Context Aware Anomalous Behaviour Detection in Crowded Surveillance

Michael J. V. Leach\textsuperscript{a,1}, Ed. P. Sparks\textsuperscript{b}, Neil M. Robertson\textsuperscript{a}

\textsuperscript{a}Heriot-Watt University, Edinburgh Campus, Edinburgh, Scotland
\textsuperscript{b}Roke Manor Research, Romsey, Hampshire, United Kingdom

Abstract

This work addresses the detection of human behavioural anomalies in surveillance. We address in particular the problem of detecting subtle behaviour in a crowded behaviourally heterogeneous surveillance scene. We novel methods of extracting scene context and social context to improve the detection of behavioural anomalies, and in particular permit the detection of subtle behavioural anomalies. Our approach is unsupervised, detecting statistical outliers in a dataset by comparison of the outlier to the full data. We find that our context aware method performs significantly better than the equivalent method without contextual information. We observe that the use of a contextual information leads to the detection of instances of subtly abnormal behaviour, which otherwise remain indistinguishable from normal behaviour.

Keywords: behavior analysis, visual surveillance, security, context

1. Introduction

As a society we have the need to monitor public and private space in order to prevent criminal behaviour and identify security threats. The scale at which surveillance is undertaken, the density of information in video results in a huge amount of data - the analysis of which using human resources is often prohibitively expensive. The solution is to automate human surveillance [12]. Such systems have the potential to monitor a greater amount of data at a lower operating cost and mitigate the burden on security staff. In order to expose salient behaviour in a video we look for abnormal observations. An abnormal event is one which has a low statistical representation in the training data [9]. We motivate our by this definition with emphasis upon contextual information as a method of creating distinct separation between otherwise only subtly distinct behaviours. A good behaviour representation should encode the dataset in such a way that homogeneous clusters of behaviour can be segmented from the heterogeneous mass of data. Equally a poor behaviour representation is incapable of measuring the distinction between desired subgroups of data. Subtle behaviours provide a greater challenge because the information required to segment them from the greater set is not readily measurable. Subtle behaviours can be handled in the following two ways; firstly by measuring more relevant information which better segments the data into homogeneous subsets, or secondly by implementing a better suited model which is capable of fitting the nuances of the data domain. In this research we tackle the former point; inspired by work in Scene Modelling and Social Signal Processing we demonstrate the extraction and use of high level surveillance information which provides a contextual basis to identify abnormal behaviour which is only subtly distinct from normal scene behaviours. The contextual information enables the segmenting of heterogeneous behaviour sets into homogeneous sub groups. Simple surveillance scenes may not contain much contextual information, in fact at its simplest a surveillance scene can be said to have only one contextual state. In such cases a simple trajectory matching algorithm may be appropriate to detect outlier behaviour. However, a dynamic or crowded surveillance scene may be heterogeneous and behaviour in one context may not be representative of behaviour in a different context. In any non-trivial surveillance scene behavioural anomalies are often dependent upon contextual information such as scene region, social context, periodic events, entry and exit points, or other scene dynamics determined from observations [21]. We can use this contextual information to provide further means of segmenting abnormal behaviours from the mass of data, and perhaps providing the means to segment subtle behaviours from the mass of data. For a more general discussion on contextual anomaly detection see [5, 24].

In this study we evaluate whether social and scene region contextual knowledge improves the detection of behavioural anomalies and permits the detection of subtle behavioural anomalies. We define ‘anomalies’ as belonging to the set of most abnormal behaviours in a dataset. A subtle anomaly is one which does not contrast the main statistical distribution of normal behaviour in its directly measurable motion information, such that it may require some further prior understanding of the environment to be segmented from other behaviours. Subsequently, a non-subtle anomaly is one which can be observed to be in direct contrast to other observed behaviours. We focus upon two types of contextual information in this study: social con-
text and scene region context. The scene regions provide an understanding of portions of the scene in which we would expect normal behaviours to be different from other areas [12]. Social contextual information grants the ability to learn the distinction between normal behaviour for groups and individuals independently. The social model provides an additional benefit; it ensures that the behaviour of each individual is analysed in reference to people external to the same social group. Thus, a homogeneous group of individuals all acting abnormally cannot be self-justifying. Furthermore, social information enables us to create likelihood dependencies between individuals in a social group. Thus, if one individual in a group is behaving abnormally, the expectation of other group members behaving abnormally goes up. We demonstrate our system’s capability to detect subtle behavioural anomalies within a moderately complex and crowded human surveillance scene. We consider our first dataset PETS 2007 [17] to be moderately crowded, and our second dataset, the Oxford town center data [2], to be sparsely crowded. The PETS 2007 data contains a total of 573 individuals over 11902 frames, averaging 24 people in the scene at any given frame in a space measuring 16.2 meters by 7.2 meters. The Oxford data contains 430 tracked pedestrians over 4500 frames. There are an average of 15 individuals in any given frame, with a minimum of 5 and a maximum of 29. We select the second dataset, the Oxford data, to test our social context approach for failure modes. The sparser Oxford data is highly structured and thus the trajectories of socially unconnected pedestrian are often very similar, and often close in proximity - giving the appearance of social connectivity. We expect this will produce false positive social context information.

