Robust abandoned object detection integrating wide area visual surveillance and social context

Article

Accepted Version

Robust abandoned object detection integrating wide area visual surveillance and social context

James Ferryman[a,*], David Hogg[b], Jan Sochman[b], Ardhendu Behera[b], José A. Rodriguez-Serrano[c], Simon Worgan[d], Longzhen Li[a], Valerie Leung[f], Murray Evans[a], Philippe Cornic[e], Stephane Herbin[e], Stefan Schlenger[g], Michael Dose[g],

[a] *Computational Vision Group, School of Systems Engineering, University of Reading, RG6 6AY, UK*
[b] *School of Computing, University of Leeds, UK*
[c] *Xerox Research Centre Europe, 6 Chemin de Maupertuis, 38240 Meylan, France*
[d] *formerly University of Leeds*
[e] *Department of Information Processing and Modelling, ONERA, BP 80100, 91123 Palaiseau Cedex, France*
[f] *MathWorks, Les Montalets, 2 rue de Paris, 92190 Meudon, France*
[g] *L-1 Identity Solutions, Universitaetsstr.160, 44801 Bochum, Germany*

Abstract

This paper presents a video surveillance framework that robustly and efficiently detects abandoned objects in surveillance scenes. The framework is based on a novel threat assessment algorithm which combines the concept of ownership with automatic understanding of social relations in order to infer abandonment of objects. Implementation is achieved through development of a logic-based inference engine based on Prolog. Threat detection performance is conducted by testing against a range of datasets describing realistic situations. The proposed system represents the approach employed in the EU SUBITO project (Surveillance of Unattended Baggage and the Identification and Tracking of the Owner).

Keywords:

☆This document is a collaborative effort.

*Corresponding author, email: james@computer.org

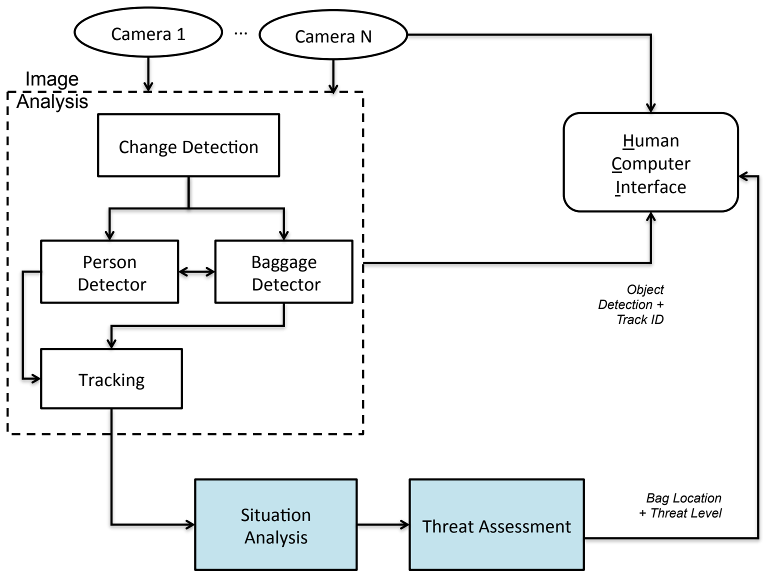


Figure 1: General framework of the automated threat detection system

wide area video surveillance, behaviour analysis, abandoned objects

1. Introduction

In recent years there have been a number of incidents where terror organisations have planted explosive devices in ordinary baggage to cause immense disruption in mass transportation networks and other areas of critical infrastructure. Due to the potentially devastating consequences of such terrorist activity, the monitoring and surveillance of unattended baggage has become a priority for the security operators of mass transportation networks and other critical infrastructure. The overriding goal is to minimise the number of false alarms. Towards this goal, the main contribution of this work is the development and evaluation of behaviour analysis methodology permitting robust identification of a baggage-owner while minimising false positives. The approach taken advances the state of the art in abandoned bag detection by introducing the concept of ownership and combines it with automatic understanding of social groups to infer abandonment. To achieve the goal, a framework (see Figure 1) has been developed consisting of a complete four-fold process, detection - tracking - situation analysis - threat assessment. This paper is divided as follows. Firstly, in Section 2 related research is detailed, followed in Sections 3-5 by descriptions of the system components. In Section 6 the datasets used and results of experiments are presented before concluding in Section 7 with conclusions and recommendations for future research.

