Contextual anomaly detection in crowded surveillance scenes

Leach, Michael; Sparks, Ed; Robertson, Neil

Published in:
Pattern Recognition Letters

DOI:
10.1016/j.patrec.2013.11.018

Publication date:
2014

Contextual Anomaly Detection in Crowded Surveillance Scenes

Michael.J.V.Leach[a,1], Ed.P.Sparks[b], Neil.M.Robertson[a]

[a] *Heriot-Watt University, Edinburgh Campus, Edinburgh, Scotland*
[b] *Roke Manor Research, Romsey, Hampshire, United Kingdom*

Abstract

This work addresses the problem of detecting human behavioural anomalies in crowded surveillance environments. We focus in particular on the problem of detecting subtle anomalies in a behaviourally heterogeneous surveillance scene. To reach this goal we implement a novel unsupervised context-aware process. We propose and evaluate a method of utilising social context and scene context to improve behaviour analysis. We find that in a crowded scene the application of Mutual Information based social context permits the ability to prevent self-justifying groups and propagate anomalies in a social network, granting a greater anomaly detection capability. Scene context uniformly improves the detection of anomalies in both datasets. The strength of our contextual features is demonstrated by the detection of subtly abnormal behaviours, 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 and

Email addresses: michael.leach@roke.co.uk (Michael.J.V.Leach), ed.sparks@roke.co.uk (Ed.P.Sparks), n.m.robertson@hw.ac.uk (Neil.M.Robertson)

[1] Corresponding Author at all stages of refereeing and publication, also post-publication: Michael Leach. +44 1794 833937

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 [7]. Due to advances in pedestrian detection and robust tracking long term human centred tracks are becoming more prevalent [18, 10]. It is becoming plausible to autonomously profile the behaviour of a single, or multiple, humans over time. An abnormal event in automated surveillance is one which has a low statistical representation in the training data [4]. Our approach is motivated by this definition with an emphasis upon contextual information as a method of creating 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 directly 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 [7] and Social Signal Processing [5] we demonstrate the extraction and use of high level surveillance information which provides a *contextual* basis to identify subtly abnormal behaviour. 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 thus behaviour in one context may not be representative of behaviour in a different context. In any non-trivial surveillance scene contextual information such as scene region, social context, periodic events, and entry or

exit points impact the dynamics of behaviour [13]. We can use this contextual information to provide further means of segmenting abnormal behaviours from the mass of data, and perhaps provide the means to segment subtle behaviours from the mass of data. For a more general discussion on contextual anomaly detection see [2, 16].

With this work we demonstrate the significance of inferring social links between people in a surveillance application. We provide further validation of the growing trend in automatic scene understanding, additionally providing a novel approach. Furthermore we demonstrate a novel social context based anomaly detection procedure. We evaluate our systems capability to detect *subtle* behavioural anomalies within a complex and crowded human surveillance scene. Our main contributions are a novel method of acquiring scene structure information in surveillance, the development of a novel mutual information social group metric, and the demonstration that social and scene contextual information is effective in combination at anomaly detection.

1.1. Related Work

We focus upon social and scene region contextual knowledge as a means of improving the detection of subtle behavioural anomalies. The scene regions provides an understanding of portions of the scene in which we would expect normal behaviours to be different from other areas [7]. Previous approaches such as Li et al. develop a scene segmentation method which divides the scene into regions based upon behavioural dissimilarity [3]. Similarly, Chen Loy segments a scene into spatial regions of similar behaviour by virtue of behaviour correlation [4]. 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 [17]. The second source of contextual information we use is social context. Social Context 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. To estimate social groupings Ge et al. uses a proximity and velocity metric to associate individuals into pairs, iteratively adding additional individuals to groups using the Hausdorff distance as a measure of closeness [15]. Yu et al. implements a graph cuts based system which uses the feature of proximity alone [14]. 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. 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 [9]. Certain actions are declared group activities and thus groups can be constructed from individuals via mutual engagement in a grouping action. Robertson and Reid utilise gaze direction in order to determine whether individuals are within each other's field of view [8]. Gaze direction is significant as it departs from the use of motion features alone by taking into account visual interest [6]. For a comprehensive and complete review of the emerging field of social signal processing see the work of Cristani [5].

