Video Behaviour Profiling and Abnormality Detection without Manual Labelling

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

Abstract

A novel framework is developed for automatic behaviour profiling and abnormality sampling/detection without any manual labelling of the training dataset. Natural grouping of behaviour patterns is discovered through unsupervised model selection and feature selection on the eigenvectors of a normalised affinity matrix. Our experiments demonstrate that a behaviour model trained using an unlabelled dataset is superior to those trained using the same but labelled dataset in detecting abnormality from an unseen video.

1. Introduction

Given 24/7 continuously recorded video or online CCTV input, the goal of automatic behaviour profiling is to learn a model that is capable of detecting unseen abnormal behaviour patterns whilst recognising novel instances of expected normal behaviour patterns. In this context, we define abnormality as atypical behaviour patterns that are not represented by sufficient samples in a training dataset but critically they satisfy the specificity constraint to abnormal patterns. This is because one of the main challenges for the model is to differentiate abnormality from outliers caused by noisy visual features used for behaviour representation. The effectiveness of a behaviour profiling algorithm shall be measured by (1) how well abnormality can be detected (i.e. measuring specificity to expected patterns of behaviour) and (2) how accurately and robustly different classes of normal behaviour patterns can be recognised (i.e. maximising between-class discrimination).

We develop a novel framework for automatic behaviour profiling and abnormality sampling/detection without any manual labelling of the training dataset. On behaviour profiling and learning, our approach is significantly different from most existing approaches which rely upon labelled datasets for model training [8, 7, 5]. On abnormality sampling and detection, there has been very little reported work in the literature. Recently, an intentional, goal-based behaviour modelling approach was proposed in [1] for detecting unusual behaviours. This approach relies on hard-wired rules established with human invention. It is thus difficult to implement for an unconstrained, complex scenario. Zhong et. al [17] proposed an unsupervised method based on co-embedding the prototype image features for classifying a group of behaviour patterns as normal or abnormal. Abnormal behaviour must be included in the training datasets for model learning. This is not required by our approach where a behaviour model automatically detects unseen normal and abnormal behaviours.

There are three key motivations for behaviour profiling using unlabelled data: (1) Manual labelling of behaviour patterns is laborious and often rendered impractical. (2) More critically though, manual labelling of behaviour patterns could be inconsistent and error prone. This is because a human tends to interpret behaviours based on a priori cognitive knowledge of what should be present in a scene rather than solely based on what is visually detectable in the scene. This introduces bias due to differences in experience and mental states. (3) A model trained based on manual labelling may have an advantage in explaining data that are well-defined. However, training using labelled data does not necessarily help a model with identifying novel instances of atypical behaviour patterns as the model tends to be brittle and less robust in dealing with instances of behaviour that are not clear-cut in an open-world scenario (i.e. the number of expected normal and abnormal behaviours cannot be pre-defined exhaustively).

Due to the space-time nature of behaviour patterns and their variable duration, we need to develop a compact and effective feature representation scheme and to deal with time-warping. We adopt a discrete scene event based image feature extraction approach [5]. This is different from most previous approaches such as [12, 8, 7, 1] where image features are extracted based on object tracking. A discrete event based behaviour representation aims to avoid the difficulties associated with tracking under occlusion in noisy scenes [5]. Each behaviour pattern is modelled using a Dynamic Bayesian Network [4] which provides a suitable means for time-warping and measuring the affinity among behaviour patterns.

The natural grouping of training behaviour patterns can be automatically discovered using the eigenvectors of the normalised affinity matrix [11]. A number of affinity matrix based clustering techniques have been proposed recently [13, 11, 16]. However, these approaches require known number of clusters. Given an unlabelled dataset, the number of behaviour classes are unknown in our case. To automatically determine the number of clusters, we propose to first perform unsupervised feature selection to eliminate those eigenvectors that are irrelevant/redundant in behaviour pattern grouping. A novel feature selection algorithm is derived which makes use of the a priori knowledge on the relevance of each eigenvector. Our algorithm differs from the existing techniques such as [6, 3] in that it is very simple and robust and thus able to work more effectively even with very sparse and noisy datasets.

2. Behaviour Pattern Representation

2.1. Video Segmentation

The goal is to automatically segment a continuous video sequence $\mathbf{V}$ into N video segments $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n, \ldots, \mathbf{v}_N\}$ such that ideally each segment contains a single behaviour pattern. The nth video segment $\mathbf{v}_n$ consists of T_n image frames represented as $\mathbf{v}_n = \{\mathbf{I}_{n1}, \ldots, \mathbf{I}_{nt}, \ldots, \mathbf{I}_{nT_n}\}$ where $\mathbf{I}_{nt}$ is the tth image frame. Depending on the nature of the video sequence to be processed, various segmentation approaches can be adopted. Since we are focusing on surveillance video, the most commonly used shot change detection based segmentation approach is not appropriate. In a not-too-busy scenario, there are often non-activity gaps between two consecutive behaviour patterns which can be utilised for activity segmentation. In the case where obvious non-activity gaps are not available, an on-line segmentation algorithm proposed in [14] can be adopted. Alternatively, the video can be simply sliced into overlapping segments with a fixed time duration [17].