2. Related Work

There exists a significant body of academic research addressing the task of robustly identifying abandoned baggage in public spaces. Most authors treat detection of abandoned (or left) objects, especially luggage, as the task of static object detection, with (Birch et al., 2011; Tian et al., 2010) or without (e.g. (Evangelio and Sikora, 2011; Porikli et al., 2008)) the application of tracking. Tian et al. (2010) present a framework to detect abandoned and removed scene objects based on background subtraction and foreground analysis, combined with tracking output to reduce false positives. Birch et al. (2011) employ motion segmentation based on a GMM with fast learning and a Motion History Image (MHI). For tracking of stationary objects, the edge map (3x3 Sobel filter) for each pixel is computed and matched) by correlation of edge directions. A comparative evaluation of stationary foreground detection algorithms based on background subtraction is given in Bayona et al. (2009).

There has been some attempt at human activity recognition and association to scene objects. In Lu et al. (2007) moving objects are tracked using shape and colour features and Kalman-based filtering, and classified using eigen features and Support Vector Machine. A package is defined as a non-human object and package ownership analysis performed using HMM-based human activity recognition.

2.1. Dataset Based Challenges

The most widely used datasets with which to evaluate approaches to abandoned bag detection have been from (PETS2007; PETS2006) and from the UK Home Office i-LIDS (2007). The dataset provided for the PETS2006 challenge consists of 7 multi-camera scenarios involving an increasing number of people and passers-by. Most of the submissions to PETS2006 were based on background subtraction combined with a blob tracker (Auvinet et al., 2006; Guler and Farrow, 2006; Krahnstoever et al., 2006; Li et al., 2006; Martínez-del-Rincín et al., 2006; Smith et al., 2006), with the exception of Lv et al. (2006) who rely on a more realistic human model by incorporating a human detector. Most often, when an object is not moving and its size is beneath a given threshold, it is assumed to be a standing bag. Smith et al. (2006) propose a probabilistic approach in which people and bags are classified based on the immediate history of their size and velocity. Another approach from PETS2006 is to use a slow-decay background model to detect stationary objects (Guler and Farrow, 2006). To be able to apply the PETS2006 rules for abandoned baggage (the owner is further than *a* metres for more than b seconds), the owner is usually defined as the nearest tracked object when the standing bag appears (Krahnstoever et al., 2006; Lv et al., 2006) or by examining blob splits during tracking (Auvinet et al., 2006; Guler and Farrow, 2006; Smith et al., 2006). When a standing bag and its owner are identified, it is straightforward to apply the PETS2006 abandoned-bag rules. The simplicity of the scenarios allows very limited situation awareness and was designed mainly to test if the low level processing stages are sufficient to cope with real-world scenarios.

The PETS2007 challenge focusses on two additional scenarios: theft and loitering. The videos are much more challenging from the tracking point of view as the scenes are more crowded. There are 8 scenarios, each viewed from 4 cameras. Two submissions to the challenge go beyond classical approaches to blob tracking and split-track analysis (such as (Arsic et al., 2007; Dalley et al., 2007)) and slowly/quickly adapting background models (such as Porikli and Yin (2007)). Firstly, Ribeiro et al. (2007) use a Temporal-JointBoost algorithm for each blob being tracked to classify it into a person-walking, not moving, a person picking-up/leaving a bag, or an abandoned bag. The basic idea is to incorporate temporal features (optical flow, motion energy) into the classification process over some temporal window. Secondly, Ardo and Astrom (2007) use an HMM to improve the temporal consistency of the tracking and show how to use an HMM efficiently in this setting. These approaches demonstrate the potential advantages of considering a longer temporal window for activity analysis. Nevertheless, the situation awareness in the PETS 2007 challenge is again very simple - reduced to comparing the distance of a bag to its owner (abandoned bag, theft) or measuring the time for which a person stays in the scene (loitering).