2. Method

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, and tracking of targeted pedestrians. Detection is achieved using the Felzenszwalb part based detector [10]. Tracking of human targets in the image plane is achieved with the use of the Predator TLD tracker [18]. We track the heads of pedestrians in the crowded PETS-2007 scene, see Figure 1 (a). for the second dataset, the Oxford data, we use the published tracking results provided by Benfold [1]. We select the TLD tracker due to high performance amongst state of the art trackers [19] and utilise its capability to learn a target model and discriminate between potential targets in a crowded surveillance scene. The pedestrian tracking performance of the TLD tracker is extensively tested against alternative recent tracking procedures in the author's paper [19].

Scene Context: Building upon the work of Makris [7] 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 direction, 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:

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

Where θ 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 $\bar{N}$ 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.

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

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 $\bar{T}$ the mean number of frames for all regions. The speeds of trajectories observed in location x, y is denoted by histogram 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 releases them allowing them to diverge. Typically such regions are appended to the ends of a traffic lane.

$$C_{xy} = \frac{\frac{1}{\pi} \sum \sqrt{(\theta_{xy} - \bar{\theta}_{xy})^2}}{-\sum P(\theta_{xy}) log(P(\theta_{xy})} \tag{3}$$

Where θ 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. 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.

Social Context The basis of our social model is the premise that a high degree of shared trajectory information implies a social dependence between two individuals. Our social model is geared towards effective detection of social groups in a moving crowd. 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. For an illustration of typical social pairs see Figure 2 (b).

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 two individuals: the mutual information of direction ($I\Theta_{ijt}$), the mutual information of speed (IV_{ijt}), the proximity between two individuals (ΔP_{ijt}) and the temporal overlap ratio between two individuals (τ_{ijt}). We train a set of weighting variables $\alpha_{\Delta P}, \alpha_{IV}, \alpha_{I\Theta}, \alpha_{\tau}$ which weight each feature in the social metric based upon the classification score of each feature independently on the ground truth training data. The feature weights are distributed proportional to each features classification score. The features which compose the pairing metric are defined

as:

$$\Delta P_{ijt} = \alpha_{\Delta P} e^{-\frac{\frac{1}{N}\sum_n |S_{it} - S_{nt}| + \frac{1}{N}\sum_n |S_{jt} - S_{nt}|}{2\ |S_{it} - S_{jt}|}} \tag{4}$$

For 2 tracked individuals i and j at frame t where S_{ij} is the distance between trajectory i and j at time t. The proximity between any two individuals ΔP is scaled by the distance between i and j to the set of all other individuals N 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.

$$\Delta \tau_{ijt} = \alpha_T e^{-\frac{|T_i - T_j|}{2T_{ij}}} \tag{5}$$

Where τ_{ijt} 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_{ij} is the number of frames in which both i and j have coexisted.

Whilst ΔP_{ijt} and $\Delta \tau_{ijt}$ are direct measures of trajectory statistics it is important to note that both $IV_{ijt}, I\Theta_{ijt}$ are more complex in nature. We use mutual information (MI) instead of the Euclidean distance as it handles non-linear and non-Gaussian random variables effectively and provides a principled method of comparing orthogonal feature dimensions. We define the Gaussian distributions of speed $P(v)$ and direction $P(\theta)$ as the Maximum Likelihood Estimation (MLE) derived from 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.

$$\begin{aligned} IV_{ijt} = &- \alpha_{IV} \sum_{b} P(v^i(b)) log_2(P(v^i(b))) \\ &- \alpha_{IV} \sum_{b} P(v^j(b)) log_2(P(v^j(b))) \\ &+ \alpha_{IV} \sum_{b} P(v^{ij}(b)) log_2(P(v^{ij}(b))) \end{aligned} \tag{6}$$

Where v^i is the MLE distribution over speed for person i over the most recent time window. The mutual information calculation for direction $I\Theta_{ijt}$ is structured identically to the above, replacing the MLE speed distribution v^i with the MLE direction distribution θ^i.