2.2. Event-based Behaviour Representation

Firstly, an adaptive Gaussian mixture background model [12] is adopted to detect foreground pixels which are modelled using Pixel Change History (PCH) [15]. Secondly, the foreground pixels in a vicinity are grouped into a blob using the connected component method. Each blob with its average PCH value greater than a threshold is then defined as a scene-event. A detected scene-event is represented as a 7-dimensional feature vector $\mathbf{f} = \{\bar{x}, \bar{y}, w, h, R_f, M_p x, M_p y\}$ where $(\bar{x}, \bar{y})$ is the centroid of the blob, (w, h) is the blob dimension, R_f is the filling ratio of foreground pixels within the bounding box associated with the blob, and $(M_p x, M_p y)$ are a pair of first order moments of the blob represented by PCH. Among these features, $(\bar{x}, \bar{y})$ are location features, (w, h) and R_f are principally shape features but also contain some indirect motion information, and $(M_p x, M_p y)$ are motion features capturing the direction of object motion.

Thirdly, classification is performed in the 7D scene-event feature space using a Gaussian Mixture Model (GMM). The number of scene-event classes K_e captured in the videos is determined by automatic model order selection based on Bayesian Information Criterion (BIC) [10]. The learned GMM is used to classify each detected event into one of the K_e event classes. Finally, the behaviour pattern captured by the nth video segment $\mathbf{v}_n$ is represented as a feature vector $\mathbf{P}_n$, given as

$$\mathbf{P}_n = \{\mathbf{p}_{n1}, \ldots, \mathbf{p}_{nt}, \ldots, \mathbf{p}_{nT_n}\}, \quad (1)$$

where the t-th element $\mathbf{p}_{nt}$ is a K_e dimensional variable: $\mathbf{p}_{nt} = \{p_{nt}^1, ..., p_{nt}^k, ..., p_{nt}^{K_e}\}$. $\mathbf{p}_{nt}$ corresponds to the tth image frame of $\mathbf{v}_n$ where p_{nt}^k is the posterior probability that an event of the kth event class has occurred in the frame given the learned GMM.

3. Behaviour Profiling

3.1. Affinity Matrix

The behaviour profiling problem can now be defined formally. Consider a training dataset $\mathbf{D}$ consisting of N feature vectors $\mathbf{D} = \{\mathbf{P}_1, \ldots, \mathbf{P}_n, \ldots, \mathbf{P}_N\}$ where $\mathbf{P}_n$ is defined in Eqn. (1) representing the behaviour pattern captured by the nth video segment $\mathbf{v}_n$. The problem to be addressed is to discover the natural grouping of the training behaviour patterns upon which a model for normal behaviours can be built. This is essentially a data clustering problem with the number of clusters unknown. There are two aspects that make this problem challenging: (1) Each feature vector can be of different length. Conventional clustering approaches such as K-means and mixture models require that each data sample is represented as a fixed length feature vector. These approaches thus cannot be applied directly. (2) A definition of a distance/affinity metric among these variable length feature vectors is nontrivial.

Dynamic Bayesian Networks (DBNs) provide a solution for overcoming the above-mentioned difficulties. More specifically, each behaviour pattern in the training set is modelled using a DBN. To measure the affinity between two behaviour patterns represented as $\mathbf{P}_i$ and $\mathbf{P}_j$, two DBNs denoted as $\mathbf{B}_i$ and $\mathbf{B}_j$ are trained on $\mathbf{P}_i$ and $\mathbf{P}_j$ respectively using the EM algorithm [2, 4]. The affinity between $\mathbf{P}_i$ and $\mathbf{P}_j$ is then computed as:

$$S_{ij} = \frac{1}{2}\left\{\frac{1}{T_j}\log P(\mathbf{P}_j|\mathbf{B}_i) + \frac{1}{T_i}\log P(\mathbf{P}_i|\mathbf{B}_j)\right\}, \quad (2)$$

where $P(\mathbf{P}_j|\mathbf{B}_i)$ is the likelihood of observing $\mathbf{P}_j$ given $\mathbf{B}_i$, and T_i and T_j are the lengths of $\mathbf{P}_i$ and $\mathbf{P}_j$ respectively. DBNs of different topologies can be used for modelling each behaviour pattern. In this paper, we employ a

Multi-Observation Hidden Markov Model (MOHMM) [5] shown in Fig. 1. The number of hidden states for each hidden variables in the MOHMM is set as K_e, i.e., the number of event classes.

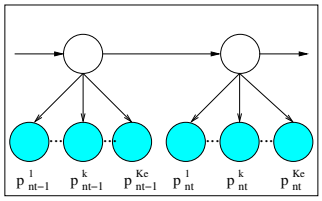


Figure 1: A MOHMM used for modelling the nth behaviour pattern. Observation nodes are shown as shaded circles and hidden nodes as clear circles.

An $N \times N$ affinity matrix $\mathbf{S} = [S_{ij}]$ where $1 \leq i, j \leq N$ provides a new representation for the training dataset, denoted as $\mathbf{D_s}$. Specifically, the nth behaviour pattern is now represented as the nth row of $\mathbf{S}$, denoted as $\mathbf{s}_n$. We thus have

$$\mathbf{D_s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n, \ldots, \mathbf{s}_N\} \tag{3}$$

Consequently each behaviour pattern is represented as a feature vector of a fixed length N (dynamically warped by a DBN). Taking a conventional data clustering approach, model selection is performed firstly to determine the number of clusters, which is then followed by data grouping using either a parametric approach such as Mixture Models or a nonparametric K-Nearest Neighbor model. However, since the number of data samples is equal to the dimensionality of the feature space, dimension reduction is necessary to avoid the 'curse of dimensionality' problem.