The UK Home Office have developed an image library (i-LIDS, 2007) to help researchers and designers to evaluate video based detection systems to meet Government requirements. The i-LIDS library includes an abandoned luggage dataset including several challenges of single instances of left luggage on a metro platform in the presence of passing passengers and trains. While the dataset is useful for evaluating detection algorithms it remains limited because it is monocular and also does not contain examples of specific behavioural interactions.

2.2. Limitations of Existing Approaches

It is clear that a global analysis of the situation rather than just examining each agent's behaviour independently, would be beneficial in many situations. The motivation for this is illustrated by a scenario similar to that of (PETS2007) where a family or a group of friends comes together and one of them leaves his/her bag with the others. Any threat detection system treating the individuals independently would inevitably report an abandoned bag, as the criteria specified in (PETS2006) that the bag is abandoned if the owner is further than a metres for more than b seconds, is fulfilled. For treating these more complex scenarios, the approaches described above may be insufficient and it may be necessary to derive a more complete activity analysis. A significant corpus of the computer vision and artificial intelligence literature attacks the problem of understanding activities from visual input. While logic and grammar-based representations, with or without combinination with stastistical approaches, (Hongeng et al., 2004; Ivanov and Bobick, 2000; Joo and Chellappa, 2006; Shet et al., 2005) organise knowledge in a flexible, powerful and clean way, one drawback of these approaches is that they are unable to propagate the uncertainty in the primitive detections. Hidden Markov Models (Brand et al. (1997)) and other flavours of dynamic Bayesian network provide a powerful generalisation of stochastic finite state automata to deal with such uncertainty. Another related approach is the so-called propagation network (Shi et al., 2004). In recent work, Damen and Hogg (Damen, 2012) first specify activities using a multiset attribute grammar and then convert it to an equivalent Bayesian network. A more general tool which converts first-order logic predicates into an equivalent Bayesian network is the framework of Markov logic networks (Richardson and Domingos, 2006), which have also been applied to activity analysis (Tran, 2008). An entirely different approach is to detect events from image pixels directly rather than by reasoning about the interactions between specific agents, for instance (Li, 2008; Wang, 2009). Whilst these approaches are easily configured to output whether an activity is normal or abnormal, they lack the explanatory power of grammar and logic-based methods (i.e. why it is abnormal).

None of the approaches described in the literature, however, have combined the concept of ownership with recognition of social groups, to reduce the number of false positives in detection of abandoned objects.

3. Object Detection and Tracking

The framework, shown in Figure 1, supports application of a range of object detectors and trackers including the POM person detection method of Berclaz et al. (2009) and tracking-by-detection of Breitenstein et al. (2011), both of which operate at low frame rates (2-4fps) or offline. While detection and tracking is not the main contribution of this paper, brief descriptions are given to methods which have been developed to permit the overall framework to operate online and with multiple cameras.

