anomalies that are ambiguous or supported by weak evidence in a wide-area scene.

Vulnerable to error and noise in behaviour representation: Due to the limited availability of abnormal data samples, existing approaches rely on normal data samples for model construction. Since both a noise contaminated normal pattern and a real abnormality cannot be explained by the trained model, it is critical that a behaviour model is robust to error and noise in behaviour representation. Conventional DBN models learn directly from a noisy observation space and there is no mechanism to stop the error propagation through the model topology. As a result, high false alarm rate is expected given the inevitably noisy inputs from a real-world busy scenario.

In contrast, the proposed framework is advantageous in the following aspects:

- (i) The framework is able to *discriminate different types of anomalies* whilst existing techniques fail. The same capability is not available in a single DBN model since it is either tailored towards one type of anomaly or there is no mechanism to differentiate the anomalies.
- (ii) The proposed framework allows for *explicit modelling of behaviour duration*. The advantages of such a method are two-fold: (a) by modelling the duration explicitly, the proposed framework is more sensitive to durational anomalies, as compared to existing DBN models that model

duration based on implicit and non-parametric approximation of the true duration distribution; (b) it is also computationally more tractable for complex behaviour modelling.

- (iii) For complex multi-object behaviour modelling, behaviour-based decomposition avoids the occlusion problem commonly faced by object-based decomposition. Importantly, the proposed framework is *computationally more tractable and more scalable*.
- (iv) The proposed framework is *more robust to noise and error* in behaviour representation.

It is worth pointing out that cascade structure of DBNs has been considered for activity analysis [31,32]. Oliver et al. propose a layered HMM (LHMM) to capture different levels of temporal details when recognising human activity [31]. The LHMM is essentially a cascade of HMMs, in which each HMM accepts observation vectors processed with different time scales. Zhang et al. [32] present a similar framework based on LHMM with each stage of the cascade being employed to learn different levels of actions exhibited from individual to group of people. The CasDBNs formulated in this paper differ significantly from previous work [31,32] in the following aspects: (1) our framework decomposes behaviours based on temporal characteristics and visual context, a different cascading strategy is thus formulated; (2) our ultimate goal on detecting and discriminating video anomalies are different than that of [31,32]. Apart from DBNs, cascade structure based on topic models has been employed for activity analysis [33]. Despite the method has been shown to be capable in detecting anomalies in a global context, it is limited to modelling static causal relationships without taking the temporal ordering of behaviours into account. The model is thus unable to detect anomalies embedded in the temporal structure of correlation.

Compared to our earlier version of this work [34], we formulate in this paper a generic framework for discriminating different temporal causes of anomalies apart from detecting abnormal correlation. Besides, in the experiment on abnormal correlation detection, more extensive evaluations are conducted to compare the proposed framework with alternative models.

3. Cascaded Dynamic Bayesian Networks

In the proposed framework, the decomposed behaviours are modelled using a cascade of DBNs (CasDBNs), with each stage of the cascade being connected to the next stage via its inferential outputs. In this section, we describe the model structure and training strategy of CasDBNs, and how on-line filtering can be carried out for anomaly detection.

3.1. Model structure

The proposed CasDBNs combine two stages of DBNs in a cascade. Fig. 3 illustrates the generic structure of the proposed framework. The framework is flexible in that different types and numbers of DBNs can be employed in different stages. For simplicity and clarity of explanation in this section, we use first-order hidden Markov models (HMMs) as an example of first-stage models of the framework and a multi-observation HMM (MOHMM) [9] as the second-stage model. The HMMs at the first stage are denoted as Λ^1 , where $\Lambda^1 \in \{\lambda_r | r = 1, \dots, R\}$, and R corresponds to the number of HMMs, which varies for different

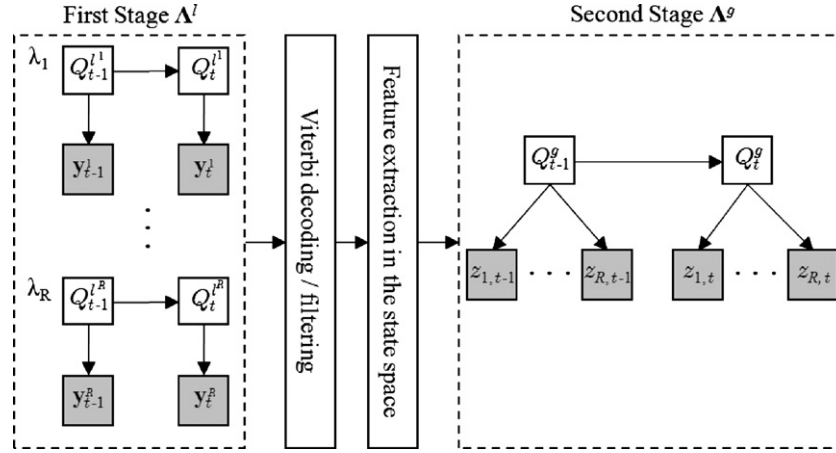


Fig. 3. The proposed cascade model with two time slices unrolled at each stage of the model. Different types of DBNs can be employed in different stages. In this example, Hidden Markov models are used as the first-stage models and a Multi-Observation HMM is employed in the second stage. Observation nodes are shown as shaded squares and hidden nodes are shown as clear squares.

computational tasks (see Sections 4 and 5). The MOHMM in the second stage is denoted as Λ^g .

In this example, the hidden variable of the r th HMM in the first stage λ_r is a discrete random variable denoted as $Q_t^r \in \{q_i^r | i = 1, \dots, K^r\}$, where K^r represents the number of hidden states. Similarly, the hidden variable of Λ^g is denoted as $Q_t^g \in \{q_i^g | i = 1, \dots, K^g\}$. Observations of both stages can be discrete or continuous inputs. In this example, the observations of the first stage are discrete values and denoted as \mathbf{y}_t^r , whilst the inputs to the second stage are also discrete, denoted as $\mathbf{z}_t \in \{z_1^t, \dots, z_R^t\}$, where R here represents the number of observation nodes in the second-stage MOHMM. Note that in this example, the number of observation nodes in the second-stage MOHMM equals to the number of HMMs in the first stage. These numbers may be different given different feature extraction scheme in state space (see Section 3.2).

We assume that all models are first-order Markov, i.e. $P(Q_t^r | Q_{1:t-1}^r) = P(Q_t^r | Q_{t-1}^r)$ and $P(Q_t^g | Q_{1:t-1}^g) = P(Q_t^g | Q_{t-1}^g)$. We also assume that the observations are conditionally first-order Markov, i.e. $P(\mathbf{y}_t | Q_{1:t-1}^r) = P(\mathbf{y}_t | Q_{t-1}^r)$ and $P(\mathbf{z}_t | Q_{1:t-1}^g) = P(\mathbf{z}_t | Q_{t-1}^g)$. It is assumed that the conditional probability distributions (CPDs) between a discrete observation node and discrete hidden variable being multinomial distribution, whilst CPDs between a continuous observation node and discrete hidden variable being conditional linear Gaussian distribution [35].

3.2. Model learning

The learning of the first-stage models precedes that of the second-stage models. Specifically, to learn a first-stage model λ_r , a training sequence of length T , $\mathbf{y}_{1:T}^r = (\mathbf{y}_1^r, \dots, \mathbf{y}_T^r)$, in which \mathbf{y}_t^r is a multi-dimensional feature vector for behaviour representation, is used to estimate model parameters through the Baum-Welch algorithm [36]. To prevent the algorithm from converging to a poor local optimum, parameters must be initialised properly. This is achieved by performing k -means clustering on the training data with the number of clusters k corresponding to the number of hidden states in the model. Subsequently, the model parameters are initialised based on the clustering results.

After the first-stage training, we proceed to the second phase of the training. First, with the same training sequence, the most probable explanation (MPE) in the state space of the first-stage

model λ_r is inferred by using the Viterbi algorithm [37], given as

$$Q_{1:T}^r = \underset{Q_{1:T}^r}{\operatorname{argmax}} P(Q_{1:T}^r | \mathbf{y}_{1:T}^r). \quad (1)$$

To train the second-stage model, the MPE is transmitted to a state feature extraction component (see Fig. 3). As an important interface between the second-stage model and first-stage models, the component is responsible for extracting important features from the MPE to form an observation vector for the second-stage model. The feature extraction function is written as

$$\mathbf{z}_t = f(Q_t^1, \dots, Q_t^r, \dots, Q_t^R). \quad (2)$$

Note that the feature extraction function may vary for different computational tasks, ranging from simple concatenation of most probable states, state duration extraction, to more elaborate methods such as principal component analysis [38] or neural networks [39]. The intermediate observation vector \mathbf{z}_t is then used as inputs by the second-stage model Λ^g for model learning. Parameters estimation for the second-stage model is carried out with the similar steps applied to the first-stage models.