Our main contributions are:
1. A novel method of acquiring scene structure information in surveillance
2. The development of a novel social group metric
3. The demonstration that contextual information is effective at isolating subtle behavioural anomalies

1.1. Related Work

There are four main components to our anomaly detection process: Tracking in a video sequence, building contextual scene models, building a social model, and detecting abnormal behaviour. We address the related work individually.

Detecting abnormal behaviour: Due to sensor noise and the time series nature of behaviour statistical models such as Hidden Markov Models (HMM) and more general Dynamic Bayesian Networks (DBN), commonly feature in the machine modelling of human behaviour [26, 13, 6, 3, 7, 1]. HMMs however have a number of drawbacks: poor scalability for multiple interacting objects; inflexible temporal scaling; and are vulnerable to structured noise in the training dataset [3]. An alternative behaviour model proposed by Li et al. [8] is that of behavioural correlation. The behaviour ontology is not dependent upon object centered sustained tracks but instead short term atomic actions. Such a system seeks to cope with expectation that it is not possible to maintain long tracks in surveillance by tracking short term foreground blobs in the image. However due to advances in pedestrian detection and robust tracking long term tracks are becoming feasible [27, 15]. Similar in method to the work of Li [8] and Loy [9] is that of Hospedales, Xiang and Gong. Hospedales et al. [20] introduces a novel Markov clustering topic model which performs hierarchical clustering of visual events into behaviours, and exposes salient behaviour. The system is robust to cluttered and noisy input, however actions do not necessarily pertain to an individual, making such a method unfavourable towards the inclusion of social context.

Scene Context: Li et al. develops a scene segmentation method which divides the scene into regions based upon behavioural dissimilarity [8]. However, the scene segmentation is solely dependent upon observed behaviour and is in effect an extension of behavioural clustering. Chen Loy segments a scene similarly into spatial regions of similar behaviour by virtue of correlated behaviour [9]. This work introduces a second line of contextual scene knowledge: temporal state. This contextual information is particularly apt for the traffic junction, in which behaviour is clearly temporally segmented in short time intervals. However, it is far less applicable to many human surveillance environments where the periodicity of behaviour is far less structured, if at all. Wang et al. uses a Dual Hierarchical Dirichlet Process to cluster behaviours spatially, learning both
observation and trajectory clusters simultaneously, in a computationally efficient model [25]. Makris et al. segments a scene into regions which fit a semantic scene model [12] based on observed trajectories. A dynamic surveillance scene is segmented into entry/exit zones, junctions, paths, routes (a series of paths), and stop zones based upon observed trajectory features.

**Social:** The second contextual component we address is that of social groupings between individuals. Ge et al. uses a proximity and velocity metric to associate individuals into pairs, and then iteratively add additional individuals to groups using the Hausdorff distance as a measure of closeness to a group [23]. Yu et al. implement a similar graph based system which uses the feature of proximity alone to encode the strength of edges between nodes [22]. A connectivity graph is constructed from the strengths of observed connections from which groups are segmented via graph cuts. However modelling social groups by positional information alone is perilously primitive and prone to finding false social connections when individuals are within close proximity due to external influences such as queuing or high traffic regions. Oliver et al. uses a Coupled HMM to construct a-priori models of group events such as Follow-reach-walk together, or Approach-meet-go separately [14]. Certain actions are declared group activities and thus groups can be constructed from individuals via mutual engagement in a grouping action. On the same methodology the work of Hosie et al. classifies feature segments into primitives at the sample rates of 10, 20 and 30 frames [18]. In the work of Robertson and Reid, social activities such as meeting utilise the feature of gaze direction estimates in order to determine whether individuals are within each others field of view [13]. The use of gaze direction is significant as it departs from the use of motion features alone, and takes into account visual interest. Farenzena presents a novel method of inferring social interactions between individuals in a crowd based upon the cues of gaze direction and proximity [11]. For a comprehensive and complete review of the emerging field of social signal processing see the work of Cristani [10].

2. Method

2.1. Feature Extraction

The extraction of pedestrian trajectories from surveillance video is non-trivial, particularly when there is occlusion and crowding. It is not our goal to develop a novel low level feature extractor and for that reason we rely upon the large amount of research in computer vision already devoted to producing tracking solutions. Extracting pedestrian trajectories requires two main stages: detection of pedestrians in the image plane, and tracking of targeted pedestrians. Detection is achieved using the Felzenszwalb part based detector [15]. Tracking of human targets in the image plane is achieved with the use of the Predator TLD tracker [27] on pedestrian detections in the PETS 2007 data [17]. We track the heads of pedestrians in the crowded PETS-2007 scene, see figure 2(a). We used the published tracking results provided by Benfold [2] in the Oxford dataset. We select the TLD tracker due to high performance amongst state of the art trackers, however more importantly its ability to learn a target model and discriminate between potential targets was critical in a busy surveillance scene such as the PETS 2007 dataset. The pedestrian tracking performance of the TLD tracker is extensively tested against alternative recent tracking procedures in the author’s paper [28].