Each feature is used independently to classify pair connections between tracked individuals and scored with against the ground truth classification. We observed that the features of proximity between two individuals (ΔP) and the temporal overlap ratio between two individuals (T_{ijt}) present a significant ability to classify pairs in the test data. The overall performance is improved with the inclusion of the mutual information measures for direction and speed, see Figure 3. Whilst the individual features of mutual information speed and direction provide better classification we find there is a lack of correlation with the true positives exemplified by the Euclidean features of proximity and temporal overlap in this dataset. In this dataset the impact is a slightly reduced true positive rate. However we select the mutual information metric over Euclidean distance as it is a more principled method and scores better than the Euclidean features.

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, θ) for 2D ground plane position vector S, speed scalar v and direction of trajectory in radians θ. 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 κ between two individuals i and j at time T is:

$$\kappa_{ijt} = \frac{1}{T} \sum_{t}^{T} IV_{ijt} I\Theta_{ijt} \Delta P_{ijt} \tau_{ijt} \tag{7}$$

$\tau_{ijt}, IV_{ijt}, I\Theta_{ijt}, \Delta P_{ijt}$ 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). We classify the social state S, for $S = \{0, 1\}$, by applying social strength threshold λ which is set empirically from the training data. Connections between individuals which score higher than λ are considered socially connected, providing the binary social context state used in the anomaly detection stage.

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 is represented by a four part feature vector $x = \mathfrak{R}^4$, 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. 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: As our approach is unsupervised anomalies are discovered due to their contrasting nature to previously observed behaviour. Much work to date has focused upon a frequency based analysis to determine the normality of behaviour observations. 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 distance to the nearest K instances of supporting evidence *n*ot the frequency of observation for that behaviour.

Whilst a nearest neighbour approach could be expected to segment out anomalies with strong contrary motions, a subtle anomaly may not be distant from the set of normal behaviour with

regard to the majority of features. A subtle anomaly may be abnormal for only a subset of features, and furthermore only 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, 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. To represent each feature in the context of another we reduce our 4D histogram feature space to a set of 1D feature distributions $Y_n^{f1,f2}$ detailing the distribution of feature f1 given the currently observed value for feature f2 for person n at frame t. For a feature vector x_i with dimensionality D there are $D^2 - D$ feature context pairs 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. To reduce the dimensionality of X_i to 1 for a particular feature context pair we sum the distribution X_i for all dimensions f in the set of dimensions F where $f_1 \neq f_2$ resulting in a 2D 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 $Y_n^{f1,f2}$ 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 $Y_n^{f1,f2}$ and the set of all distributions Y to determine the nearest neighbour $Y_m^{f1,f2}$ to $Y_n^{f1,f2}$ for each possible feature context pairing $\{f_1, f_2\} \in F$. The Nearest Neighbour distance metric specified is the Bhattacharyya co-

efficient. The nearest neighbour distance metric for feature context pair $\{f_1, f_2\}$ is thus defined as:

$$B(Y_n, Y_m) = \sum_h \sqrt{Y(h)_n^{f1,f2} Y(h)_m^{f1,f2}} \tag{8}$$

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 $\{m \in N : n \neq m\}$.

$$NN(Y_n) = \{Y_m \in Y | \forall Y_p \in Y : B(Y_n, Y_m) \geq B(Y_n, Y_p)\} \tag{9}$$

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 Bhattacharya 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 recombine the independent feature context pairs to generate a single value for the abnormality coefficient $A(n, t)$ 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 nearest neighbour:

$$A(n, t) = argmin_{f1,f2} B(Y_n^{f1,f2}, Y_m^{f1,f2}) \tag{10}$$

A consequence of segmenting subgroups is that an observation may be the only member of a context defined sub group. Ideally in operation an active learning methodology would be implemented to determine the normality of an observation in a new area of the behaviour space. However, in our application we chose to suspend judgment of new instances of behaviour, specifying that no evidence of an alarm is not an alarm. It would be equally valid to select the opposite,

the effect of which would be to place a bias upon highlighting rare behaviour.