3.3. On-line filtering

On-line filtering is an inference process to recursively estimate the belief state. It is known as ‘filtering’ because we are filtering out the noise from the observations [35]. In this study, the purpose of performing filtering is to compute the likelihood values on-the-fly with respect to the cascade model. This process does not require the past inputs before the current time instance for computation; the computational time and the required memory space are thus constant over time. Specifically, given an unseen sequence $\mathbf{y}_{1:T}^r$, our aim is to obtain the normalised log-likelihood LL_t^r at time t with respect to the first-stage model and the normalised log-likelihood LL_t^g with respect to the second-stage model, which are given as

$$LL_t^r = \frac{1}{t} \log P(\mathbf{y}_{1:t}^r | \lambda_r), \quad (3)$$

$$LL_t^g = \frac{1}{t} \log P(\mathbf{z}_{1:t} | \Lambda^g), \quad (4)$$

where $P(\mathbf{y}_{1:t}^r | \lambda_r)$ and $P(\mathbf{z}_{1:t} | \Lambda^g)$ are obtained from the computation of marginal probabilities in the filtering process. In particular, the marginal probability $P(Q_t^r | \mathbf{y}_{1:t}^r)$ of a first-stage model is computed

as a function of current input \mathbf{y}_t^r and prior belief state $P(Q_{t-1}^r | \mathbf{y}_{1:t-1}^r)$:

$$P(Q_t^r | \mathbf{y}_{1:t}^r) \propto P(\mathbf{y}_t^r | Q_t^r, \mathbf{y}_{1:t-1}^r) P(Q_t^r | \mathbf{y}_{1:t-1}^r) \\ = P(\mathbf{y}_t^r | Q_t^r) \left[\sum_{Q_{t-1}^r} P(Q_t^r | Q_{t-1}^r) P(Q_{t-1}^r | \mathbf{y}_{1:t-1}^r) \right]. \quad (5)$$

Based on Markovian assumption, we can replace $P(\mathbf{y}_t^r | Q_t^r, \mathbf{y}_{1:t-1}^r)$ with $P(\mathbf{y}_t^r | Q_t^r)$. Under the same assumption, $P(Q_t^r | \mathbf{y}_{1:t-1}^r)$ can be computed from the prior belief state. The normalising constant, which makes the probabilities sum up to 1, is denoted as $c_t^l = P(\mathbf{y}_{1:t}^r)$. It is obtained during the forward message passing phase in the on-line filtering. By multiplying all the normalising constants arising during the filtering, we can compute $P(\mathbf{y}_{1:t}^r | \lambda_r) = \prod_{t=1}^T c_t^l$ and obtain LL_t^r according to (3).

To compute LL_t^g , we need to estimate the local hidden state Q_t^r instantly at every time t . To estimate Q_t^r , the probabilities $P(Q_t^r = q_i^r | \mathbf{y}_{1:t}^r)$ are first computed using (5). The most likely hidden state is then determined by choosing the hidden state that yields the highest probability:

$$Q_t^r = \underset{q_i^r}{\operatorname{argmax}} P(Q_t^r = q_i^r | \mathbf{y}_{1:t}^r). \quad (6)$$

With the most likely hidden state Q_t^r obtained using (6), the observation input for the second-stage model is computed using (2). Subsequently, we compute marginal probability of stage-two model by replacing Q_t^r with Q_t^g and \mathbf{y}_t^r with \mathbf{z}_t^r in (5). We can then obtain $P(\mathbf{z}_{1:t} | \Lambda^g)$ by multiplying the normalising constants c_t^g and compute LL_t^g following (4).

Note that we used the Viterbi algorithm to obtain the MPE for training because it provides more accurate estimation of hidden state path for the training of the second-stage model. In the testing stage, however, we employed the on-line filtering method and find the most likely state which has the maximum probability. A set of such states may not be exactly the same as those obtained using the Viterbi algorithm, but they do give a good approximation based on our experimental results. More importantly, the computational cost is much lower than the Viterbi algorithm, and it permits the on-line estimation of the log-likelihoods.

4. Discriminating different temporal causes of anomalies

We decompose a complex behaviour with hierarchical structure based on two key temporal characteristics, i.e., temporal order and temporal duration. This is achieved by using different DBNs to model different temporal characteristics of a given behaviour sequence.

4.1. Model structure and learning

In the first stage of CasDBNs, we employ a two-layer HHMM represented as a DBN [40] to model the hierarchical structure of a complex behaviour (see Fig. 4). HHMM is chosen because it is effective in modelling different stochastic levels and length scales that present in a complex behaviour [41,22]. In particular, the children states at the bottom layer of HHMM are used to learn the constituent parts of atomic actions and the parent states at the top layer of HHMM are employed to learn the atomic actions. The state of the HHMM in layer d at time t is denoted as $Q_{d,t}^r$. Since the number of first-stage model $R=1$ in this case, the superscript r is omitted. Consequently, the states of the whole model at time t can be represented by a vector $[Q_{1,t}^l, Q_{2,t}^l]$. The binary indicator F_t is introduced here to enforce the fact that the top layer of HHMM can only change state when the state at the bottom layer is finished [40]. This DBN structure offers several advantages compared to the original model proposed by Fine et al. [42], of which the two main advantages are: (1) it allows one to use generic DBN learning and inference methods [35]; (2) the exact inference time complexity of the DBN structure is only $O(Q^D T)$ compared to $O(Q^D T^3)$ of original HHMM implementation, where D is the total number of layer in an HHMM, Q is the number of states in each layer, and T is the length of a sequence.

The first stage HHMM is employed to detect atypical temporal order in a behaviour sequence. Specifically, we train the model using normal sequences so that the hidden state transitional probability captures the temporal order of normal activity sequences. Consequently, a sequence with irregular temporal order is assigned with a low log-likelihood since its temporal order is poorly explained by the trained model. The normal behaviour sequences used for training are manually segmented into atomic actions. Note that this step is only performed during the training phase. Once trained, the model can be used for automatic temporal segmentation. Alternatively, one can perform automatic temporal segmentation during the training phase using the methods proposed in [21,43].

A hybrid input MOHMM [44] is employed in the second stage of the CasDBNs (see Fig. 4) which is designed to model the temporal duration aspect of a behaviour explicitly. The MOHMM has two observation nodes, one being discrete and the other being continuous. To obtain the observation features for the MOHMM, the MPE through the first stage HHMM is first estimated using the Viterbi algorithm using (1). The intermediate observation feature vector is then obtained using the feature extraction function f defined in (2), which has become a function to convert the one-dimensional sequence of state values $Q_{1:T}^l$ into a two-dimensional sequence of observation features encoding the state label and duration of the state, given as

$$\mathbf{z}_i = (z_{1,t}, z_{2,t}) = (q_i^l, |q_i^l|), \quad (7)$$

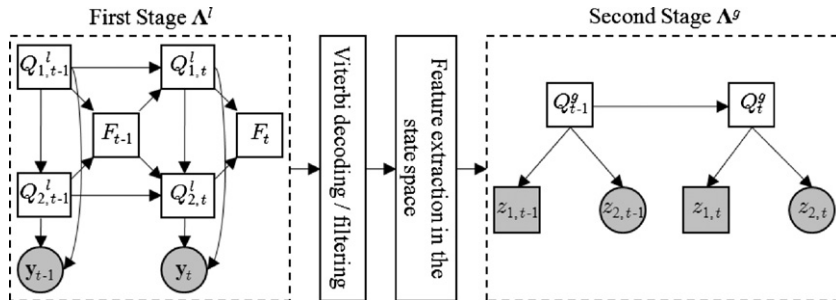


Fig. 4. The DBN representation of the model employed for detecting and differentiating different types of anomalies. Continuous-valued nodes are represented in circle and discrete-valued nodes are shown in square. Shaded nodes are observed whilst the remaining nodes are hidden.

where q_i^l is the state label and $|q_i^l| \geq 0$ is the corresponding state duration and they correspond to the discrete and continuous observation nodes, respectively, in the second-stage MOHMM (see Fig. 4). Note that the variable \mathbf{z} defined in (7) differs from the one in (2) in that, the \mathbf{z} in (2) is formed at every time slices, whilst the \mathbf{z} in (7) is only available after the end of an atomic action is automatically segmented by the first stage HHMM. After we obtain the intermediate inputs of the second-stage model, the learning process of the model proceeds as described in Section 3.

With this structure, the CasDBNs in Fig. 4 are able to model the duration explicitly with a continuous observation density. Specifically, the probability $P_i(d)$ of state occupancy for d consecutive frames at the i th state follows a Gaussian distribution:

$$P_i(d) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(d-\mu_i)^2}{2\sigma_i^2}\right), \quad (8)$$

where μ is the mean duration and σ^2 is the variance. Note that we can use a mixture of Gaussians to approximate a more complex duration distribution. However, a single Gaussian is found to be sufficient for our experiments.

4.2. Anomaly detection and differentiation

For anomaly detection, given an unseen sequence, we first compute its normalised log-likelihood LL_T^l and LL_T^g at time T with respect to the first stage HHMM and second stage MOHMM using (3) and (4), respectively. We also perform an additional step to normalise LL_T^l and LL_T^g with respect to the respective log-likelihood range obtained during the training stage. The normalisation aims to minimise the dominance of the first-stage log-likelihood over the second-stage log-likelihood due to discrepancy in the size of input feature vector. Anomaly detection is then achieved by summing the log-likelihoods of the two stages, $LL_T^{joint} = LL_T^l + LL_T^g$. Specifically, if

$$LL_T^{joint} < Th^{sum}, \quad (9)$$

where Th^{sum} is a pre-defined threshold, the unseen sequence is detected as an anomaly.