2.2. Context Aware Anomaly Detection

Figure 1 shows the process our method follows. Our approach is unsupervised; observations of behaviour find support from the set of all behaviours observed, if possible. Anomalies are thus discovered due to their contrasting nature to previously observed behaviour.
2.3. Scene Context

Building upon the work of Makris and Ellis [12] our scene model consists of four potential regions: Traffic lanes, idle areas, convergence/divergence regions, and general area. Convergence and divergence is synonymous as there is no temporal direction. Each region is defined to isolate a different dynamic of a scene, and is captured as a relation between the directional, speed, persistence (the number of frames a trajectory last for), and energy and entropies of trajectories through the scene. For each of the four potential regions a heat map is constructed on the ground plane and a threshold segments positive regions from negative. Scene regions are mutually exclusive of each other. We define each of the four scene context regions as follows:

Traffic Lanes: A traffic lane represents an area of the scene which contains a high number of trajectories in a structured motion. The traffic region is defined as:

$$T_{xy} = \frac{N_{xy}}{N} - \frac{1}{\sum P(\theta_{xy}) \log(P(\theta_{xy})) + \frac{1}{2} \sum (\theta_{xy} - \bar{\theta}_{xy})^2}$$

(1)

Where $\theta$ is a histogram of directions populated by all target trajectories to go through region $x, y$ in the scene. The numerator $N_{xy}$ gives the number of trajectories through the location $x, y$, and $N$ gives the mean number of trajectories for any given location. High scoring traffic locations coincide with regions displaying a high number of trajectories, low directional entropy and low trajectory energy.

Idle Regions: The idle region captures the area of the scene which hold enough evidence of near stationary trajectories that the region is considered a legitimate place to remain idle.

$$I_{xy} = \frac{T_{xy}}{T} \frac{v_{xy}}{\sum \sqrt{(v_{xy} - \bar{v}_{xy})^2 + \sum v_{xy}}}$$

(2)

The mean temporal persistence $T_{xy}$ provides the mean numbers of frames that trajectories persist for in the region $x, y$, this coefficient is balanced by the denominator $T$ the mean number of frames for all regions. The speeds of trajectories observed in location $x, y$ is denoted by $v$. We define likely idle regions as those with a high mean temporal persistence, low speed and low speed energy.

Convergence Divergence areas: These areas of the scene are responsible for imposing a force which brings trajectories together or frees them allowing them to diverge. Typically such regions are appended to the ends of a traffic lane.

$$C_{xy} = \frac{1}{2} \sum (\theta_{xy} - \bar{\theta}_{xy})^2 - \sum P(\theta_{xy}) \log(P(\theta_{xy}))$$

(3)

Where $\theta$ is the histogram of direction observed at $x, y$. We define the convergence region by a high directional energy low
directional entropy region. Thus a structured splitting of trajectories over a region would be considered a likely candidate for a convergence or divergence region.

**General Area:** having scored the scene with the above region definitions we normalise the region intensity maps between [0,1], and apply a threshold to segment active regions. We found in experiment a threshold of 0.5 to be appropriate. The remaining area of the scene not classified as any of the above regions is considered the general area. The interpretation of the general area is as the region which does not impose any influence on the motion vector of tracked pedestrians or ships.

### 2.4. Social Context

The basis of our social model is a metric capable of measuring the apparent strength of a social connection between any two people. We proceed on the basis that a high degree of shared trajectory information implies a social dependence between two individuals. To test and train our model we generated an independent social ground truth for both the PETS 2006 [15], and PETS 2007 [17]. To set the weighting function and train the model an initial one off training phase is required. We trained upon the PETS 2006 sequence which consisted of 28 people with 14 social connections between them of varying strength. We use a subset of data to test the social model (PETS 2007, scene 4) which contains 152 tracked individuals, 44 social connections. For an illustration of typical social pairs see figure 4(b). Our social model is geared towards effective detection of social groups in a moving crowd. It is essential that potentially socially connected individuals are moving such that trajectory information can be extracted. Furthermore crowded surveillance provides an environment in which socially connected individuals are more likely to move together, and thus display more similar trajectory information. The more entropic the underlying motion of the crowd is the more salient similar trajectories will be.