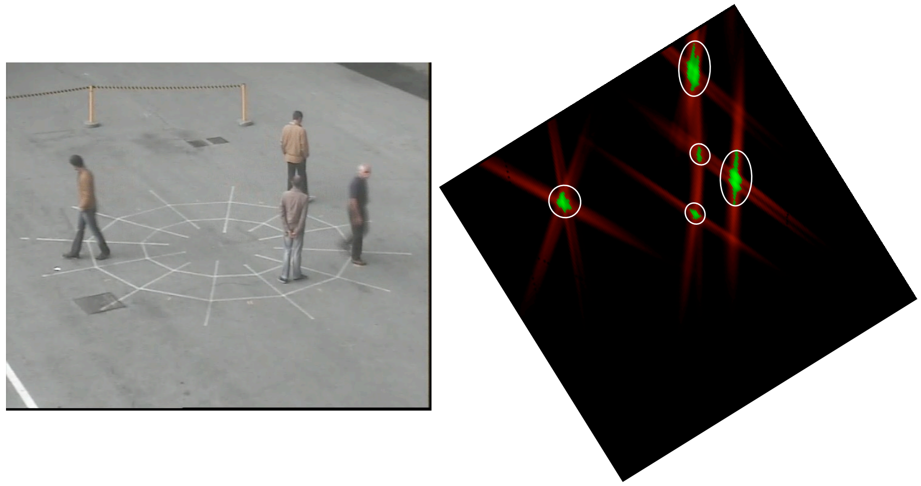
3.1. Baggage Detection

Baggage hypothesis generation is based on static change detection using the dual background approach of Porikli et al. (2008) adapted to use the efficient implementation of the Gaussian Mixture Model in Zivkovic (2004). Bag verification consists of application of a combination of filters including both 2D and 3D geometric filters and foreground/background similarity filter, and temporal filtering to check for peristence of the static regions.

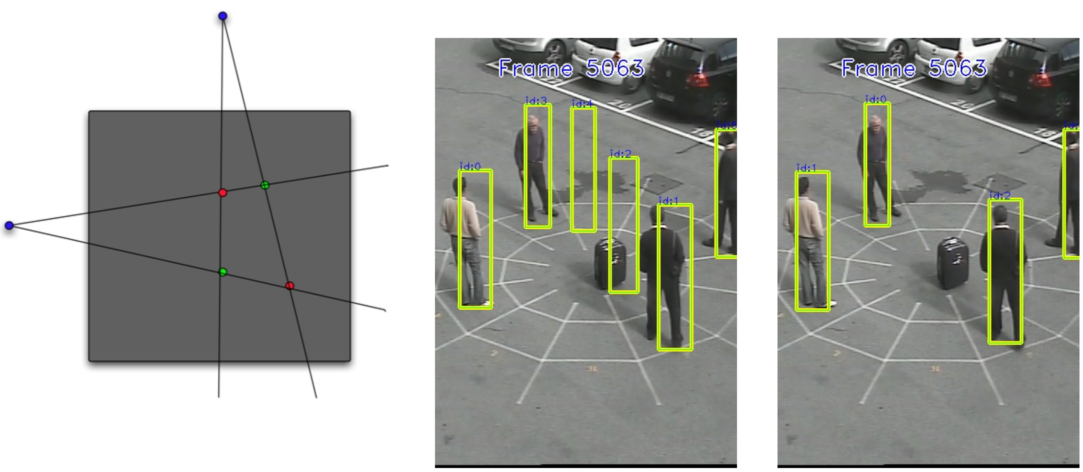
3.2. Person Detection

Person detection is based on the homography based multi-camera approach of Yildiz and Akgul (2010), extended with a novel approach for ghost suppression. First, a synergy map, the result of projecting detected foreground from each camera view to a single plane, is created, as shown in Figure 2. In practice, the reverse process is used with sampled cells on the synergy map, each corresponding to a vertical cuboid in space of fixed person height, back-projected to the bounded rectangles in the original images. The process is applied for an image resolution-limited ”infinite” number of planes in a very efficient and fully real-time manner without hardware acceleration.

For a given location (x, y) in the Synergy map (which corresponds to a small rectangular region on the ground plane), the value $S(x, y)$ accumulating the evidence of a person's presence can be calculated as:



(a)



(b) (c) (d)

Figure 2: Synergy map: (a). Detection of all pedestrians requires a threshold on synergy map to be set to value that permits ghost detection to pass thorough. (b). Ghost positions (red) can be predicted if correct positions (green) are known or can be estimated. (c-d). Bounding boxes resulting from detections without (c) and with (d) ghost prediction and suppression, for the same frame of video.