With the behaviour-based decomposition, the proposed cascade model can also be utilised for anomaly differentiation. In particular, our CasDBNs are designed in a way that the DBNs at the two stages have different levels of sensitivity to different types of anomalies. Specifically, the first stage HHMM is more sensitive to irregular temporal orders than atypical durations, whilst the second stage MOHMM, specially designed for more accurate duration modelling, is more sensitive to durational anomalies. The different characteristics of the two models are taken advantage of by the following procedure for anomaly differentiation. Firstly, an unseen behaviour sequence is examined if it is an anomaly using (9). Secondly, if it is classified as an anomaly, LL_T^l and LL_T^g of the unseen behaviour sequence are compared with two thresholds Th^l and Th^g , respectively. It is then classified into different types of anomaly using the decision rules listed in Table 1. The ‘-’ symbols in Table 1 imply that the framework does not rely on the log-likelihood generated by that particular stage for decision making because the corresponded

model is less sensitive to that anomaly type. The two thresholds Th^l and Th^g are determined automatically through a grid search using cross validation. More precisely, given a validation dataset, the optimal values of Th^l and Th^g are determined as those that yield the best cross-validation accuracy.

4.3. Discussion

One of the key features of our CasDBNs is that the state occupancy duration, which corresponds to the duration of atomic actions in a complex behaviour, is modelled explicitly through a hybrid input MOHMM. The formulation of hybrid input MOHMM is similar to that proposed by Kimball and Como [44] for accurate audio segmentation. Here we extend the idea for modelling activity duration and durational anomalies detection. The way we model duration is in contrast to most existing DBN-based approaches for anomaly detection [10,23], which perform implicit duration modelling. Here we provide an in-depth discussion on the pros and cons of existing DBNs on duration modelling and how explicit modelling can bring about more accurate and importantly, computationally more efficient duration modelling, therefore resulting in durational anomalies being better detected and distinguished from those caused by abnormal temporal order of atomic actions.

Let us first look at how state occupancy duration is modelled implicitly using a standard first-order HMM. The probability of staying at a certain hidden state in a first-order HMM decreases exponentially with time, with the state duration following a geometric distribution. Specifically, the probability of state occupancy for d consecutive frames at the i th state $P_i(d)$ is equal to the probability of $d-1$ self-loop transitions and a state exit transition:

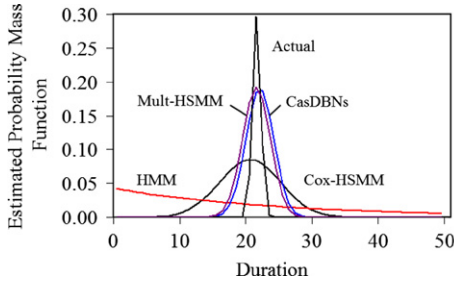
$$P_i(d) = a_{ii}^{d-1}(1-a_{ii}), \quad (10)$$

where a_{ii} is the self-transition probability. An example of the geometric distribution is shown in Fig. 5 (HMM curve). However, typical behaviour sequences rarely follow a geometric duration distribution. Instead, in most cases there will be an expected duration for each atomic actions, and duration that is either too short or too long would be considered as abnormal. First-order HMM is thus not suitable to model duration distribution of typical behaviour sequences and insensitive to durational anomalies.

Expanded state HMMs (ESHMMs) [28], such as multinomial HSMM (Mult-HSMM) [23] and Coxian HSMM (Cox-HSMM) [24] aim to provide a more accurate modelling of duration of arbitrary distribution through expanding a hidden state into interconnected sub-phases. However, the computational cost of learning and inference for an ESHMM increases drastically as the length of a sequence increases. This makes ESHMMs unsuitable for modelling behaviour sequences with long duration. In particular, for a Mult-HSMM, the number of sub-phases required has to be the same as the maximum duration of a behaviour sequence, which in turn determines the number of free parameters needed to describe the model. Fig. 5 show that although it can approximate the actual distribution accurately, it needs a large number of parameters and thus computationally expensive. To overcome this problem, Cox-HSMM was proposed which requires fewer parameters in duration modelling (see the table in Fig. 5) by approximating the duration distribution with fewer sub-phases. However, as can be seen from Fig. 5, fewer number of sub-phases leads to poorer approximation as compared to Mult-HSMM. In essence, Cox-HSMM requires more sub-phases to maintain the same level of approximation accuracy given sequence with increased duration. This will lead to the increase of parameters in the model. To make the ESHMMs computationally tractable,

Table 1
Decision rules employed in anomaly differentiation.

$LL_T^l < Th^l$?	$LL_T^g < Th^g$?	Decision
Yes	-	Abnormal temporal order
-	Yes	Atypical duration



Model	HMM	Cox-HSMM	Mult-HSMM	CasDBNs
No. of Parameter	5	41	63	12

Fig. 5. Comparing different DBNs with our CasDBNs for modelling state occupancy distribution. The distribution of the actual duration follows a Gaussian distribution with a mean of 20 time slices and standard deviation of 5. Estimated duration distributions are obtained using different models, namely HMM, Mult-HSMM, Cox-HSMM and CasDBNs. For the Mult-HSMM, the number of phases was set to 30, which was the maximum duration obtained from the synthetic sequences, while the number of phases of Cox-HSMM was set to 10 in the study. Single Gaussian was used in the observation node of second-stage model in CasDBNs. The table next to the plot summarises the number of parameters of different models used for duration modelling.

compromise has to be made to reduce the number of sub-phases, therefore sacrificing the approximation accuracy and detection performance.

The only way to achieve both accurate duration modelling and small number of model parameters regardless of the length of a behaviour sequence is to model the duration explicitly in a parametric form. Our CasDBNs have done exactly that through the second stage MOHMM. In particular, the framework requires a fixed number of Gaussian mixtures (even though a single Gaussian was used in our study) to model a duration distribution, and importantly the number of mixtures is determined by the complexity of the duration distribution rather than the length of the sequence. Therefore, the proposed framework only needs to adjust its Gaussian parameters given sequences with arbitrary duration length. Fig. 5 show that with a much smaller number of parameters, our CasDBNs can achieve the same accuracy as Mult-HSMM and outperforms Cox-HSMM. Thanks to the decomposition strategy, in most cases the number of training data needed for our CasDBNs would not be substantially different from an HMM.

5. Detecting abnormal correlations

In this section, the proposed framework is formulated for detecting abnormal correlations among multiple objects in a wide-area outdoor scene.

5.1. Behaviour decomposition based on visual context

Based on visual context learning, we decompose behaviours in a complex wide-area scene into regions in which distinctive behaviour patterns are detected and represented as discrete local atomic events. This is achieved by using a discrete event based approach introduced by our previous work [8]. Without relying on object segmentation and tracking, the approach is not affected by the severe occlusions commonly occurred in a busy outdoor scene. We provide a brief description of the approach here to facilitate explanations in other sections appeared later.

A continuous video sequence \mathbf{V} is first segmented uniformly into T non-overlapping video clips $\mathbf{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_t, \dots, \mathbf{v}_T\}$, with each video clip \mathbf{v}_t containing N_f frames. Foreground blobs are represented as 10 dimensional feature vectors which include object centroid (\bar{x}, \bar{y}) , width and height of bounding box (w, h) , occupancy (R_f) , ratio of the dimension (R_d) , the mean optical flow of the bounding box (u, v) , and the scaled optical flow $(R_u = u/w, R_v = v/h)$, given as $\mathbf{f}_{blob} = [\bar{x}, \bar{y}, w, h, R_f, R_d, u, v, R_u, R_v]$. The blobs are then clustered using k -means into a set of atomic events (e.g. vehicles stop at the middle of intersection waiting for right-turning) across all the frames in a clip \mathbf{v}_t . Upon obtaining atomic events for all clips, global clustering based on Gaussian mixture

model (GMM) is performed to group the atomic events into coarse global event classes. Based on the distributions of the clusters, spatial scene segmentation is carried out to decompose a scene \mathbf{S} into R semantic regions with R automatically determined through model selection, where $\mathbf{S} = \{s_1, \dots, s_r, \dots, s_R\}$. Specifically, two pixels are considered to be similar and grouped together if similar classes of events occurred there. Consequently, behaviour patterns within each segmented region are similar to each other whilst being different from those in other regions. To detect the atomic events more accurately, the aforementioned clustering method is repeated within each region with automatic feature selection to group foreground blobs into finer regional event classes. As a result, a video clip \mathbf{v}_t is spatially represented by segmented regions (see Fig. 6(b)), each of which contains a set of correlated regional atomic events. To construct the input features for the proposed cascade model, we represent the behaviours captured in a video clip \mathbf{v}_t using R binary vectors $\{\mathbf{y}_t^1, \dots, \mathbf{y}_t^r, \dots, \mathbf{y}_t^R\}$, which correspond to the occurrence of regional events in each region. Specifically, a binary vector \mathbf{y}_t^r is given as $\mathbf{y}_t^r = (y_{1,t}^r, \dots, y_{n,t}^r, \dots, y_{N_r,t}^r)$. We have

$$y_{n,t}^r = \begin{cases} 1 & \text{if atomic event } e_n^r \text{ occurs in region } s_r, 1 \leq n \leq N_r \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where e_n^r is the atomic event belonging to the n th regional event class in region s_r and N_r is the total number of regional event classes in region s_r .