We use a novel metric to identify the strength of pair-wise social connections consisting of the weighted product of multiple features. We identified 4 features as effective at detecting pair connections between to individuals: the mutual information of direction (IΘ), the mutual information of speed (IV), the proximity between two individuals (ΔP) and the temporal overlap ratio between two individuals (T). We train a set of weighting variables αΔP, αIV, αIΘ, ατ which weight each feature in the social metric based upon the classification score of each feature independently on the ground truth training data. The features weights are distributed proportional to the classification score. The features which compose the pairing metric are defined as: For 2 individuals i,j at time t.

\[
\Delta P_{ij} = \alpha \Delta P e^{-\frac{\sum_{i,j} |P(i,j) - P(i')j|}{\sum_{i,j} |P(i,j)|}} 
\]

\[
\Delta I = \alpha I e^{-\frac{\sum_{i,j} |I(i,j) - I(i')j|}{\sum_{i,j} |I(i,j)|}} 
\]

\[
\Delta IV_{ij} = -\alpha IV \sum_{b} P(v(b))log_2(P(v(b))) 
- \alpha IV \sum_{b} P(v(b))log_2(P(v(b))) 
+ \alpha IV \sum_{b} P(v(i,b))log_2(P(v(i,b))) 
\]

\[
\Delta IΘ_{ij} = -\alpha IΘ \sum_{b} P(θ(b))log_2(P(θ(b))) 
- \alpha IΘ \sum_{b} P(θ(b))log_2(P(θ(b))) 
+ \alpha IΘ \sum_{b} P(θ(i,b))log_2(P(θ(i,b))) 
\]

ΔP_{ij}, ΔI_{ij}, ΔIV, ΔIΘ are the Euclidean difference in proximity, temporal overlap and mutual information measure for speed and direction between person i and person j at frame t. τ_{ij} is the temporal overlap ratio between i and j up to the current frame t, which is to say the ratio of time both individuals have existed contemporaneously to total time of existence, thus rewarding individuals who enter and exit the scene at similar times. T_i and T_j is the frame length of trajectory i and j respectively, and T_i is the number of frames in which both i and j have coexisted. The proximity between any two individuals ΔP is scaled by the distance between i and j to all other individuals in the scene. Thus we incorporate a measure of scene density which places a bias upon pairs being closer together in denser areas, and allows pairs to drift apart in sparse areas.
connection strengths to identify how each feature is at

We define the distribution $P(x_\cdot y)$ using the Maximum Likelihood Estimation for Gaussian parameters $\sigma$ and $\mu$ using the most recent 1 second of trajectory data. The joint probability is calculated as the MLE Gaussian for the combined data of both person $i$ and $j$ over the last second. The mutual information between individual $i$ and $j$ is calculated for a number of temporal offsets thus permitting an individual reaction time to the trajectory it has dependence upon. Thus we calculate the mutual information between each individual with set time offsets of 10 frames consecutively forwards and backwards, and take the mutual information for all time offsets.

We analyse each feature against a hand built ground truth of pair connections between observed individuals and comparing the classification against the ground truth. We observed that the mutual information speed and direction metrics outperform the Euclidean distance feature metrics in overall true positive classification.

Whilst $\Delta P_{ij}$ and $\Delta \tau_{ij}$ are direct measures of trajectory statistics it is important to note that both $IV_{ij}, J\Theta_{ij}$ are more complex in nature. We use mutual information (MI) as it handles non-linear and non-Gaussian random variables effectively and provides a principled method of comparing orthogonal feature dimensions. The key to a successful MI metric is in the definition of the joint probability distribution $P(xy) \neq P(x \cdot y)$. We define the distribution $P(v)$ and $P(\theta)$ as the Maximum Likelihood Estimation for Gaussian parameters $\sigma$ and $\mu$ using the most recent 1 second of trajectory data. The joint probability is calculated as the MLE Gaussian for the combined data of both person $i$ and $j$ over the last second. The mutual information between individual $i$ and $j$ is calculated for a number of temporal offsets thus permitting an individual reaction time to the trajectory it has dependence upon. Thus we calculate the mutual information between each individual with set time offsets of 10 frames consecutively forwards and backwards, and take the maximal mutual information for all time offsets. We analyse each feature against a hand built ground truth of pair connection strengths to identify how effective each feature is at identifying pairs in the training data. Analysis of the features was carried out by using each feature independently to classify pair connections between observed individuals and comparing the classification against the ground truth. We observed that the features of proximity between two individuals ($\Delta P$) and the temporal overlap ratio between two individuals ($\Delta \tau$) present a significant ability to classify pairs in the test data. However there the overall performance is improved with the inclusion of the mutual information measures for direction and speed, see figure 5.

To measure the overall social connection strength between two individuals we utilise the pairwise strength in the previous step in the following way. A trajectory of length $T$ frames consists of $T$ tuples $(S, v, \theta)$ for 2D ground plane position vector $S$, speed scalar $v$ and direction of trajectory in radians $\theta$. We can calculate the pair strength at frame $T$ between any two individuals $i$ and $j$, for $i, j \in N$ where $N$ is the set of all individuals in the scene for all frames. The social connection strength $\kappa$ between two individuals $i$ and $j$ at time $T$ is:

$$\kappa_{ij} = \frac{1}{T} \sum_{t} IV_{ij} J\Theta_{ij} \Delta P_{ij} \tau_{ij}$$

$\tau_{ij}, IV_{ij}, J\Theta_{ij}, \Delta P_{ij}$ are the temporal overlap, mutual information for speed, mutual information for direction and proximity difference between person $i$ and $j$, as detailed in the feature equations (4), (5), (6) and (7). We classify the social state $S$, where $S \in [0,1]$, by applying social strength threshold $\lambda$ which is set empirically from the training data. Connections between individuals which score higher than $\lambda$ are considered socially connected, providing binary the social context used in the anomaly detection stage.