Anomaly detection: Threshold μ upon $A(n,t)$ separates anomalies from normal observations and in effect represents the sensitivity of the system. If we seek to detect only anomalies then μ represents the expectation of abnormal behaviour in the sequence. For the end user μ represents a constant surveillance workload for the operator. Variable μ 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:

$$A(n,t) = \delta(A(n,t)) = \begin{cases} 1, & A(n,t) < \mu \\ 0, & A(n,t) \geq \mu \end{cases} \tag{11}$$

Based upon the assumption that there is dependence between the behaviour of individuals within the same social group we utilise the social contextual information in an additional two ways. Firstly we ensure that the behaviour of each individual is only analysed in reference to people external to their social group. Thus a behaviourally homogeneous group of individuals all acting abnormally cannot be self-justifying. We enforce this by 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

We wish 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 which is itself designed to detect subtle anomalies.

The publicly available PETS 2007 dataset [11] offers a source of multi camera real world surveillance footage. The datasets consists of 8 sequences each captured from 4 different viewpoints. We consider the PETS 2007 data to be a crowded scene. The 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. Behavioural anomalies in this dataset are characterised by strong motion abnormality such as a group running across part of the scene, or subtle anomalies such as a single individual standing still in a busy area, or a group loitering amongst a crowd. We specifically chose this data due to its behavioural complexity for anomaly detection. The second dataset selected is the Oxford dataset. 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 consider this data as sparsely populated. The trajectory motion in the Oxford data is far more structured; the vast majority of individuals travel at walking pace in one of two directions. We select the second dataset, the Oxford data, to test our social context approach for failure modes. In the Oxford data the trajectories of socially unconnected pedestrians 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. We evaluate upon 3 non-sequential videos from the PETS 2007 selected due to the ground truth behaviour

abnormalities present. PETS Scene 02 consists of 4500 images, Scene 04 is 3500 images long, and Scene 07 is 3000 images in length. All three are imaged at 25fps. The single scene from the Oxford dataset is captured at 25fps and 4500 frames in length. each sequence is treated individually. We apply the tracking procedure outlined earlier upon the jpeg the format images with no other pre-processing.

Scene Segmentation 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. For clarity we illustrate the scene segmentation, see Figure 4. 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. As our approach is data driven, scene regions are defined by virtue of being a tool for segmenting the behaviour space rather than fitting an intuitive interpretation of scene regions.

Social Context We test the social context classification against an 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. Classifying social connections in the PETS 2007 data using parameters trained in the PETS 2006 data achieved a true positive detection rate (TPR) of 0.92 and a false positive rate (FPR) of 0.092, see Figure 3 (a). There are a greater number of false positive social connections in the Oxford data. The optimal result found 0.412 TPR and 0.0149 FPR. However beyond this true positive rate the false positives escalated greatly.

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

TPR and FPR, for detection of groundtruth anomalies. See Table 1 for a full list of anomalies. For examples of subtle anomaly detection see Figure 5. The anomaly ground truth reveals 12 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 μ threshold from 0 to 1 in small increments to adjusts the systems sensitivity to unlikely observation. Figure 6 (a) (b) and (c) demonstrates the anomaly detection success in the PETS 2007 dataset. Figure 7 illustrates the results on the Oxford data.

4. Evaluation

The final TPR and FPR 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, the inclusion of social context, and impact of propagating anomalies through a social group and denying self-justifying social groups. In the three PETS-2007 datasets we observe that the addition of scene context improves the TPR over FPR detection of anomalies over all datasets in comparison to the no-context baseline. This is most significantly observed in Scene 04, Figure 6 (c). The inclusion of social context alone into the PETS-2007 data demonstrates a reduction in anomaly detection capacity in Scene 02, Figure 6 (c). PETS-2007 Scene 02 shows only a minor improvement. 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, reducing the FPR from 0.2 to 0.03, see Figure 6 (c). The overall classification score with both social and scene context for all PETS-2007 data is shown in Figure 8. We

recorded a drop in the false positive rate of 0.13 for the optimal classification rate of 0.78 when applying the social and scene context.