5.2. Model structure and learning

The first stage of the CasDBNs for multi-object behaviour modelling is composed of multiple MOHMMs (see Fig. 7), each of which is used to model the temporal evolution of regional atomic events within a single region. The second stage consists of a MOHMM for modelling the state sequences inferred from the first stage MOHMMs, and is responsible for learning the global correlations among decomposed local behaviours across regions. The MOHMMs in both stages are ergodic models having discrete-valued observation nodes and discrete hidden variables. Note that discrete MOHMMs are chosen because of the discrete representation of the regional events. In contrast to conventional HMM that emits a symbol in a given state, a MOHMM allows a state to produce multiple symbols in every time step. It is thus ideal for modelling multiple atomic events temporally in each region in the first stage, and modelling the temporal correlations of local behaviours across multiple regions in the second stage.

Parameter estimation is carried out using Baum–Welch algorithm. After the first-stage training, the MPE through a first-stage model λ_r is obtained by using the Viterbi decoding



Fig. 6. Scene segmentation according to spatial visual context. (a) A traffic scene, (b) segmented regions.

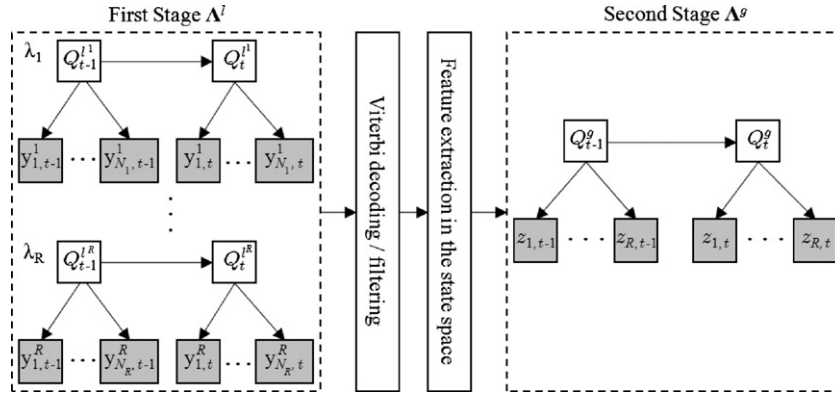


Fig. 7. The proposed cascade model for detecting abnormal correlations among multiple objects.

according to (1). We use the feature extraction function f to concatenate the MPEs obtained across first-stage models to form the second-stage input \mathbf{z}_t :

$$\mathbf{z}_t = (Q_t^{11}, \dots, Q_t^{1r}, \dots, Q_t^{1R}) \quad (12)$$

where $r=1, \dots, R$. The second-stage input \mathbf{z}_t is then used for second-stage model training. The parameter estimation procedures follow the same steps implemented during first-stage training.

5.3. Detecting abnormal correlations

In the detection phase, both normalised log-likelihood values LL_t^1 and LL_t^g can be used for anomaly detection. However, since we are interested in global behaviour anomalies that are defined in the correlations of decomposed behaviours, the use of second-stage log-likelihood LL_t^g for anomaly detection would be more appropriate since the second-stage model implicitly models the correlations among local atomic events and collectively learn the evidences obtained from all first-stage models. Abnormal correlation detection is thus achieved by computing LL_t^g according to (4) and comparing the obtained value against a pre-defined threshold Th . Specifically, if $LL_t^g < Th$, the unseen sequence is detected as anomalous.

6. Experiments

We first examine the effectiveness of the behaviour-based decomposition and the CasDBNs in detecting and discriminating different types of anomalies in indoor environments. The performance of the proposed framework is then evaluated on

detecting abnormal correlations among multiple objects in a busy traffic junction.

6.1. Discriminating different temporal causes of anomalies

Two experiments were conducted. In the first experiment, we studied the abnormality detection capability of CasDBNs. In the second experiment, we examined the performance of CasDBNs on discriminating different types of abnormal behaviours. The experimental results were then compared with those obtained using alternative models including a first-order HMM, an HHMM and a S-HSMM [10], which is an hierarchical extension of the Coxian HSMM [24].

Datasets: Two video datasets, collected in an office scenario and a café scenario, respectively, were employed in the experiments. The videos were recorded at 15 Hz and have a frame size of 320×240 pixels and 320×256 pixels, respectively. The office dataset consists of 60 sequences (32 998 frames) in total, including 25 normal behaviour sequences, 15 sequences containing atypically long duration and 20 sequences with atypical temporal order. Examples of the office sequences can be seen in Fig. 1. A total of 60 sequences (31 338 frames) were collected for the café dataset, including 30 normal behaviour sequences, 10 sequences containing atypically long duration, 10 sequences containing atypically short duration and 10 sequences with abnormal temporal order. The typical temporal order and duration (computed as the mean duration in the normal sequences) of the atomic actions involved in both datasets are given in Fig. 8.

Background subtraction was performed using adaptive Gaussian mixture background modelling [45]. Feature extraction proceeded by extracting the object centroid (\bar{x}, \bar{y}) , occupancy (R_f) ,

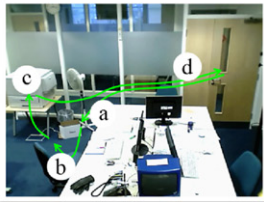
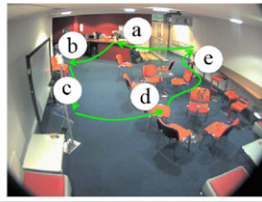
Office dataset	Café dataset
 <p>a. entering the office (6 sec) b. working at the desk (12 sec) c. printing (12 sec) d. leaving the office (6 sec)</p>	 <p>a. entering the café and getting a cup (6 sec) b. getting water at water dispenser (6 sec) c. choosing a magazine (6 sec) d. reading at the table (12 sec) e. leaving the café (6 sec)</p>

Fig. 8. Atomic actions in the datasets and the associated typical durations spent (in second). The trajectories illustrate the normal temporal order of atomic actions.

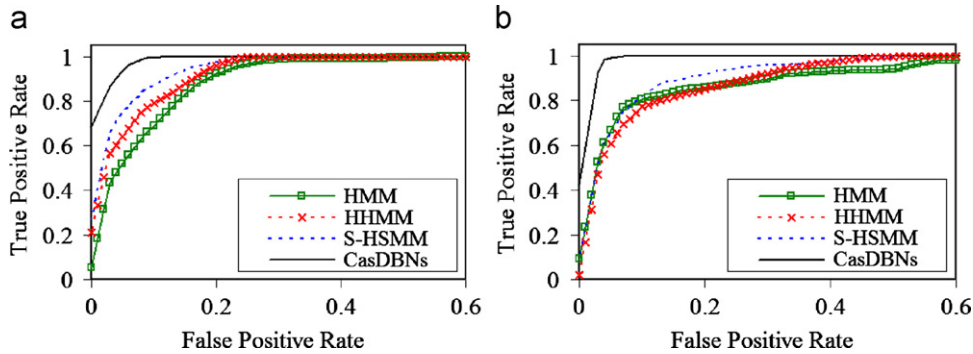


Fig. 9. Averaged ROC curves for the HMM, HHMM, S-HSMM and CasDBNs based on (a) office dataset, (b) café dataset.

ratio of the dimension (R_d), and ratio of the minor axis to the major axis (R_e), of an ellipse fitted to the blob. The features extracted were represented as a feature vector $\mathbf{f}_{blob} = [\bar{x}, \bar{y}, R_f, R_d, R_e]$ and used as the input to the first stage HHMM in the CasDBNs (see Fig. 4).

We applied random sample cross validation in this experiment. In particular, in each cross-validation run, we randomly selected 15 normal behaviour sequences from the office dataset and 20 normal sequences from the café dataset to train the CasDBNs, a first-order HMM, an HHMM and a S-HSMM. We varied the number of hidden states in each model and report the best results obtained from each model. Specifically, the number of hidden states in the first-order HMM was varied from 2 to 15. It turned out that an HMM with 12 states gave the best result on office dataset, whilst an HMM with 13 states yielded the best result on café dataset. For the standalone HHMM, S-HSMM and the first stage HHMM in CasDBNs, we set the number of parent states $|Q_1^p|$ at the top layer as equal to the number of atomic actions involved in the dataset. The number of children states $|Q_2^c|$ corresponding to each parent state was varied from 2 to 6. The optimal numbers of children states were 4 and 3 for the office dataset and the café dataset, respectively. For the second-stage MOHMM, the number of hidden states was set to the number of atomic actions involved.

An additional parameter to determine for S-HSMM is the number of sub-phases for modelling the behaviour duration. We tested the performance of S-HSMM with different number of sub-phases and found that 20 sub-phases for each children state yielded a reasonable balance between the accuracy and computational complexity. It is worth pointing out that the total number of sub-phases of a S-HSMM is enormously large even with a small number of parent states and children states. For instance, a S-HSMM with four parent states with three children states each would have 240 sub-phases when the number of sub-phases is set to 20 for each children state.