$$S(x,y) = \frac{1}{|I|}\sum_{i\in I}\frac{\sum_{u=u_0}^{u_1}\sum_{v=v_0}^{v_1}p(u,v,i)}{A(Z(x,y,i))} \tag{1}$$

where I is the set of images into which the cuboid can be visibly projected, $Z(x,y,i) = \{(u_0,v_0),(u_1,v_1)\}$ is the bounding box projection of the cuboid corresponding to a specific synergy map pixel (x,y) into image i as defined by two extreme corner points. $A(s)$ is a function to calculate the area of any shape s, and

$$p(u,v) = \begin{cases} 1, \text{if } I(u,v) \text{ is foreground} \\ 0, \text{otherwise} \end{cases} \tag{2}$$

Candidate objects are represented by peaks in the synergy map, obtained via thresholding. Ghost detections can occur where lines from different cameras to different objects intersect. To prevent ghosts becoming new tracking targets, a suppression map is generated in the regions of high ghost probability and subtracted from the synergy map. Frame-to-frame tracking of peaks further reinforces probable objects' location.

3.3. Tracking

A multi hypothesis tracker is used Blackman (2004) modified for application to tracking of extended objects. First, to handle short-term occlusions and the merging of measurements from different persons in the detection process, measurement-sharing between track hypotheses is allowed. This concept is illustrated in Figure 3 (Top). Secondly, the measurement-to-track association cost is modified to allow image features, specifically two hue-saturation histograms corresponding to the top and bottom halves of a person, to be used in addition to a simple Brownian motion model. Each model is updated using the Exponentially Weighted Moving Average (EWMA). The association score between a predicted state and a measurement is a product of the normalised histogram intersection distance between their histograms and the normalised Euclidean distance between their positions in 3D.

To overcome track fragmentations caused by long-term or complex patterns of interaction between people, long term tracking based on tracklet association is used. The approach is based on a Markov Logic Network (MLN) (Leung and Herbin (2011)) where the notion of a group to account for generic interaction between people is introduced. The scores for possible associations are calculated using both spatial-temporal constraints and appearance information. Associations are not only considered for tracklets that can be directly joined together; but are extended to tracklets separated by a group in space and time. It therefore handles the formation and splitting of groups, reducing track fragmentations and allowing longer tracks to be formed. Examples of the tracklet association rules are shown in Figure 3 (Middle) and example final tracking output in Figure 3 (Bottom).

4. Situation Analysis

Situation analysis is an intermediate step towards threat assessment and is defined as the description of the relationships between people and bags that can be inferred from the behaviour of the participating agents. This contribution focusses on two kinds of relationship: who owns each bag, and who knows who. The analysis takes object tracks and class information as input and describes the state of the world (i.e. the scene) in terms of the observed agents and their behaviour. The following stage (threat assessment)