3.2. Eigendecomposition

Dimension reduction on the N dimensional feature space defined in Eqn. (3) can be achieved through eigendecomposition of the affinity matrix $\mathbf{S}$. The eigenvectors of the affinity matrix are then used for data clustering. However, it has been shown in [13, 11] that it is more desirable to perform clustering based on the eigenvectors of the normalised affinity matrix $\bar{\mathbf{S}}$, defined as:

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

where $\mathbf{L} = [L_{ij}]$ is an $N \times N$ diagonal matrix with $L_{ii} = \sum_j S_{ij}$. It has been proven in [16, 13] that the largest K eigenvectors of $\bar{\mathbf{S}}$ (i.e. eigenvectors corresponding to the largest eigenvalues) are sufficient to partition the dataset into K clusters. Representing the dataset using the K largest eigenvectors reduces the data dimensionality from N (i.e. the number of behaviour patterns) to K (i.e. the number of behaviour pattern classes). For a given K, standard clustering approaches such as K-means or Mixture Models can be adopted. The remaining problem is to determine the K, which is unknown. This is solved through automatic model selection.

3.3. Model Selection

We assume that the number of clusters K is between 1 and K_m. K_m is a number sufficiently larger than the true value of K. Suppose that we set $K_m = \frac{1}{5}N$ where N is the number of samples in the training dataset. This is a reasonable assumption since as a rule of thumb a more sparse dataset would make any data clustering algorithm unworkable. The training data set is now represented using the K_m largest eigenvectors, denoted $\mathbf{D_e}$, as follows:

$$\mathbf{D_e} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n, \ldots, \mathbf{y}_N\} \tag{5}$$

with the nth behaviour pattern being represented as a K_m dimensional feature vector

$$\mathbf{y}_n = [e_{1n}, \ldots, e_{kn}, \ldots, e_{K_m n}] \tag{6}$$

where e_{kn} is the nth element of the kth largest eigenvector $\mathbf{e_k}$. Since $K \ll K_m$, it is guaranteed that all the information needed for grouping K clusters is preserved in this K_m dimensional feature space.

We model the distribution of $\mathbf{D_e}$ using a Gaussian Mixture Model (GMM). The log-likelihood of observing the training dataset $\mathbf{D_e}$ given a K-component GMM is computed as

$$\log P(\mathbf{D_e}|\theta) = \sum_{n=1}^{N} \left(\log \sum_{k=1}^{K} w_k P(\mathbf{y}_n|\theta_k) \right), \tag{7}$$

where $P(\mathbf{y}_n|\theta_k)$ defines the Gaussian distribution of the k-th mixture component. The model parameters θ are estimated using the EM algorithm. The Bayesian Information Criterion (BIC) is then employed to select the optimal number of components K determining the number of behaviour classes. For any given K, BIC is formulated as:

$$BIC = -\log P(\mathbf{Y}|\theta) + \frac{C_K}{2} \log N \tag{8}$$

where C_K is the number of parameters needed to describe a K-component Gaussian Mixture.

However, it is found in our experiments that in the K_m dimensional feature space, BIC tends to underestimate the number of clusters (see Fig. 2(c) for an example and more in Section 4). This is not surprising because BIC has been known for having the tendency of underfitting the model given sparse data [9]. A dataset of N samples represented in a $K_m = \frac{1}{5}N$ dimensional feature space can always be considered as sparse. Our solution to this problem is to reduce the dimensionality through unsupervised feature selection.

3.4. Feature Selection

Now we need to derive a suitable criterion for measuring the relevance of each eigenvector. Since only the first K largest eigenvectors are needed for grouping K clusters, there are certainly redundant/irrelevant features in the K_m dimensional feature space defined in Eqn. (6). It is safe to say that a smaller eigenvector is less likely to be relevant in data clustering. It has also been shown in [13] that each of the K largest (i.e., relevant) eigenvectors of the normalised affinity matrix is able to separate one cluster from others while other eigenvectors are not. This suggests that a feature selection strategy should be employed based on measuring the relevance of each eigenvector according to how well it can separate the dataset into two clusters.

We denote the likelihood of the kth eigenvector $\mathbf{e_k}$ being relevant as $R_{\mathbf{e_k}}$ with $0 \leq R_{\mathbf{e_k}} \leq 1$. We assume that the elements of $\mathbf{e_k}$, e_{kn} follow two different distributions depending on whether $\mathbf{e_k}$ is relevant. It is thus natural to formulate the distribution of e_{kn} using a mixture model of two components. The likelihood of observing e_{kn} given the distribution parameters $\theta_{e_{kn}}$ can then be written as:

$$P(e_{kn}|\theta_{e_{kn}}) = (1 - R_{\mathbf{e_k}})P(e_{kn}|\theta^1_{e_{kn}}) + R_{\mathbf{e_k}}P(e_{kn}|\theta^2_{e_{kn}}) \tag{9}$$

where $P(e_{kn}|\theta^1_{e_{kn}})$ is the distribution of e_{kn} when $\mathbf{e_k}$ is irrelevant/redundant and $P(e_{kn}|\theta^2_{e_{kn}})$ when it is relevant. $R_{\mathbf{e_k}}$ acts as the weight or mixing probability of the second components. We assume the distribution of e_{kn} to be a single Gaussian when it is irrelevant:

$$P(e_{kn}|\theta^1_{e_{kn}}) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{k1}} \exp\left[-\frac{1}{2}\left(\frac{e_{kn} - \mu_{k1}}{\sigma_{k1}}\right)^2\right] \tag{10}$$