Anomaly detection: The first experiment was to test the performance of the proposed framework on anomaly detection. In each cross-validation run, normalised log-likelihood LL^{pint} was

computed and compared with varying threshold Th^{sum} . The receiver operating characteristic (ROC) curve averaged over 10 runs of the four models are shown in Fig. 9. The results show that the proposed framework outperforms the other three models. As expected, with the intrinsic hierarchical structure of the behaviours being explicitly modelled, HHMM outperforms HMM. On top of the hierarchical structure modelling, S-HSMM provides more accurate duration modelling via state expansion. Consequently, S-HSMM achieves better performance than HHMM. As pointed out in Section 4.3, our CasDBNs offer more accurate duration modelling at a much lower computational cost compared to an expanded state HMM such as S-HSMM (214 parameters compared to 799 for S-HSMM), thanks to the explicit duration modelling enabled by the cascade model structure. In particular, it is noted that our CasDBNs are more sensitive to durational anomalies, leading to the better detection performance.

Anomaly differentiation: The objective of the second experiment is to evaluate the capability of behaviour-based decomposition and CasDBNs in distinguishing different types of anomalies, i.e. anomalies in temporal order and duration. The datasets were divided into training set, validation set and testing set. The number of training sequences was the same as that in the previous experiment. We randomly picked 15 sequences from office data set and 20 sequences from café dataset as validation set to find the optimal values of thresholds Th^d and Th^o . The rest of the dataset were reserved for testing. The experimental results on the office dataset show that the CasDBNs are able to distinguish durational abnormality and temporal order abnormality at an accuracy rate of 82.40%, whilst an accuracy rate of 95.33% is obtained in the experiment on the café dataset.

Figs. 10 and 11 show the levels of sensitivity of HMM, S-HSMM, the first stage HHMM and the second stage MOHMM towards the two different types of anomalies. Each log-likelihood value in these plots corresponds to one behaviour sequence. From Figs. 10(a–c) and 11(a–c), it is evident that HMM, HHMM and S-HSMM are insensitive to abnormal temporal duration, whilst being sensitive to abnormal temporal orders. Importantly, when

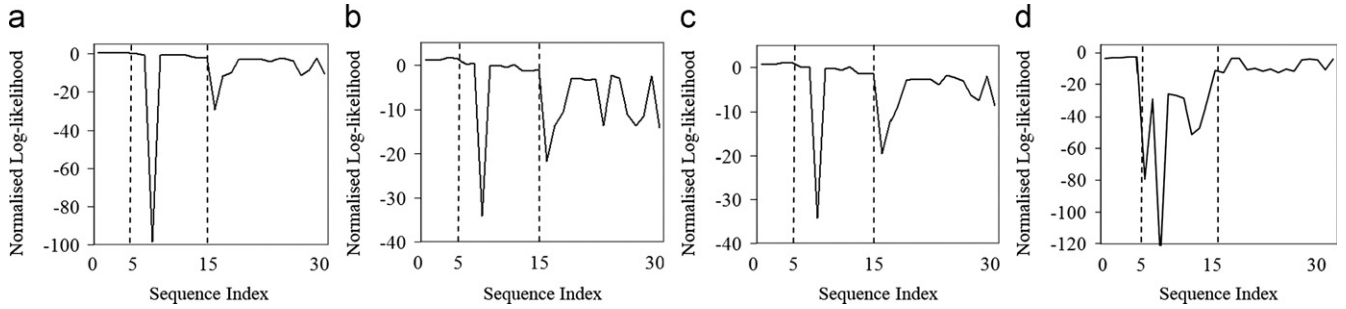


Fig. 10. Normalised log-likelihood plot (averaged over 10 runs) for (a) HMM, (b) S-HSMM, (c) CasDBNs Stage 1—HHMM and (d) CasDBNs Stage 2—MOHMM based on office dataset. Y-axis represents the normalised log-likelihood, while X-axis represents the index of test sequences. The first 5 sequences are normal sequences, 6–15 are sequences with atypical duration, and 16–30 are sequences with abnormal temporal order.

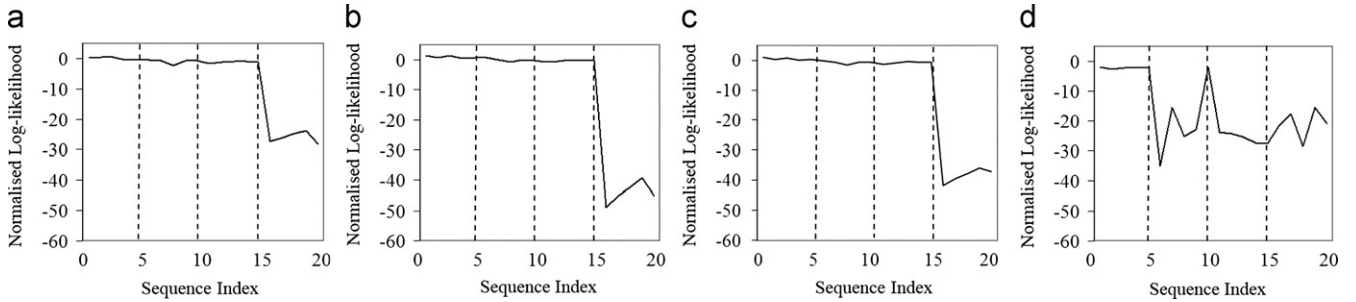


Fig. 11. Normalised log-likelihood plot (averaged over 10 runs) for (a) HMM, (b) S-HSMM, (c) CasDBNs Stage 1—HHMM and (d) CasDBNs Stage 2—MOHMM based on café dataset. The first 5 sequences are normal sequences, 6–10 are sequences with atypical long duration, 11–15 are sequences with atypical short duration, and 16–20 are sequences with abnormal temporal order.

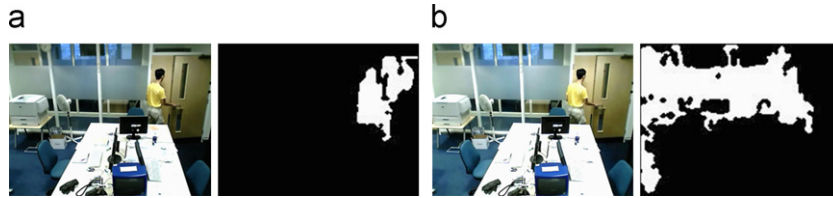


Fig. 12. The frames with the corresponding detected foreground blob. Imperfect blob detection caused by lighting change, which in turn causes the sequence being wrongly classified. (a) Frame 33, (b) Frame 34.

used alone, they cannot be used to distinguish the two types of anomalies. In comparison, the second stage MOHMM is sensitive to durational anomalies. Combined with the first stage HHMM, they act as filters that are selective to different anomaly types, thus offering a solution to the behaviour differentiation task.

Most of the misclassifications by our CasDBNs were caused by the noise in behaviour representation due to the changing lighting conditions. The adaptive background subtraction method was configured to have a slow adaptation rate; it thus could not adapt to the sudden change of illumination. An example of such errors is depicted in Fig. 12, from which we can observe a drastic lighting change between two consecutive frames. This erroneous feature input triggered both the first stage HHMM and the second stage MOHMM to believe that the current atomic action has finished and another atomic action out of order has begun. Consequently, this sequence, correctly detected as an anomaly, was wrongly classified as an anomaly with abnormal temporal order following the decision rules set out in Table 1.

6.2. Detecting abnormal correlations

Dataset: The road traffic video footage used in this experiment was recorded at 25 Hz and resized to a resolution of 360×288

pixels. The total duration of the recording is approximately 25 min, showing a busy road junction regulated by traffic lights, dominated by four types of traffic flows as illustrated in Fig. 13. Specifically, Flow A corresponds to traffic in vertical directions. Flow B, C and D are regarded as traffic flows in horizontal directions. In particular, Flow B represents left-turning and right-turning traffics by vehicles from vertical directions. Flow C corresponds to rightward traffic and Flow D corresponds to leftward traffic. The order of the traffic flow depends upon how busy the vertical traffics are. During most of the recording, the scene was extremely crowded. Consequently, Flow B can only take place after Flow A finishes and is followed by C and D, (i.e. the typical temporal order is A, B, C, D). However, it is noted that very occasionally, there is a gap in Flow A which is big enough for Flow B to take place until the gap closes. In other words, Flow A and B can occur alternatively during the vertical traffic phase. This makes global behaviour modelling and anomaly detection challenging in this scene as vehicles behaviours and the correlations among them are determined by not only the traffic light cycles, but also the traffic volume as well as the driving habits and reactions of the drivers.

A total of 123 non-overlapping clips were segmented from the video. In particular, 73 clips (21900 frames) were used for training, whilst 50 clips (15000 frames) were reserved for testing.



Fig. 13. Traffic flows observed in the dataset. (a) Flow A, (b) Flow B, (c) Flow C, (d) Flow D.

Table 2
Ground truth of the traffic dataset.

Category	Description	Clip no.
A	Anomalies that are visually obvious	3, 4
B	Rare and ambiguous behaviours	25, 35
C	Anomalies supported by weak evidence	10, 41

Scene segmentation was applied on the dataset resulting in six semantic regions (see Fig. 6(b)). A total of 30 local atomic events were automatically discovered in the six regions.