In the Oxford data set the use of context information does not appear to raise the ability to detect anomalies significantly. We believe this to be due to the highly structured simple nature of the Oxford data. There is in effect very little contextual information to leverage our method upon. The false positive social connections in the Oxford data has not adversely affected use of social context, however, the inclusion of denying self-justifying groups, and propagating anomalies through social groups has a notable negative impact. The impact 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. We observe this failure mode in the Oxford data, see Figure 7 which reflects our original prediction that our social model, geared towards crowds, would present a failure mode in the highly structured motion of Oxford data. To further test our approach we applied our context aware algorithm to maritime AIS shipping data in Southampton Harbour. The social context depicted mutual dependencies such as tugs pulling ships and convoy behaviour. Scene context was directly comparable. We achieved a true positive anomaly detection rate of 0.98 with a false positive rate of 0.17 over 66 hours of data. However as the focus of our approach is computer vision we do not discuss the results further in this work.

In the PETS-2007 data anomalies such as loitering are subtle behavioural anomalies as the 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, particularly the scene context. The output of our system is displayed in Figure 5. Images (a) through (c) show correct identification of anomalies. Image (a) shows an example of a context independent anomaly: running through the scene. Image

(b) shows two examples of context dependent anomalies. The motion features pertaining to the anomaly are common within the entire scene, requiring scene context for them to be detected as anomalies.

To see 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. We select the WSJTM as it is designed specifically to detect *subtle* abnormal behaviour similar in style to our own work. Furthermore, it is based upon a different behaviour representation whilst its use of positional information makes it comparable to our scene contextual information. For a detailed account of this work see [12]. We use the code provided by the author to make the comparison. The results from our own 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. The WSJTM is capable of finding gross motion anomalies better than our method however it fails to detect subtle anomalies such as loitering. We observe that our method achieves a better overall TPR over FPR.

5. Conclusion

We successfully demonstrated the capability to detect anomalies based upon contextual information and trajectories in two scenes, presenting distinctly different behavioural environments. The application of social context provides a improvement in anomaly detection in the crowded PETS-2007 data. However, failure of the social model can result in a negative impact upon anomaly detection, as witnessed in the Oxford dataset. We found that our context aware method performs significantly better than the equivalent method without contextual information; reducing the false positive rate from 0.2 to 0.03. We show an overall true positive classification rate

of 0.78 over 0.19 false positives on the PETS-2007 data, a reduction in the false positive rate of 0.13 due to the inclusion of contextual information. We conclude that in a crowded scene the application of social context to prevent self-justifying groups and propagate anomalies is highly relevant. Scene context uniformly improved the detection of anomalies in both datasets, and provided the ability to detect subtle context dependent anomalies. 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 used with other anomaly detection systems revealing subtle anomalies that otherwise may be missed.

[1] B. Benfold, I. R., 2011. Stable multi-target tracking in real-time surveillance video.

[2] Chandola, V., 2009. Anomaly detection: A survey. ACM Computing Surveys.

[3] J. Li, J. L., Xiang, T., 2008. Scene segmentation for behaviour correlation. European Conference on Computer Vision.

[4] Loy, C. C., 2010. Activity understanding and unusual event detection in surveillance videos. Ph.D. thesis, queen Mary University of London.

[5] M. Cristani, R. Raghavendra, A. D. B. V. M., 2012. Human behavior analysis in video surveillance: a social signal processing perspective. Neurocomputing.

[6] M. Farenzena, A. Tavano, L. B. D. T., 2011. Social interactions by visual focus of attention in a three-dimensional environment. Expert Systems.