and a mixture of two Gaussians when it is relevant:

$$P(e_{kn}|\theta^2_{e_{kn}}) = \prod_{n=1}^{N} \left(\begin{array}{l} \frac{w_k}{\sqrt{2\pi}\sigma_{k2}} \exp\left[-\frac{1}{2}\left(\frac{e_{kn}-\mu_{k2}}{\sigma_{k2}}\right)^2\right] \\ +\frac{1-w_k}{\sqrt{2\pi}\sigma_{k3}} \exp\left[-\frac{1}{2}\left(\frac{e_{kn}-\mu_{k3}}{\sigma_{k3}}\right)^2\right] \end{array} \right) \tag{11}$$

where $\mu_{k1}, \mu_{k2}, \mu_{k3}, \sigma_{k1}, \sigma_{k2}$ and σ_{k3} are the means and variances of the three Gaussians in (10) and (11), w_k is the weight of the first component of the two Gaussians when $\mathbf{e_k}$ is relevant. There are 8 parameters required for describing the distribution of e_{kn}:

$$\theta_{e_{kn}} = \{R_{\mathbf{e_k}}, \mu_{k1}, \mu_{k2}, \mu_{k3}, \sigma_{k1}, \sigma_{k2}, \sigma_{k3}, w_k\}$$

The maximum likelihood (ML) estimate of $\theta_{e_{kn}}$ can be estimated using the following algorithm. First, the parameters of the first mixture component $\theta^1_{e_{kn}}$ are estimated as $\mu_{k1} = \frac{1}{N}\sum_{n=1}^{N} e_{kn}$ and $\sigma_{k1} = \frac{1}{N}\sum_{n=1}^{N}(e_{kn} - \mu_{k1})^2$. The rest 6 parameters are then estimated iteratively using EM. The ML estimate $\hat{R_{\mathbf{e_k}}}$ thus provides a real-value measurement of the relevance of $\mathbf{e_k}$. Since a 'hard-decision' is needed for dimension reduction, we eliminate the kth eigenvector $\mathbf{e_k}$ among the K_m candidate eigenvectors if

$$\hat{R_{\mathbf{e_k}}} < 0.5 \tag{12}$$

Since the EM algorithm is essentially a local (greedy) searching method and a mixture model is not unimodal, the EM algorithm could be sensitive to parameter initialisation especially in the presence of noise [6]. To overcome this problem, a priori knowledge on the relevance of each eigenvector can be utilised to set the initial value of $R_{\mathbf{e_k}}$. Specifically, we tie the initial value of $R_{\mathbf{e_k}}$, denoted as $\tilde{R_{\mathbf{e_k}}}$ with the corresponding eigenvalue:

$$\tilde{R_{\mathbf{e_k}}} = \bar{\lambda}_k, \tag{13}$$

where $\bar{\lambda}_k$ is the normalised eigenvalue for $\mathbf{e_k}$. The value of $\bar{\lambda}_k$ ranges from 0 to 1 with $\bar{\lambda}_1 = 1$ and $\bar{\lambda}_{K_m} = 0$. The other parameters of $\theta_{e_{kn}}$ are initialised using K-means.

After eliminating those irrelevant eigenvectors, the selected relevant eigenvectors are used to determine the number of clusters K and perform clustering based on GMM and BIC as described in Section 3.3. Each behaviour pattern in the training dataset is then labelled as one of the K behaviour classes. A synthetic dataset experiment is presented in Fig. 2 to illustrate the importance of feature selection on clustering using eigenvectors of a normalised affinity matrix. It is worth pointing out that for a noise-free synthetic dataset, the number of relevant eigenvectors is equal to the number of clusters and the BIC based model selection process becomes redundant. However, given a noisy dataset arising from a real problem, BIC based model selection after feature selection becomes crucial for determining the correct number of expected normal behaviour classes.

3.5. Abnormality Detection

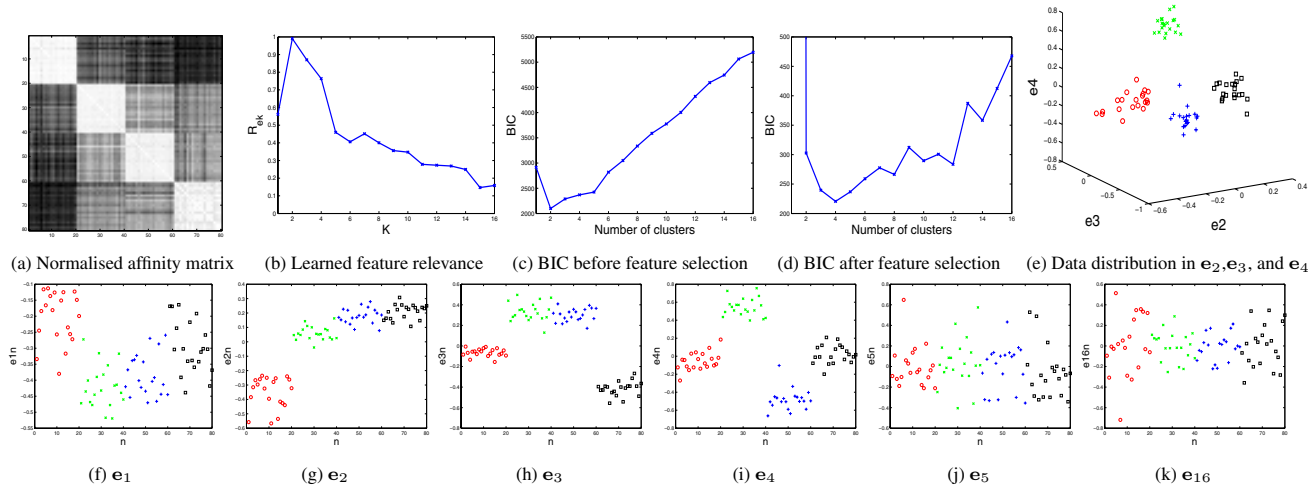
Now each of N behaviour patterns in the training set are labelled as one of the K classes. Firstly, a MOHMM $\mathbf{B}_k$ (see Fig. 1) is employed for modelling the kth behaviour class. The parameters of $\mathbf{B}_k$, $\theta_{\mathbf{B}_k}$ are estimated using all the patterns in the training set that belong to the kth class. A normal behaviour model $\mathbf{M}$ is then formulated as a mixture of the K MOHMMs for the K behaviour classes. Given an unseen behaviour pattern, represented as $\mathbf{P}$ (Eqn. (1)), the likelihood of observing $\mathbf{P}$ given $\mathbf{M}$ is:

$$P(\mathbf{P}|\mathbf{M}) = \sum_{k=1}^{K} \frac{N_k}{N} P(\mathbf{P}|\mathbf{B}_k) \tag{14}$$

where N is the total number of training behaviour patterns and N_k is the number of patterns belonging to the kth class. An unseen behaviour pattern is detected as abnormal if

$$P(\mathbf{P}|\mathbf{M}) < Th_A \tag{15}$$

where Th_A is a threshold.



(a) Normalised affinity matrix (b) Learned feature relevance (c) BIC before feature selection (d) BIC after feature selection (e) Data distribution in $\mathbf{e}_2$,$\mathbf{e}_3$, and $\mathbf{e}_4$

(f) $\mathbf{e}_1$ (g) $\mathbf{e}_2$ (h) $\mathbf{e}_3$ (i) $\mathbf{e}_4$ (j) $\mathbf{e}_5$ (k) $\mathbf{e}_{16}$

Figure 2: Clustering a synthetic dataset. 80 data samples were randomly generated using four MOHMMs with different random parameters. There are thus 4 data classes each of which has 20 samples. (a): the normalised affinity matrix ordered according to the data classes for illustration. (b): the learned relevance for the K_m largest eigenvectors. The first 4 largest eigenvectors were determined as relevant. (c) and (d) show the BIC model selection results before and after feature selection respectively. The number of clusters was determined as 4 after feature selection. (e): the 80 data sample plotted using the three most relevant eigenvectors, i.e. $\mathbf{e}_2$,$\mathbf{e}_3$, and $\mathbf{e}_4$. The 4 clusters can be easily separated. (f)-(k): the distributions of some eigenvectors. Elements corresponding to different classes are color coded in (e)-(k).

When an unseen behaviour pattern is detected as normal, the normal behaviour model $\mathbf{M}$ can also be used for recognising it as one of the K behaviour pattern classes learned from the training set. More specifically, an unseen behaviour pattern is assigned to the $\hat{k}$th behaviour class when

$$\hat{k} = \arg\max_k \{P(\mathbf{P}|\mathbf{B}_k)\}. \tag{16}$$

4. Experiments

Dataset and feature extraction — A CCTV camera was mounted on the ceiling of an office entry corridor, monitoring people entering and leaving the office area (see Figure 3). The office area is secured by an entrance-door which can only be opened by scanning an entry card on the wall next to the door (see middle frame in row (b) of Figure 3). Two side-doors were also located at the right hand side of the corridor. People from both inside and outside the office area have access to those two side-doors. Typical behaviours occurring in the scene would be people entering or leaving either the office area or the side-doors, and walking towards the camera. Each behaviour pattern would normally last a few seconds. For this experiment, a dataset was collected over 5 different days consisting of 6 hours of video totalling 432000 frames captured at 20Hz with 320×240 pixels per frame. This dataset was then segmented into sections separated by any motionless intervals lasting for more than 30 frames. This resulted in 142 video segments of actual behaviour pattern instances. Each segment has on average 121 frames with the shortest 42 and longest 394. Given these video segments, discrete events were detected and classified using automatic model order selection in clustering, resulting in four classes of events corresponding to the common constituents of all behaviours in this scene: 'entering/leaving the near end of the corridor', 'entering/leaving the entrance-door', 'entering/leaving the side-doors', and 'in corridor with the entrance-door closed'. Examples of detected events are shown in Figure 3 using colour-coded bounding boxes. It is noted that due to the narrow view nature of the scene, differences between the four common events are rather subtle and can be mis-identified based on local information (space and time) alone, resulting in large error margin in event detection. The fact that these events are also common constituents to different behaviour patterns reinforces the assumption that local events treated in isolation hold little discriminative information for behaviour profiling.