2.5. Anomaly Detection

Anomaly detection splits into three distinct segments: the behaviour ontology, the method for calculating normality of observations, and the algorithm for detecting anomalies.

Behaviour Ontology: Our behaviour ontology consists of a four part feature vector $x = \Re^4$, consisting of a bivariate motion component [speed, persistence], and the two contextual states [social state, scene region]. Speed is measured in meters per second on the ground plane, and social state is a binary state describing whether the individual is part of a social group or not. The persistence of an individual is a measure in frames of how long an individual has remained in the scene for. Lastly, the scene region identifies the scene context region in which
the individual resides, denoted by a numerical identifier. For an individual with trajectory length $T$ frames we have $T$ feature vector observations. The observations are accumulated to a discrete 4 dimensional feature space representing a 4D histogram, termed the behaviour profile $X_i$, for individual $i$. Defined in this way $X_i$ consists of a feature distribution from a large number of observations. The advantage to this is that it hides short-term measurement noise resulting in a behaviour ontology which is more robust to random sensor noise. Furthermore, as measurement noise is often correlated rather than Gaussian white noise, the order independent nature of the behaviour profile $X_i$ overcomes the appearance of anomalies that arise from structured noise. Our behaviour profile provides flexible temporal scaling of behaviours; something DBNs struggle with, however it results in the loss of time series information which may reduce the descriptive capacity of the ontology.

**Normality of behaviour observations:** Much work to date has focused upon a frequency based analysis to determine the normality of behaviour observations. With a suitably long data stream all possible normal variations of behaviour are observed, and thus the legitimacy of any observable behaviour can be determined by its frequency of occurrence in the data [3]. However, frequency-based anomaly detection suffers under the following assumption: that the normality of any observed behaviour is proportional to the relative frequency of observations of the behaviour. Whilst we can expect abnormal events to be rare, it is not the case that normal events are all frequent, and proportionally represented. We wish to distinguish here between the *normality* of a behaviour and the *expectation* of a behaviour. The expectation of a behaviour is how likely it is to occur next, whereas the normality of a behaviour is how permitted the behaviour is in the scene; how legitimate it is. A frequency based analysis reveals expectation of each behaviour to occur next, not the intrinsic normality of the behaviour itself, thus missing the mark. We instead implement a Nearest Neighbour method to search for supporting evidence for an observation from others within the data. The normality of any behaviour is based upon its similarity to the nearest K instances of supporting evidence not the frequency of observation for that behaviour.

A subtle anomaly may not be distant from the set of normal behaviour with regard to the majority of features. For example a subtle anomaly may pertain to only the speed and scene region, and this may not be realisable if we include social state and persistence in the distance measure. Additionally a subtle anomaly may be an outlier for only one feature when seen in the context of another feature. For example the speed is abnormal only when seen in the context of a specific scene region, rather than the speed and scene region both being independently abnormal. As such we need to assign a normality score to each feature in context of each other feature, independently of every other feature. Determining the outliers in each feature independently enables the identification of differences between behaviours which manifest in only one feature or context, a step critical to detecting subtle differences between behaviours. This step enables us to see context dependent distinctions between behaviours which when viewed in the full feature space are too subtle to impact a distance calculation. Thus we measure the normality of each observed feature given every other feature for all of the four features speed, persistence, social state and scene region. To represent feature context behaviour we reduce our 4D histogram feature space for person $n$ and frame $t$ to a set of 1D feature distributions $Y_n^{T_1}$, detailing the distribution of feature $f_1$ given the currently observed value for feature $f_2$. For a feature vector $x_i$. With dimensionality D there are $D^2-D$ feature pair distributions covering each $\{f_1, f_2\}$ feature pairing, when $f_1 \neq f_2$. In our 4D feature space 12 individual feature pairs are assessed at each frame for each individual, each representing a different observation given context pairing. Thus, we reduce the dimensionality of $X_i$ to 2 by summing the distribution $X_i$ for all dimensions $f$ in the set of dimensions $F$ where:

$$f = \{ f \in F : f \neq f_1 \land f \neq f_2 \}$$

resulting in a joint distribution $Y_n$ of observation feature $f_1$ and context feature $f_2$. We then take a further step reducing the 2D distribution to the target 1D distribution by taking the distribution through the current context feature value $f_2(i)$ only. Thus our resulting distribution $Y_n^{1/2}$ details the distribution of observed feature values for observation feature dimension $f_1$ given the context feature state $f_2(i)$. An example of which would be the distribution of the speed feature given the scene feature of idle region.