Ground truth: Prior to the evaluation of the proposed method, ground truth was first obtained by performing exhaustive frame-wise examination on both the training and testing set. Consequently, six out of 50 testing clips were labelled as anomaly. They are summarised in Table 2.

Following the definitions given in Section 5.3, clips 3 and 4 were categorised in Category-A since they contain anomalies featured with abnormal correlations that are visually obvious. In particular, clip 3 captures the event where all the vehicles in vertical traffic flow stopped moving either because the drivers heard the siren and/or saw the fire engine approaching the junction from the left entrance (see Fig. 14(a)). In clip 4, the fire engine entered the junction and caused interruptions to the vertical traffic at both directions (see Fig. 2(a)). Clips 25 and 35 correspond to abnormal traffic flows where vehicles did not follow the typical temporal order of A–B–C–D. In particular, clip 25 shows a motorbike making a left turn during the vertical traffic flow A. In clip 35, vehicles were using a gap in the middle of traffic flow A to make right turn and left turn at the same time interval. Both clips were grouped into Category-B because they belong to rare/unusual behaviour with low frequency of occurrence in the training set (out of 73 training clips, only three clips correspond to left-turn and two clips correspond to turning both ways). Clip 10 (see Fig. 14(c) and also Fig. 2(b)) shows a white van running the red light from the left to right horizontally (see the caption of Fig. 2 for details), and Clip 41 is featured with a car jumping the red light in the vertical direction. Both clips were labelled as anomalies belonging to Category-C, which are undetectable even by human without comprehensive examination of the traffic cycle duration over time.

Model construction: For CasDBNs, the number of hidden states in the second-stage MOHMM was set to 2 with each of them representing the horizontal and vertical traffic flows, respectively. To obtain the optimal number of hidden states in each first-stage model, we varied the number of states from 2 to 10, and observed the matching accuracies of the global traffic phases (corresponding to the red–green traffic light phases) inferred using the training data in each setting with the ground truth global phases. The first-stage model with five hidden states yielded the best accuracies. It is observed that different states of a model correspond to different stages of a regional behaviour (e.g. vehicles waiting in the region, vehicles start moving).

We employed a MOHMM, a CHMM, a PaHMM, and an hierarchical MOHMM (HMOHMM) as baseline methods in this experiment. The number of hidden states of MOHMM was set to 2 since the global traffic flows have two phases (vertical and horizontal). We let the MOHMM learned directly from the observation space and without behaviour-based decomposition. Thus, each hidden state of the model consisted of 30 observation nodes, each of which corresponds to one class of atomic events detected. Note that it is possible for a MOHMM to learn from the regional information, i.e. to let each observational node encode all possible combinations of regional events. However, it was found from our experiment (not reported here) that the false detections obtained were unacceptably high due to the sparse observational vectors. We implemented a CHMM with six coupled chains with each of them correspond to one segmented region. Following the same setting as in the first-stage MOHMM models of CasDBNs, each chain in CHMM had five hidden states. The number of observation nodes per hidden state in each chain was equal to the number of atomic events detected in respective region. The PaHMM had a similar setting as the CHMM but with the chains decoupled. As for the HMOHMM, there are multiple MOHMMs, of which their hidden states were conditionally dependent on another common hidden variable. Together they formed two hidden layers that had the same definition as respective hidden layer of the two stages in the CasDBNs. The key difference between the HMOHMM and the CasDBNs is that the former has additional dependencies between the two hidden layers, whilst the two stages of CasDBNs were coupled using the inferential outputs of the first stage. Note that the CHMM, PaHMM and HMOHMM took advantage of our behaviour-based decomposition to reduce the computational cost. The key difference against our CasDBNs is therefore on the model structure.

Experimental results: The normalised log-likelihoods for each test clip was computed and compared against a threshold Th for anomaly detection. Th was varied to obtain the ROC curves of the models, as shown in Fig. 15. As can be seen from the ROC curves, the detection result of CasDBNs is significantly better than those obtained using the baseline methods.

To gain some insights into the causes of the misdetections and false alarms, the normalised log-likelihoods obtained using the five models are plotted in Fig. 16. As can be observed, all models except the CasDBNs suffered from high false alarm rate. In particular, although the CHMM was able to detect anomalies that are supported by strong visual evidences (i.e., clips 3 and 4), it missed other anomalies that are either ambiguous (Category-B) or those that are supported by weak visual evidences (Category-C). The poorest result is observed in the HMOHMM. As compared to the CasDBNs, the structure of the HMOHMM is inevitably more complex due to the additional dependencies between the first stage and the second stage. As a result, the model is not able to learn the ‘true’ dependencies given limited amount of training data, leading to poor result. In comparison, Fig. 16 suggests that our CasDBNs are more selective. In particular, anomalies belonging to Categories A and B (Clips 3, 25, and 35) can be easily detected as abnormal behaviours with only 1 false positive.



Fig. 14. Example clips for abnormal behaviours that are visually obvious featured with apparent abnormal correlations (Category-A), abnormal behaviours that are ambiguous (Category-B), and clips that contain a subtle anomaly that is almost undetectable by human (Category-C). The corresponding objects that caused the anomalies are highlighted using bounding boxes. (a) Clip 3 (Category-A), (b) Clip 35 (Category-B), (c) Clip 10 (Category-C), (d) Clip 41 (Category-C).

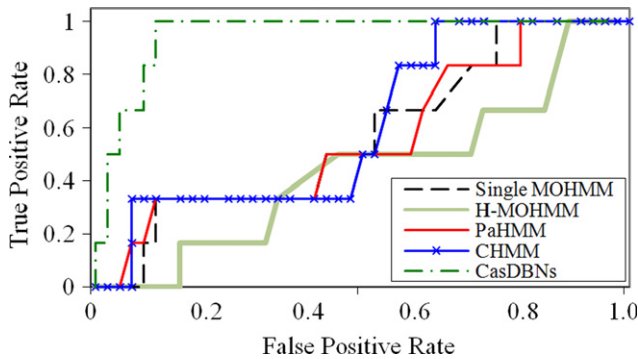


Fig. 15. The receiver operating characteristic (ROC) curves. The area under ROC (AUROC) achieved by using the single MOHMM, PaHMM, CHMM and HMOHMM was 0.5720, 0.5644, 0.6080, and 0.4224, respectively, compared with 0.9280 obtained by using the proposed framework.

More importantly, the model correctly detected anomalies supported by weak visual evidence, namely clip 10 at the cost of 4 false positives, and clip 41 at the cost of 2 false positives. In the analysis of clip 41, we found that the events triggered in regions 2 and 5, i.e. the route taken by the car, were not correlated with events occurrence in other regions. The conflict of local events occurrence was successfully captured by CasDBNs in the first-stage inferential outputs, and was collectively passed to the second-stage model for global anomaly detection. The second-stage model of CasDBNs was then able to detect that the inferred states from stage-one models were out of the normal temporal order, resulting in a low log-likelihood for clip 41. The capability of combining the visual evidences from local regions for global anomaly detection is *unique* to CasDBNs and explains its superior ability to detect anomalies in Category-C. For instance, to detect

clip 41, CHMM, PaHMM, MOHMM, and HMOHMM recorded more than 15 false positives.

The high false alarm rate (see Fig. 15) observed in the baseline methods was mainly caused by the fact that they are susceptible to noise, since all models learn directly from the observation space without any mechanism to prevent error propagation. An example can be seen in Fig. 17, which shows that in clip 13, vehicles in region four were mistakenly grouped as a large blob with vehicles in region 1 (highlighted using the black bounding box) causing errors in event occurrence in region 4, which consequently led to a false alarm using these models. In contrast, the CasDBNs were able to cope with this error by estimating the most probable state and the influence of the error was further reduced by collectively considering all the evidences from different segmented regions. As can be seen from Fig. 16(e), clips 13 yielded much higher log-likelihood value using the CasDBNs, indicating the model is able to cope with the erroneous input effectively.

Our experiments on multi-object correlation anomaly detection demonstrate that the CasDBNs are more robust to noise and errors exhibited in behaviour representation than the alternative models. Given erroneous features caused by shadows, occlusions, and changing lighting condition, conventional DBN models are unable to filter out these errors in the observation space or prevent them from propagating to the state space. As a result, these models suffer from the problem of high false alarm rate. Our results also suggest that even with behaviour-based decomposition, alternative models for multi-object correlation modelling (e.g. CHMM and PaHMM) are unable to accurately capture the temporal dynamics of the causal relationships between objects. They thus failed to detect subtle behavioural anomalies supported by weak visual evidence such as Clips 10 and 41 in Fig. 14. On the contrary, the first stage in the CasDBNs is connected to the second

stage via its inferential outputs, which can minimise the propagation of noise from one stage to the next. Therefore, the model can work well given noisy data.