Model training — A training set consisting of discrete events extracted from 80 video segments was randomly selected from the overall 142 segments without any behaviour class labelling of the video segments. The remaining 62 segments were used for testing later. This model training exercise was repeated 20 times and in each trial a different model was trained using a different random training set. This is in order to avoid any bias in the abnormality detection and normal behaviour recognition results to be discussed later. For comparative evaluation, alternative models were also trained using labelled datasets as follows. For

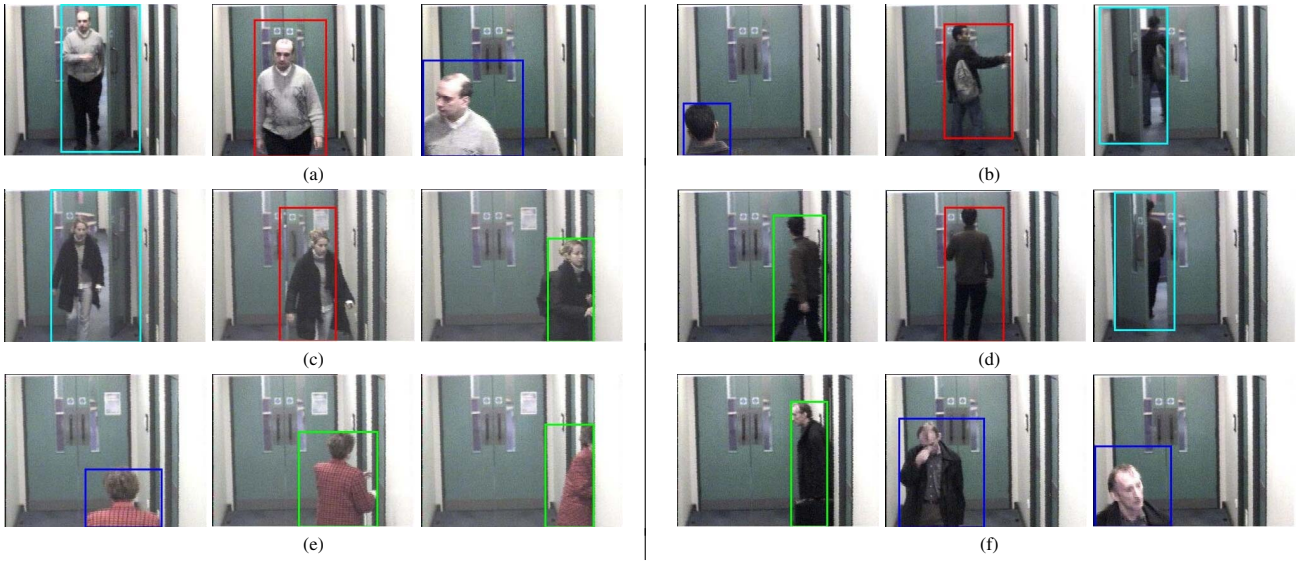


Figure 3: Behaviour patterns in a corridor scene: (a)–(f) show three representative frames of different typical behaviour patterns C1–C6 as listed in Table 1. Events detected during each behaviour pattern are shown by colour-coded bounding boxes in each frame.

each of the 20 training sessions above, a model was trained using identical training sets as above. However, each data sample in the training sets was also manually labelled as one of the manually identified behaviour classes. On average 12 behaviour classes were manually identified for the labelled training sets in each trial. Six classes were always identified in each training set (see Table 1). On average they accounted for 83% of the labelled training data.

C1	From the office area to the near end of the corridor
C2	From the near end of the corridor to the office area
C3	From the office area to the side-doors
C4	From the side-doors to the office area
C5	From the near end of the corridor to the side-doors
C6	From the side-doors to the near end of the corridor

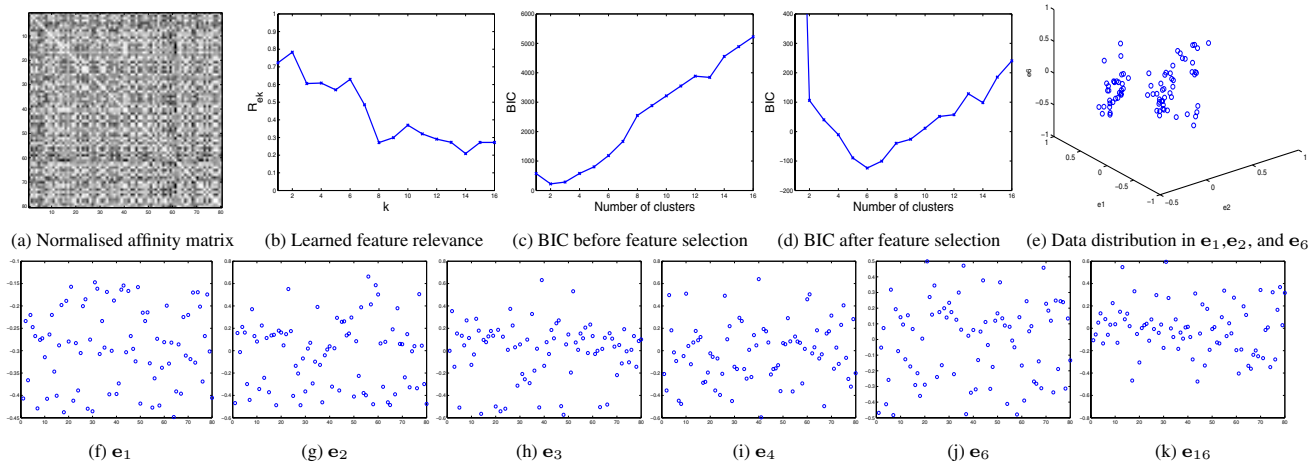
Table 1: The 6 classes of behaviour patterns that most commonly occurred in a corridor scenario.