We apply the Nearest Neighbour (NN) function to distribution $Y_n$ and the set of all distributions $Y$ to determine the nearest neighbour $Y_m$ to $Y_n$ for each feature context pairs $\{f_1, f_2\}$, $F$. The Nearest Neighbour distance metric specified is the Bhattacharyya coefficient. The nearest neighbour distance metric for feature context pair $\{f_1, f_2\}$ is thus defined as:

$$B(Y_n, Y_m) = \sum_h \sqrt{Y(h_1)f_1^{1/2}Y(h_2)f_2^{1/2}}$$

Where we sum over all histogram bins $h$ for feature dimension $f_1$. Thus given a feature vector for individual $n \in N$ at frame $t \in T$ we find the nearest neighbour $m$ where $\{m \in N : n \neq m\}$

$$NN(Y_n) = \{Y_m \in Y|YY_{p} \in Y : B(Y_n, Y_m) \geq B(Y_n, Y_p)\}$$

The nearest neighbour equation specifies $m$ the index of the least distant behaviour profile of $n$ for feature context pair $\{f_1, f_2\}$ and $B$ the resultant Bhattacharyya coefficient. As the Bhattacharyya coefficient is a measure of similarity, scoring more similar distributions higher, the NN finds the greatest Bhattacharyya coefficient to distribution $Y_n$ from the set of all distributions $Y$ given the feature context pair $\{f_1, f_2\}$. We then combine the independent feature context pairs to generate a single value for the abnormality $A(n, t)$ for person $n$, at frame $t$. The abnormality coefficient of behaviour at frame $t$ for person $n$ is the least supported feature pairing; the lowest similarity to the
Figure 6: ROC charts for Anomaly Detection classification, with a comparison of different contextual setups. (a) shows the results from PETS-2007 Scene 00, (b) from PETS-2007 Scene 02, and (c) from PETS-2007 Scene 04.

nearest neighbour:

\[ A(n, t) = \arg\min_{f_1, f_2} B(Y_{n f_1}, Y_{n f_2}) \]  

(12)

Anomaly detection: The abnormality coefficient for each individual at each frame is used to determine anomalies. Threshold \( \mu \) separates anomalies from normal observations and in effect represents the sensitivity of the system. If we seek to detect only anomalies then \( \mu \) represents the expectation of abnormal behaviour in the sequence. For the end user \( \mu \) represents a constant surveillance workload for the operator. Variable \( \mu \) can be either set by the operator or defined empirically in an additional training phase. Anomalies \( A(n, t) \) at frame \( t \) for person \( n \) are classified by:

\[ A(n, t) = \delta(A(n, t)) = \begin{cases} 1, & A(n, t) < \mu \\ 0, & A(n, t) \geq \mu \end{cases} \]  

(13)

Based upon the assumption that there is dependence between the behaviour of individuals within the same social group we additionally utilise the social contextual information in an additional two ways. Firstly we ensure that the behaviour of each individual is analysed in reference to people external to the same social group. Thus a behaviourally homogeneous group of individuals all acting abnormally cannot be self justifying. We enforce this removing the indexes of individuals from the same social group from the nearest neighbour calculation for individuals in that group. Secondly, social information enables us to propagate the expectation of an anomaly through the entire social group. In this way each member of a social group at any given frame has the highest anomaly score for all individuals in that group. Thus if one individual in a group is behaving abnormally all group members are equally as abnormal.

We do not implement any post process alarm filtering. We justify the exclusion of this process as it may obscure the change in accuracy resulting from the inclusion and exclusion of contextual information.

3. Experiment

From the outset we wished to evaluate whether social and scene region contextual knowledge improves the detection of behavioural anomalies and permits the detection of subtle behavioural anomalies. We now detail the results of an anomaly detection experiment on the PETS 2007 dataset with the inclusion and exclusion of contextual information. Furthermore we test against a state of the art behaviour anomaly detection system.

The publicly available PETS 2007 dataset [17] offers a source of multi camera real world surveillance footage. The datasets consists of 8 sequences each captured from 4 different view points. The PETS data presents a difficult anomaly detection task due to tracking and behavioural complexity. There is a high degree of occlusion and variation in appearance and sensor/compression noise, whilst many of the tracking targets are small, ranging from 100 to 200 pixels. Behavioural anomalies in this scene are characterised by gross motion abnormality such as a group running across part of the scene, or subtle anomalies such as a single individual standing still in a flow of traffic, or a group loitering amongst a crowd. We specifically chose this data due to its behavioural complexity for anomaly detection.

3.1. Scene Segmentation

We first report upon the results of scene segmentation. As detailed in the methodology section we address 4 scene context regions: Idle regions, traffic regions, convergence/divergence regions and the general region. We found well defined regions for the idle, divergence and traffic region in the PETS data which fit with the intuitive interpretation of the scene, see Fig.3. The Oxford data held well defined areas for the traffic region and the divergence region. However the idle region hardly featured. This finding fits with the highly structured nature of the Oxford data in which there are very few stationary tracks, see Fig.3(d).
Figure 7: The anomaly detection results on the Oxford Dataset. We test upon the Oxford data to test for a failure mode in the social model.