[7] Makris, D., Ellis, T., 2005. Learning semantic scene models from observing activity in visual surveillance. IEEE Transactions on Systems, Man, and Cybernetics.

[8] N. M. Robertson, I. D. R., 2011. Automatic reasoning about causal events in surveillance video. EURASIP Journal on Image and Video Processing.

[9] N. Oliver, B. R., Pentland, A., 1998. Statistical modelling of human interactions. CVPR Workshop on Interpretation of Visual Motion.

[10] P. Felzenszwalb, R. Girshick, D. M., Ramanan, D., 2010. Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32.

[11] PETS2007, 2012. http://www.cvg.rdg.ac.uk/pets2007/data.html. Accessed 24/09/2012.

[12] T. Hospedales, S. G., Xiang, T., 2011. Identifying rare and subtle behaviours: A weakly supervised joint topic

model. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[13] T. Lan, Y. Wang, W. Y. G. M., 2010. Beyond actions: Discriminative models for contextual group activities. Advances in Neural Information Processing Systems.

[14] T. Yu, S. lim, K. P., Krahnstoever, N., 2009. Monitoring, recognising and discovering social networks. IEEE Computer Vision and Pattern Recognition.

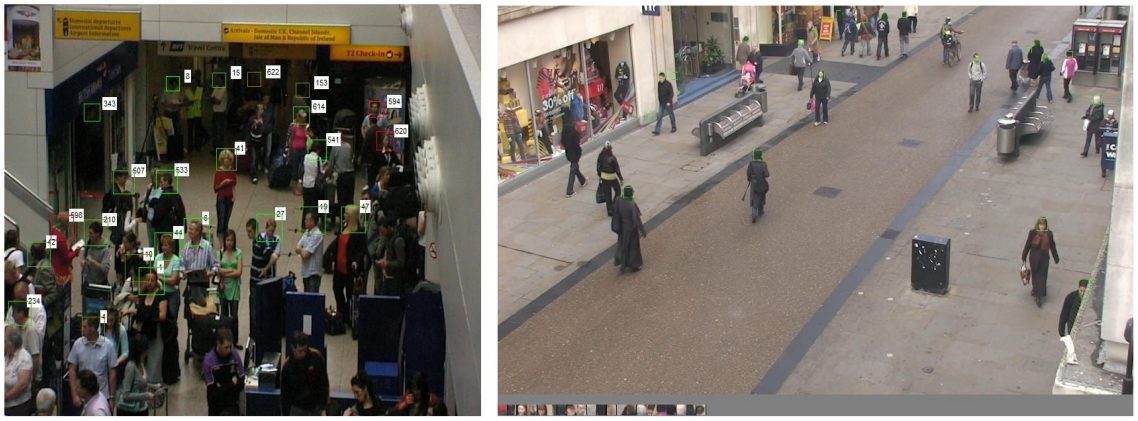
[15] W. Ge, R. T. C., Ruback, B., 2009. Automatically detecting the small group structure of a crowd. IEEE Workshop on Applications of Computer Vision.

[16] X. Song, M. Wu, C. J., Ranka, S., 2007. Conditional anomaly detection. IEEE Transactions on Knowledge and Data Engineering 19.

[17] X. Wang, K. Teck Ma, G. N. E. G., 2008. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. Computer Science and Artificial Intelligence Laboratory Technical Report.

[18] Z. Kalal, J. Matas, K. M., 2009. Online learning of robust object detectors during unstable tracking. 3rd Online Learning for Computer Vision Workshop, Kyoto, Japan, IEEE CS.

[19] Z. Kalal, J. Matas, K. M., 2010. P-n learning: Bootstrapping binary classifiers by structural constraints. 23rd IEEE Conference on Computer Vision and Pattern Recognition.