Model training using unlabelled data: Over the 20 trials, on average 6 eigenvectors were automatically determined as being relevant for clustering with smallest 4 and largest 9. All the selected eigenvectors are among the 10 largest eigenvectors of the normalised affinity matrices. The number of clusters for each training set was determined automatically as 6 over in every trial. By observation, each discovered data cluster mainly contained samples corresponding to one of the 6 behaviour classes listed in Table 1. Fig. 4 shows an example of the model training process using an unlabelled training set. For each unlabelled training set, a normal behaviour model was constructed as a mixture of 6 MOHMMs as described in Section 3.5. *Model training using unlabelled data:* For each labelled training set, a normal behaviour model was built as a mixture of MOHMMs with the number of mixture components determined by the number of behaviour classes manually identified. Each MOHMM component was trained using the data samples corresponding to one class of manually identified behaviours in each training set.

	Ab. det. rate (%)	Fal. Ala. rate(%)
unlabelled	85.4 ± 2.9	6.1 ± 3.1
labelled	73.1 ± 12.9	8.4 ± 5.3

Table 2: Comparing the performance of models trained using unlabelled and labelled data on abnormality detection. The results were obtained over 20 trials with $Th_A = -0.2$.

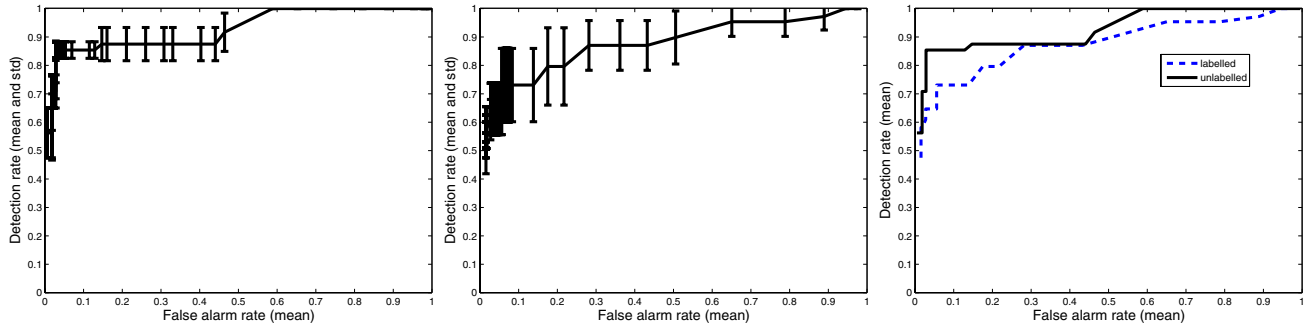
Abnormality detection — To measure the performance of the learned models on abnormality detection, each behaviour pattern in the testing sets was manually labelled as normal if there were similar patterns in the corresponding training sets and abnormal otherwise. The detection rate and false alarm rate of abnormality detection are shown in the form of a ROC curve. Fig. 5 shows that the models trained using unlabelled data clearly outperformed those trained using labelled data. It can also been seen from Fig. 5(a) & (b) that the performance of the models trained using unlabelled data was more consistent over different trials. In particular, it is found that given the same Th_A (see Eqn. (15)) the models trained using unlabelled datasets achieved higher abnormality detection rate, lower false alarm rate, and smaller variation over different trials compared to those trained using labelled datasets (see Table 2 and the last columns of the confusion matrices shown in Fig. 7). Fig. 6 shows ex-



(a) Normalised affinity matrix (b) Learned feature relevance (c) BIC before feature selection (d) BIC after feature selection (e) Data distribution in $\mathbf{e}_1$,$\mathbf{e}_2$, and $\mathbf{e}_6$

(f) $\mathbf{e}_1$ (g) $\mathbf{e}_2$ (h) $\mathbf{e}_3$ (i) $\mathbf{e}_4$ (j) $\mathbf{e}_6$ (k) $\mathbf{e}_{16}$

Figure 4: Model training using an unlabelled dataset. (b): the learned relevance for the K_m largest eigenvectors. The first 7 largest eigenvectors were determined as relevant features for clustering. (c) and (d) show the BIC model selection results before and after feature selection respectively. The number of clusters was determined as 6 after feature selection. (e): the 80 data samples plotted using the three most relevant eigenvectors, i.e. $\mathbf{e}_1$,$\mathbf{e}_2$, and $\mathbf{e}_6$. (f)-(k): the distributions of some eigenvectors.



(a) (b) (c)

Figure 5: The performance of abnormality detection measured by detection rate and false alarm rate. (a) and (b) show the mean and ± 1 standard deviation of the ROC curves obtained over 20 trials using unlabelled data and labelled data respectively. (c) compares the mean ROC curves using unlabelled and labelled data.

amples of false alarm and mis-detection by models trained using labelled data. It is noted that the lower tolerance towards event detection errors was the main reason for the higher false alarm rate of models trained using labelled data (see Fig. 6(b)&(d) for an example).

	Nor. Beh. Rec. Rate(%)
unlabelled	77.9 ± 4.8
labelled	84.7 ± 6.0

Table 3: Comparing the performance of models trained using unlabelled and labelled data on normal behaviour recognition. The results were obtained with $Th_A = -0.2$.

Recognition of normal behaviours — To measure the recognition rate, the normal behaviour patterns in the testing sets were manually labelled into different behaviour classes. A normal behaviour pattern was recognised correctly if it was detected as normal and classified into a behaviour class containing similar behaviour patterns in the corresponding training set by the learned behaviour model. Table 3 shows that the models trained using labelled data achieved slightly higher recognition rates compared to those trained using unlabelled data. Fig. 7(a) shows that when a normal behaviour pattern was not recognised correctly by a model trained using unlabelled data, it was most likely to be recognised as another class of normal behaviour pattern. On the other hand, Fig. 7(b) shows that for a model trained by labelled data, a normal behaviour pattern was most likely to be wrongly detected as an abnormality if it was not recognised correctly. This contributed to the higher false alarm rate for the model trained by labelled data.