3.2. Social Context

We test the social context classification against a independently constructed ground truth for social connections. The training data (PETS 2006) consisted of 28 people with 14 true positive unique social connections between them of varying strength. The test data (PETS 2007) contains 152 tracked individuals, 44 social connections. We found that classifying social connections in the PETS 2007 data using parameters trained in the PETS 2006 data achieved a true positive detection rate of 0.92 and a false positive rate of 0.092, see Fig. 5 (a). There are a greater number of false positive social connections in the Oxford data. We observe that this is in fact due to the high degree of structure in the motion in the Oxford data which gives the false appearance of mutual dependency of motion. In the human PETS data however the motion of individuals is far more entropic, thus structure from social dependency is far more salient.

3.3. Anomaly Detection

To demonstrate the impact context information has upon anomaly detection we determine the accuracy in four states: no contextual information, only scene context, only social context and with both types of contextual information. A comparison is made of the true positive and false positive detection rates detecting the anomalies listed in Table 1. We also demonstrate illustrative examples of subtle behaviours which are detected only when contextual information is used. The anomaly ground truth reveals 9 behavioural anomalies in the PETS 2007, and 3 anomalies over 4500 frames in the Oxford data. In both the PETS and Oxford data we vary the $\mu$ threshold from 0 to 1 in small increments to adjust the systems sensitivity to unlikely observation, thus populating the ROC curve for detection accuracy. Fig. 6 (a), (b) and (c) demonstrates the anomaly detection success in the PETS 2007 dataset. Fig 7 illustrates the results on the Oxford data.

Table 1: The behavioural anomalies in PETS 2007 (3 sequences) and Oxford Data. (1), (2) and (3) occur due to a group standing on the left of the scene looking around and suddenly dispersing in different directions. Anomalies (4) and (5) occur due to two individuals entering the scene, turning a corner and then suddenly turning around and leaving in the same place they entered. (6) is a known ground truth behavioural anomaly. One of the participants in the PETS 2007 experiment purposefully loiters in a busy scene. (6), (7) and (8) are all members of a small group of 3 running through the scene, from the top to the bottom of then scene. (9), (10), (11), and (12) are four more instances of known ground truth anomalies. Two individuals purposefully loiter in the scene whilst another two suspiciously switch bags. In the Oxford data, anomaly (13) is due to the unique behaviour of the individual interacting with a bin in the scene. Anomaly (14) captures an individual entering the scene at the bottom and loitering in the middle. Anomaly (15) captures a women meandering slowly through the scene.

<table>
<thead>
<tr>
<th></th>
<th>PETS 2007 (Scene s00)</th>
<th>PETS 2007 (Scene s02)</th>
<th>PETS 2007 (Scene s04)</th>
<th>Oxford Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Id</td>
<td>Start</td>
<td>End</td>
<td>Id</td>
</tr>
<tr>
<td>Unusual group behaviour</td>
<td>1</td>
<td>1</td>
<td>2656</td>
<td>Running through scene</td>
</tr>
<tr>
<td>Unusual group behaviour</td>
<td>2</td>
<td>1</td>
<td>2419</td>
<td>Running through scene</td>
</tr>
<tr>
<td>Unusual group behaviour</td>
<td>3</td>
<td>1</td>
<td>2714</td>
<td>Running through scene</td>
</tr>
<tr>
<td>Abrupt you turn in busy area</td>
<td>4</td>
<td>2627</td>
<td>2928</td>
<td>Bag swap, unusual motion</td>
</tr>
<tr>
<td>Abrupt you turn in busy area</td>
<td>5</td>
<td>2604</td>
<td>2928</td>
<td>Bag swap, unusual motion</td>
</tr>
<tr>
<td>ground-truth loitering</td>
<td>11</td>
<td>1</td>
<td>2596</td>
<td>ground-truth loitering</td>
</tr>
<tr>
<td>Motion + interaction with scene</td>
<td>13</td>
<td>3554</td>
<td>4349</td>
<td>Loitering</td>
</tr>
<tr>
<td>Abnormally slow movement</td>
<td>15</td>
<td>2382</td>
<td>3454</td>
<td></td>
</tr>
</tbody>
</table>

The output of our system is displayed in figure 9. Images (a) through (c) show correct identification of anomalies. Image (a) shows an example of a context independent anomaly: running through the scene. The behaviour is however more salient when self justifying groups are prevented using the social context. Image (b) shows two examples of context dependent anomalies. The motion features pertaining to the anomaly are common within the entire scene, and detected as an anomaly only when viewed in light of the high persistence feature given scene context information.