Object-based decomposition vs. behaviour-based decomposition: An experiment was carried out on the traffic dataset to highlight the inadequacy of object-based decomposition and advantage of behaviour-based decomposition. Fig. 18(a) shows the trajectories extracted from a 2-minute video clip using a Kalman filter based tracker. In Fig. 18(b), we plot the duration of each of the extracted trajectories (331 in total), and compare with the ground truth trajectories (114 in total) which were obtained by manual labelling. It is evident that the large amount of broken trajectories makes an object-based decomposition unsuitable for anomaly detection.

6.3. Computational cost

The CasDBNs need less parameter and is thus computationally more tractable than alternative DBN models such as CHMM when dealing with multiple temporal processes. Consequently the

CasDBNs can be readily applied to a wide area busy scene with complex spatial and temporal visual context. This is mainly due to the decomposition of behaviour based on visual context, and the effective modelling of global correlations using the two-stage structure. In particular, referring to the model structure given in Fig. 7, the time complexity for a standard MOHMM is $O(TN^2)$, where T is the number of time slices and N is the number of hidden states. Therefore the proposed framework exhibits $O((R+1)TN^2)$ complexity assuming that the second-stage model and all first-stage models have the same number of hidden states. In other words, the complexity of our framework scales linearly to the number of decomposed behaviours, which is much lower than conventional DBN models used for multi-object correlation modelling such as CHMM. For instance, the complexity of CHMM is exponential to the number of decomposed behaviours, given as $O(TN^{2C})$, where C is the number of coupled temporal processes corresponding the number of objects if a tracking-based representation is used [6] or the number of event classes if event based representation is deployed [11].

The training and inference (on the full training/testing sets) time needed by the models in detecting abnormal correlations

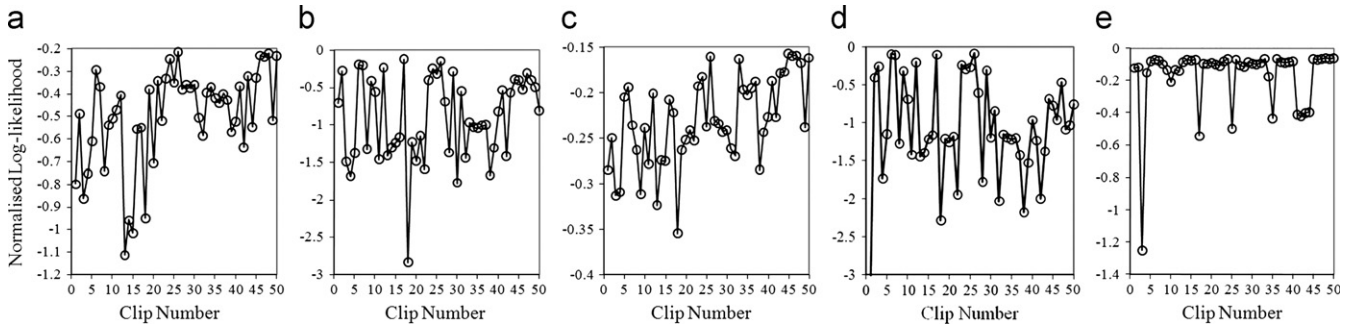


Fig. 16. The normalised log-likelihood plots. (a) Single MOHMM, (b) PaHMM, (c) CHMM, (d) HMOHMM, (e) CasDBNs.

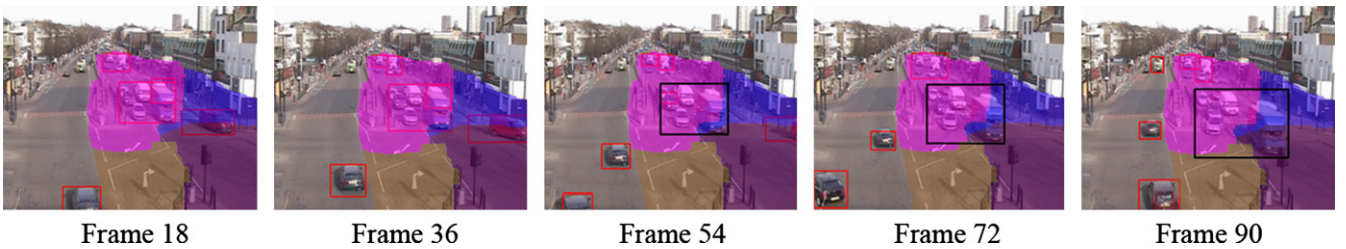


Fig. 17. An example of imperfect blob detection in clip 13 which result in local atomic events being grouped into wrong clusters. The blob is marked with a black colour bounding box.

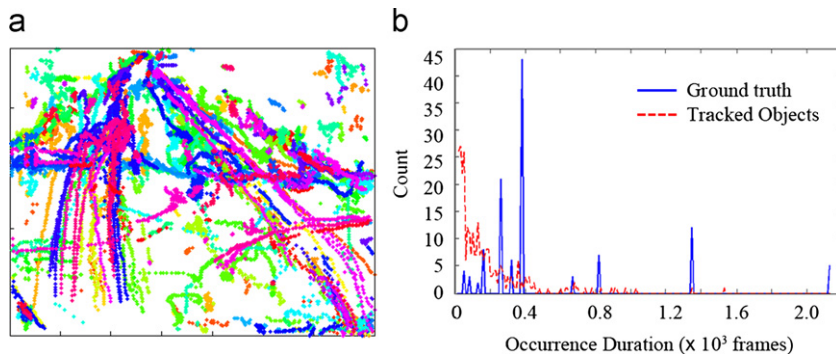


Fig. 18. Taking an object-based decomposition approach, objects are tracked in a busy traffic scene resulting in large number of broken trajectories. (a) Trajectories, (b) histogram of duration.

Table 3

The training and inference time (on the full training/testing set) averaged over 10 runs along with the standard deviation of the MOHMM, PaHMM, CHMM, HMOHMM, and CasDBNs in Matlab implementation.

Model	Training time (s)	Inference time (s)
Single MOHMM	114.5500 \pm 0.5725	2.2824 \pm 0.0009
PaHMM	404.3969 \pm 7.6180	11.1020 \pm 0.3310
CHMM	1052.8067 \pm 15.1522	30.0124 \pm 0.5455
HMOHMM	448.3300 \pm 9.5261	19.5817 \pm 0.4374
CasDBNs	160.7282 \pm 0.5488	5.8180 \pm 0.1819

(Section 6.2) are summarised in Table 3. The platform employed in the experiments has a dual-Core 3 GHz processor with 4 GB of RAM.

6.4. Limitations and possible extensions

The proposed framework has several limitations:

- (i) Even though individual stages in CasDBNs are generative, the CasDBNs itself cannot be used in a fully generative fashion due to the lack of dependencies between stages. Nevertheless, the primary objective of using CasDBNs is for detecting and discriminating anomalies, rather than generating new activity sequences. Importantly, compromising the generative capability offers computational gain that is critical for real-time detection, robustness to noise and discrimination of video anomalies.
- (ii) The proposed model can only perform atypical duration detection after a full sequence of an atomic action is automatically segmented. This is a trade-off between the per time frame detection given partial observation and the explicit modelling of the whole duration. Further exploration is needed to overcome this shortcoming.
- (iii) The first stage of CasDBNs is still sensitive to noise in the task of discriminating different types of anomalies. This is because we have to use the output from both stages to make the decision on the anomaly type, with the first-stage model still under the direct influence of the erroneous behaviour representation. To mitigate this problem, apart from improving the image pre-processing, one can monitor the reliability/confidence value of low-level features and stop the anomaly detection process temporarily when low feature reliability value (which is possibly caused by noise) is observed.

In this study, we choose MPE to form the intermediate observation for the subsequent DBN stage, due to its good trade-off between simplicity and effectiveness. Nonetheless, alternative feature extraction methods exist and need to be investigated. Although only a two-stage CasDBNs are employed in this study, the proposed framework can be generalised to have more stages to model more complex behaviours (e.g., more types of anomalies, more complex visual contexts). One of the ongoing work is to extend the framework for the detection of abnormal co-occurrence of events captured by a network of cameras, for which CasDBNs of more than two stages are required. A possible implementation is to add an additional stage on top of the second-stage model to learn the temporal dependency and co-occurrence of behaviours captured by a small local camera network. The hidden states of the third stage then correspond to the phases of global multi-camera behaviour that can only be interpreted across different camera views. Beyond that, the fourth stage will be required when behaviours are modelled across different local camera networks. When we extend the framework to more stages, we would foresee longer training time and

inference time. However, these problems can be effectively solved by using efficient approximate inference algorithms [46].

7. Conclusions

We have presented a framework for detecting anomalies exhibited in complex behaviours which are more subtle and difficult to detect owing to the complex temporal characteristics and correlation among multiple objects' behaviours. In contrast to conventional methods that perform object-based decomposition and employ standalone models for complex behaviour modelling, we have proposed to decompose complex behaviour in accordance with different temporal characteristics and visual contexts and model the decomposed behaviours with a cascade of DBNs. The experimental results have shown that the proposed framework has a unique capability of abnormality differentiation, lacking from the existing techniques. In addition, while alternative methods fail to detect durational abnormality accurately, the framework is sensitive to abnormal duration in complex behaviours. In multiple object scenarios, we have demonstrated that framework's capability in coping with the noise and errors in behaviour representation. More importantly, it has shown superior performance in detecting subtle anomalies that are ambiguous or difficult to detect when objects are viewed in isolation. Both behaviour-based decomposition and cascaded structure are crucial for achieving the superior performance. Without the behaviour-based decomposition, the correlation cannot be modelled effectively (and will be computationally less tractable). On the other hand, without the cascaded structure, the model would be susceptible to noise presented in the low-level feature space.