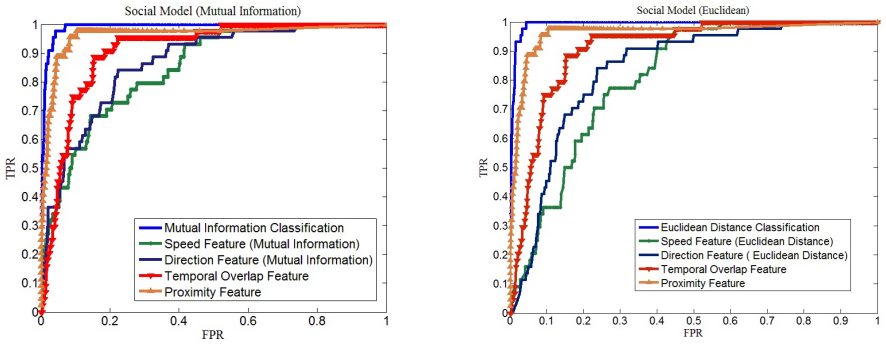
(a) (b)

Figure 1: We illustrate here the tracked dataset PETS-2007 (a), and tracked Oxford data (b). The PETS-2007 data presents a challenging crowded environment and contains far less structure in the apparent motion of individuals in the scene. In contrast the Oxford data contains very structured trajectory information, and is sparsely populated. Our social context extraction is geared towards crowded scenes such as the PETS-2007 data, however this presents a harder surveillance challenge.

(a) (b)

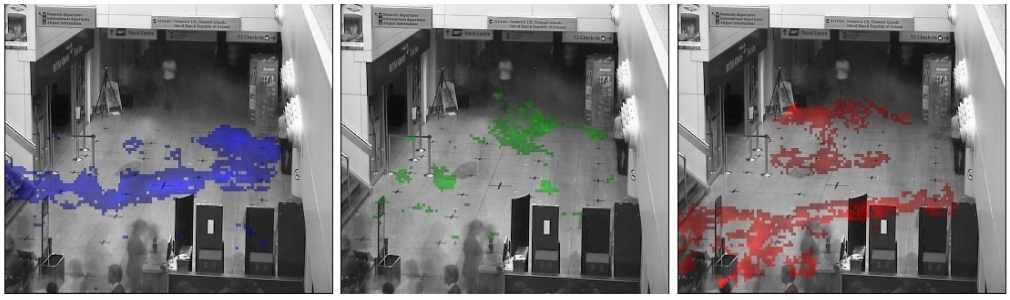
(c)

Figure 2: An example of social grouping from the Oxford data (a) and the PETS-2007 data Scene 04 (b) derived using our social connection strength metric. Both (a) and (b) show a true positive result. (c) demonstrates a failure mode.



(a) (b)

Figure 3: A comparison of the features which comprise the Mutual information social model (a) and for comparison the Euclidean distance equivalent (b) both trained upon the PETS 2006 dataset and tested upon the PETS 2007 data set. The proximity and temporal overlap in both metrics are identical. The critical difference is in the speed and direction information. We observe that the mutual information speed and direction metrics outperform the Euclidean distance feature metrics in overall true positive classification

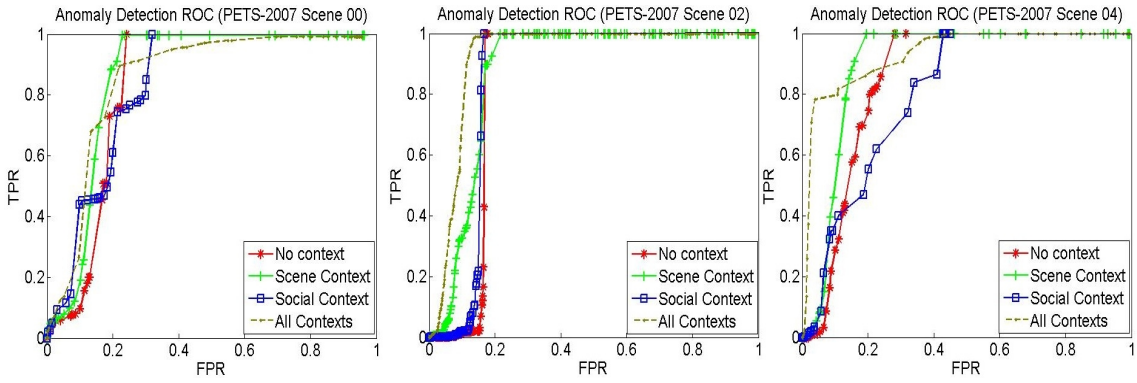


(a) (b) (c)