To see the capability of our anomaly detection system in reference to the state of the art we include an implementation of the Weakly Supervised Joint Topic Model (WSJTM) proposed and developed by T. Hospedales, Jian Li, Shaogang Gong and Tao Xiang. WSJTM was selected because it is based upon a different behaviour representation whilst its use of positional information makes it comparable to our scene contextual infor-
However, the primary reason for making comparisons to this piece of work is that the procedure is adept at detecting subtle outlier behaviour similar in style to our own work. For a detailed account of this work see Identifying Rare and Subtle Behaviours: A Weakly Supervised Joint Topic Model [19]. The comparison between our results and the WSJTM procedure can be seen in figure 8. We find that the WSJTM outperforms our method at low TPR and FPR rates. However the results sharply fall off as it is incapable of segmenting a range of anomalies from the challenging PETS-2007 data. We observe that our method, over all PETS data is capable of detecting a greater number of true positive anomalies, however at the cost of introducing false positives.

4. Evaluation

In this study we endeavoured to evaluate whether social and scene region contextual information improves the detection of behavioural anomalies and permits the detection of subtle behavioural anomalies. The final true positive and false positive classification results with the inclusion of both types of context are affected by three factors above the no-context baseline. Firstly the inclusion of scene context which will either better segment anomalies or provide false positives, the inclusion of social context which will similarly either improve or worsen the TPR rate and FPR, and impact of propagating anomalies through a social and denying self justifying social groups. We can isolate the three impacts for each dataset in the above results.

In the three PETS-2007 datasets we observe that the addition of scene context improves the true positive rate (TPR) over false positive rate (FPR) detection of anomalies over all datasets in comparison to the no-context baseline. This is most significantly observed in scene 04, Fig 5(c). The inclusion of social context alone into the PETS-2007 data demonstrates a reduction in anomaly detection capacity in two of the three PETS data, Scene 02 and Scene 04, Fig 4(a) and (c). PETS-2007 Scene 02 shows a minor improvement. We found that the inclusion of social context information did provide some additional separation between true positives and true negatives however it also generated a greater number of false positives. However the significant result is that with the inclusion of both social context and scene context the TPR is improved above the TPR of scene context inclusion alone. This is due to the inclusion of the capability introduced by the social context to deny self justifying groups and propagate anomalies within social groups. Particularly in PETS Scene 04, we observe that by propagating low likelihood scores throughout the group the bulk of true positive anomalies are discovered earlier, see Fig 6(c).

We observe very little distinction between the social context segmentation and the scene context segmentation of the Oxford data set. Due to the highly structured nature of the Oxford data there are several false positive social connections, in a number of cases large social groups are falsely identified, see Fig 4. The false positive connections in the Oxford data has not adversely affected use of social context in the nearest neighbour anomaly calculation. However, the inclusion of denying self justifying groups, and propagating anomalies through social groups has a large negative impact. The impact in the Oxford data of denying self justifying groups in the presence of false positive social groups is to remove potential training data, thus increasing the probability of false positive anomaly alarms. The impact of implementing the propagation of anomalies through social groups in the presence of false positive social connections is that false positive anomalies propagate through the data. The degree to which this happens reflects the large false positive social groups detected in the Oxford data. We observe this failure mode in the Oxford data, see Fig 7. This result reflects our original prediction that our social model, geared towards crowds, would present a failure mode in the highly structured motion of Oxford data.

In the PETS-2007 data anomalies such as loitering are subtle behavioural anomalies. The motion trajectories of these behaviours are very similar to a large number of legitimate behaviours in the scene, particular in the queuing areas. Because motion alone is not sufficient to define the behaviour as an anomaly we require extra contextual information to segment these subtle behaviours from the main body of data. In the 3 PETS scenes we evaluated the scene context information played a large role in segmenting out these subtle abnormal behaviours.

5. Conclusion

We successfully demonstrated the capability to detect anomalies based upon contextual information and trajectories of humans upon two surveillance scenes. The Oxford data and PETS-2007 data presented distinctly different behavioural environments. Our system learned a scene model, classified social connections, and applied that contextual knowledge to the detection of abnormal behaviour observations. The application of social context provides a strong improvement in anomaly detection in the crowded PETS-2007 data. However failure of the social model can result in strong negative impact upon anomaly detection, as we witnessed in the Oxford dataset. We conclude that
in a crowded scene the application of social context to prevent self justifying groups and propagate anomalies is highly relevant - granting a greater anomaly detection capability. Scene context uniformly improved the detection of anomalies in both datasets. This validates the current state of the art in which scene understanding has gained much ground. Denying abnormal behavioural homogeneous social groups and propagating anomalies throughout an abnormal social group was shown to improve the anomaly detection capability. Contextual information of this nature may well be very applicable to a large portion of surveillance tasks, particularly those in crowded environments. The metric for comparing behaviours in this work can be interchanged with other state of the art methods, the implication being that contextual information, particularly scene regions could be complimentary implemented with other anomaly detection systems revealing subtle anomalies that otherwise may be missed.

References


Figure 9: Illustrated here is three examples of anomalies detected by our system in the PETS 2007 data set. (a) shows two true positives with a false positive in the bottom left corner. The anomalies in (a) refer to anomaly Id: 6 and 7 in table . In (b) two examples of loitering are detected, anomaly Id: 11 and 12. In (c) loitering is detected, Anomaly Id: 9, and 10.