References

- [1] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, *IEEE Transactions on Systems, Man, and Cybernetics* 34 (3) (2004) 334–352.
- [2] H.M. Dee, S.A. Velastin, How close are we to solving the problem of automated visual surveillance? *Machine Vision and Applications* 19 (5–6) (2008) 329–343.
- [3] T.B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, *Computer Vision and Image Understanding* 104 (2) (2006) 90–126.
- [4] N. Johnson, D.C. Hogg, Learning the distribution of object trajectories for event recognition, *Image and Vision Computing* 14 (8) (1996) 609–615.
- [5] R. Fraile, S.J. Maybank, Vehicle trajectory approximation and classification, in: *British Machine Vision Conference*, Southampton, UK, 1998, pp. 832–840.
- [6] M. Brand, N. Oliver, A. Pentland, Coupled hidden Markov models for complex action recognition, in: *IEEE International Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997, pp. 994–999.
- [7] C. Vogler, D. Metaxas, A framework for recognizing the simultaneous aspects of American sign language, *Computer Vision and Image Understanding* 81 (3) (2001) 358–384.
- [8] J. Li, S. Gong, T. Xiang, Scene segmentation for behaviour correlation, in: *European Conference on Computer Vision*, Marseille, France, 2008, pp. 383–395.
- [9] S. Gong, T. Xiang, Recognition of group activities using dynamic probabilistic networks, in: *IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 742–749.
- [10] T. Duong, H. Bui, D. Phung, S. Venkatesh, Activity recognition and abnormality detection with the switching hidden semi-Markov model, in: *IEEE International Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 838–845.
- [11] T. Xiang, S. Gong, Beyond tracking: modelling activity and understanding behaviour, *International Journal of Computer Vision* 67 (1) (2006) 21–51.
- [12] Y. Du, F. Chen, W. Xu, Y. Li, Recognizing interaction activities using dynamic Bayesian network, in: *International Conference on Pattern Recognition*, Hong Kong, China, 2006, pp. 618–621.
- [13] T. Xiang, S. Gong, Video behaviour profiling for anomaly detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30 (5) (2008) 893–908.
- [14] Y. Shi, A. Bobick, I. Essa, Learning temporal sequence model from partially labeled data, in: *IEEE International Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, 2006, pp. 1631–1638.

- [15] M. Perše, M. Kristan, J. Perš, G. Mušič, G. Vučkovič, S. Kovačič, Analysis of multi-agent activity using petri nets, *Pattern Recognition* 43 (4) (2010) 1491–1501.
- [16] M. Brand, Understanding manipulation in video, in: *International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, USA, 1996, pp. 94–99.
- [17] Y.A. Ivanov, A.F. Bobick, Recognition of visual activities and interactions by stochastic parsing, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (8) (2000) 852–872.
- [18] G. Medioni, I. Cohen, F. Brémont, S. Hongeng, R. Nevatia, Event detection and analysis from video streams, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (8) (2001) 873–889.
- [19] L. Wang, W. Hu, T. Tan, Recent developments in human motion analysis, *Pattern Recognition* 36 (3) (2003) 585–601.
- [20] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities—a survey, *IEEE Transactions on Circuits and Systems for Video Technology* 18 (11) (2008) 1473–1488.
- [21] T. Xiang, S. Gong, Activity based surveillance video content modelling, *Pattern Recognition* 41 (7) (2008) 2309–2326.
- [22] N.T. Nguyen, D.Q. Phung, S. Venkatesh, H.H. Bui, Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model, in: *IEEE International Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 955–960.
- [23] S. Lühr, S. Venkatesh, G.W. West, H.H. Bui, Explicit state duration HMM for abnormality detection in sequences of human activity, in: *Pacific Rim International Conference on Artificial Intelligence*, Auckland, New Zealand, 2004, pp. 983–984.
- [24] T. Duong, D. Phung, H. Bui, S. Venkatesh, Efficient coxian duration modelling for activity recognition in smart environments with the hidden semi-Markov model, in: *International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, Melbourne, Australia, 2005, pp. 277–282.
- [25] M.J. Russell, R.K. Moore, Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition, in: *IEEE International Conference on Acoustics Speech and Signal Processing*, Tampa, Florida, USA, 1985, pp. 5–8.
- [26] S.E. Levinson, Continuously variable duration hidden Markov models for automatic speech recognition, *Computer Speech and Language* 1 (1) (1986) 29–45.
- [27] S. Hongeng, R. Nevatia, Large-scale event detection using semi-hidden Markov models, in: *IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 1455–1462.
- [28] M.J. Russell, A.E. Cook, Experimental evaluation of duration modelling techniques for automatic speech recognition, in: *IEEE International Conference on Acoustics Speech and Signal Processing*, Dallas, Texas, USA, 1987, pp. 2376–2379.
- [29] Y. Du, F. Chen, W. Xu, Human interaction representation and recognition through motion decomposition, *IEEE Signal Processing Letters* 14 (12) (2007) 952–955.
- [30] N. Oliver, B. Rosario, A. Pentland, A Bayesian computer vision system for modeling human interactions, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (8) (2000) 831–843.
- [31] N. Oliver, E. Horvitz, A. Garg, Layered representations for human activity recognition, in: *IEEE International Conference of Multimodal Interfaces*, Pittsburgh, Pennsylvania, USA, 2002, pp. 3–8.
- [32] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, G. Lathoud, Modeling individual and group actions in meetings with layered HMMs, *IEEE Transactions on Multimedia* 8 (3) (2004) 509–520.
- [33] J. Li, S. Gong, T. Xiang, Global behaviour inference using probabilistic latent semantic analysis, in: *British Machine Vision Conference*, Leeds, UK, 2008, pp. 193–202.
- [34] C.C. Loy, T. Xiang, S. Gong, From local temporal correlation to global anomaly detection, in: *International Workshop on Machine Learning for Vision-based Motion Analysis (European Conference on Computer Vision)*, Marseille, France, 2008.
- [35] K.P. Murphy, Dynamic Bayesian networks: representation, inference and learning, Ph.D. Thesis, University of California at Berkeley, Computer Science Division, 2002.
- [36] L.E. Baum, T. Petrie, G. Soules, N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *The Annals of Mathematical Statistics* 41 (1) (1970) 164–171.
- [37] G.D. Forney, The Viterbi algorithm, *Proceedings of the IEEE* 61 (1973) 268–278.
- [38] Y. Wang, X. Hou, T. Tan, Recognize multi-people interaction activity by PCA-HMMs, in: *Asian Conference on Computer Vision*, Hyderabad, India, 2006, pp. 160–169.
- [39] H. Hermansky, D.P.W. Ellis, S. Sharma, Tandem connectionist feature extraction for conventional HMM systems, in: *International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000, pp. 1635–1638.
- [40] K.P. Murphy, M.A. Paskin, Linear-time inference in hierarchical HMMs, in: *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, 2001.
- [41] S. Lühr, H.H. Bui, S. Venkatesh, G.W. West, Recognition of human activity through hierarchical stochastic learning, in: *International Conference on Pervasive Computing and Communication*, Fort Worth, Texas, USA, 2003, pp. 416–422.
- [42] S. Fine, Y. Singer, N. Tishby, The hierarchical hidden Markov model: analysis and applications, *Machine Learning* 32 (1) (1998) 41–62.
- [43] H. Zhong, J. Shi, M. Visontai, Detecting unusual activity in video, in: *IEEE International Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2004, pp. 819–826.
- [44] S. F. Kimball, J. Como, Cascaded hidden Markov model for meta-state estimation, US Patent Number: 6963835, 2005.
- [45] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, in: *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, Ft. Collins, CO, USA, 1999, pp. 246–252.
- [46] K.P. Murphy, Y. Weiss, The factored frontier algorithm for approximate inference in DBNs, in: *Uncertainty in AI*, Seattle, Washington, USA, 2001, pp. 378–385.

Chen Change Loy received the B.Eng degree in Electronics Engineering from University of Science, Malaysia in 2005. He is now a Ph.D candidate (Supervisors: Tao Xiang and Shaogang Gong) in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His current research interests include computer vision and machine learning, with focus on activity analysis and abnormal behaviour recognition in surveillance video.

Tao Xiang received the Ph.D degree in electrical and computer engineering from the National University of Singapore in 2002. He is a currently a lecturer in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision, statistical learning, video processing, and machine learning, with focus on interpreting and understanding human behaviour.

Shaogang Gong is Professor of Visual Computation at Queen Mary University of London, a Fellow of the Institution of Electrical Engineers and a Member of the UK Computing Research Committee. He received his D.Phil in computer vision from Keble College, Oxford University in 1989. He has published over 200 papers in computer vision and machine learning, and a book on *Dynamic Vision: From Images to Face Recognition*. His work focuses on motion and video analysis; object detection, tracking and recognition; face and expression recognition; gesture and action recognition; visual behaviour profiling and recognition.