Figure 4: (a) (b) and (c) illustrate the automatic scene segmentation we arrived at using the all trajectories from the PETS-2007 datasets. Each unique scene context is designated by a colour; Idle region - Red, Traffic region - Blue, and Divergence region - Green. Areas of the scene not included in either scene region class do not have sufficient supporting evidence to be classified and as such remain blank.

(a) (b) (c)

Figure 5: 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 1. In (b) two examples of loitering are detected, anomaly Id: 11 and 12. In (c) loitering is detected, Anomaly Id: 9, and 10.



(a) (b) (c)

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.

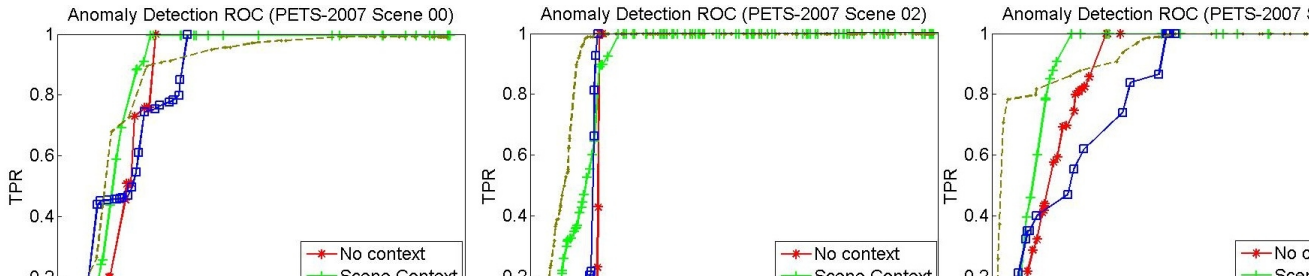


Figure 7: The anomaly detection results on the Oxford Dataset. we test upon the Oxford data to test for a failure mode in the social model.

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 the 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 baggage. 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.

PETS 2007 (Scene s00)	Id	Start	End
Unusual group behaviour	1	1	2656
Unusual group behaviour	2	1	2419
Unusual group behaviour	3	1	2714
Abrupt you turn in busy area	4	2627	2928
Abrupt you turn in busy area	5	2604	2928
PETS 2007 (Scene s02)			
ground-truth loitering	6	160	4497
PETS 2007 (Scene s04)			
Running through scene	6	109	275
Running through scene	7	130	290
Running through scene	8	148	322
Bag swap, unusual motion	9	1	3496
Bag swap, unusual motion	10	1	3496
ground-truth loitering	11	1	2596
ground-truth loitering	12	497	1726
Oxford Data	Id	Start	End
Motion + interaction with scene	13	3554	4349
Loitering	14	3867	4500
Abnormally slow movement	15	2382	3454

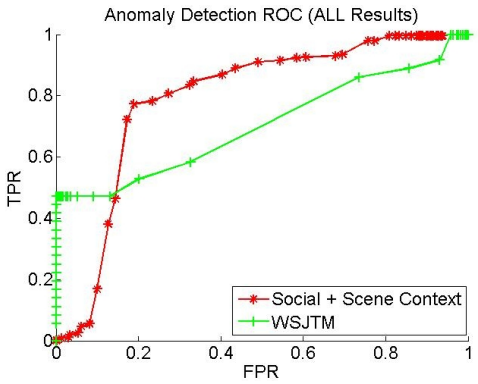


Figure 8: A comparison between the Weakly Supervised Joint Topic model and our context aware method on the challenging PETS-2007 dataset. We trained and tested against all PETS-2007 data for both datasets.