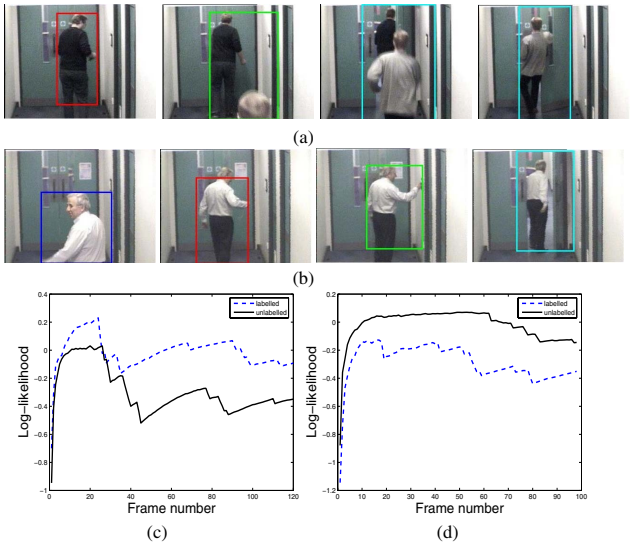
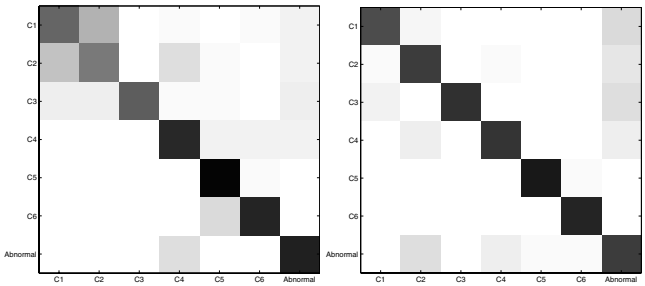


Figure 6: (a) An abnormal behaviour pattern which was detected as being abnormal by the model trained using an unlabelled dataset, but detected as being normal by the model trained using the same but labelled dataset. It shows a person sneaking into the office area without using an entry card. (b) A normal behaviour pattern which was detected correctly by the model trained using an unlabelled dataset, but detected as being abnormal by the model trained using the same but labelled dataset. (c)&(d) show the log-likelihood of observing the behaviour patterns shown in (a)&(b) given the learned models respectively.



(a) Models trained using unlabelled data (b) Models trained using labelled data

Figure 7: Confusion matrices for abnormality detection and normal behaviour recognition of the behaviour classes listed in Table 1. Each row represents the probabilities of that class being confused with all the other classes averaged over 20 trials. The results were obtained with $Th_A = -0.2$.

5. Conclusions

Our experiments show that a behaviour model trained using an unlabelled dataset is superior to a model trained using the same but labelled dataset in detecting abnormality from an unseen video. In particular, our behaviour profiling algorithm is capable of discovering natural grouping of behaviour patterns in the training data. The optimal number of behaviour pattern classes is automatically determined which enables the trained behaviour model to identify novel abnormality instances accurately and consistently. The model is also able to distinguish abnormal behaviour patterns from normal ones contaminated by errors in behaviour representation. On the contrary, the more deliberate model trained using labelled data tends to be brittle and less robust in dealing with unseen instances of behaviours that are not clear-cut in an open-world scenario. Such a model is more likely to perform poorly in interpreting unseen behaviour patterns, either normal or abnormal.

References

[1] H. Dee and D. Hogg. Detecting inexplicable behaviour. In *BMVC*, pages 477–486, 2004.

[2] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[3] J. Dy, C. Brodley, A. Kak, L. Broderick, and A. Aisen. Unsupervised feature selection applied to content-based retrival of lung images. *PAMI*, pages 373–378, 2003.

[4] Z. Ghahramani. Learning dynamic bayesian networks. In *Adaptive Processing of Sequences and Data Structures. Lecture Notes in AI*, pages 168–197, 1998.

[5] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *ICCV*, 2003.

[6] M. Law, M.A.T. Figueiredo, and A.K. Jain. Simultaneous feature selection and clustering using mixture model. *PAMI*, 26(9):1154–1166, 2004.

[7] R. J. Morris and D. C. Hogg. Statistical models of object interaction. *IJCV*, 37(2):209–215, 2000.

[8] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modelling human interactions. *PAMI*, 22(8):831–843, August 2000.

[9] S. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to Gaussian mixture modelling. *PAMI*, 20(11):1133–1142, 1998.

[10] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[11] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000.

[12] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *PAMI*, 22(8):747–758, August 2000.

[13] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *ICCV*, pages 975–982, 1999.

[14] T. Xiang and S. Gong. Activity based video content trajectory representation and segmentation. In *BMVC*, 2004.

[15] T. Xiang, S. Gong, and D. Parkinson. Autonomous visual events detection and classification without explicit object-centred segmentation and tracking. In *BMVC*, pages 233–242, 2002.

[16] S. Yu and J. Shi. Multiclass spectral clustering. In *ICCV*, pages 313–319, 2003.

[17] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR*, pages 819–826, 2004.