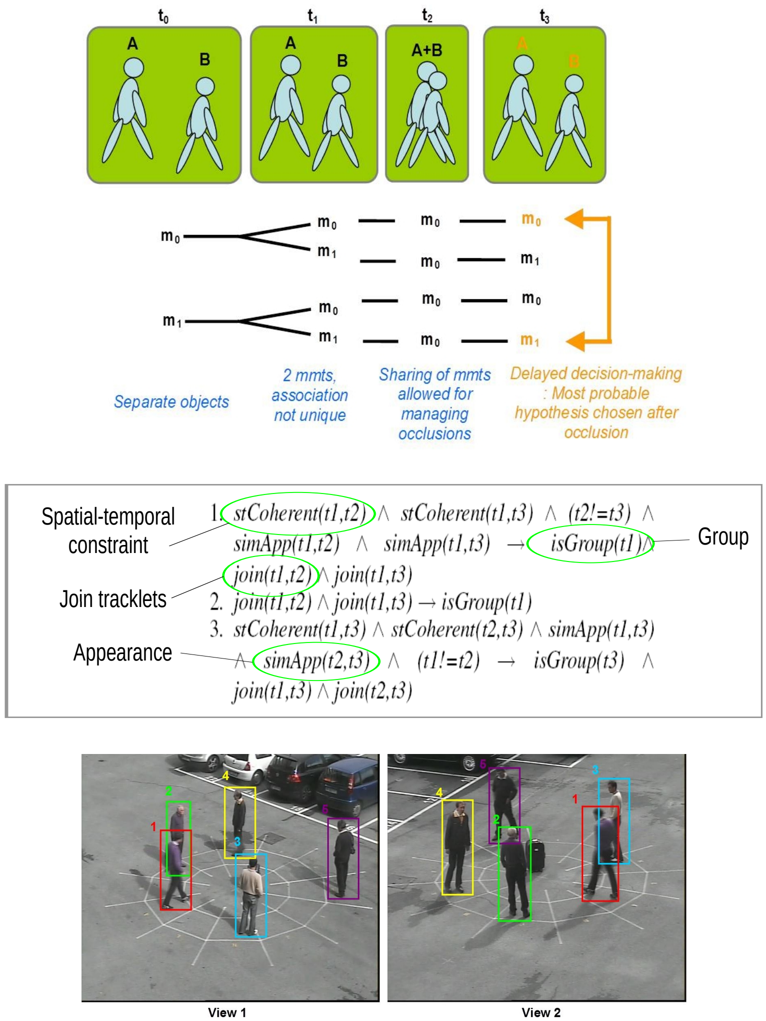


Figure 3: Tracking processes. Top: Illustrating how measurement-sharing in video-MHT overcomes short-term occlusions. Middle: Examples of tracklet association rules used in the MLN formalism. Spatial-temporal coherence and appearance information are used as inputs. The inference of groups and the joining of tracklets are two of the outputs. Bottom: Example tracking output for two cameras showing objects IDs.

determines whether the state of the world constitutes a possible threat (i.e. there is a truly abandoned bag.) The main contribution is the combination of the automatic understanding of social relationships with the concept of ownership to reduce the number of false alarms.

4.1. Bag Ownership

For the reported experiments in this paper, a bag is detected when it appears stationary in the scene, having been placed there by a person. At this stage, detection of a bag as it is carried into or out of the scene has not been incorporated. The ownership of each bag is inferred by simply looking for a person in the proximity of the bag over a fixed time interval prior to its appearance. The person is also required to be stationary at the time the bag-drop is hypothesised to occur. Specifically, in the experiments reported here, any person is assumed to be an owner if they are temporarily stationary within one metre of the bag at any point within one second prior to its appearance. Note that multiple possible owners are allowed, not because this is expected to be the case in reality but in order to reduce false alarms through taking both hypotheses through into the threat assessment.

4.2. Inference of Social Relations

Social groups are a very common phenomena in human crowds, with empirical studies suggesting that about 74% of people come in a group to a social event (Aveni (1977)) and about 50-70% (depending on the environment) are in a group during casual walking (Rudloff et al. (2011)). Despite this high percentage, the prevailing crowd behaviour models in todays simulation tools (Challenger et al. (2009)), computer graphics applications (Reynolds (1987)) and in particular in activity recognition and computer vision (PETS2006) are based on modelling each individual independently. An online algorithm has been developed for automatic detection of social groups within crowds, based on the analysis of the way the social relations influence the walking behaviour of the group members.

The method is based on the Social Force Model (SFM) (Helbing and Molnar, 1995; Moussaid et al., 2010) widely used in the crowd simulation community. In this, each individuals' movement is influenced by notional forces operating between individuals. Depending on whether two individuals (a) know each other or (b) do not know each other, the Social Force Model produces different sets of trajectories for these individuals. Until recently, these attempts were based on human designed forces without proper evaluation. Only recently, the model has been calibrated on real-world video sequences resulting in a model that realistically predicts avoidance behaviour of a walking group (Moussaid et al., 2009; Singh et al., 2009) and later in a model with all its parameters, including group behaviour, estimated from real data (Moussaid et al. (2010)).

The method employed in this work solves the inverse problem: knowing the trajectories, what are the social forces, and thus the relations, that caused that behaviour. The method is used in the framework to infer the social relations between the individuals in a scene and thereby to inform threat assessment as explained in Section 5.

The authors are aware of only two approaches aiming explicitly at social group inference (Ge et al., 2009; Jacques et al., 2007) and one paper using social groups to improve tracking (Pellegrini et al. (2010)). In Jacques et al.

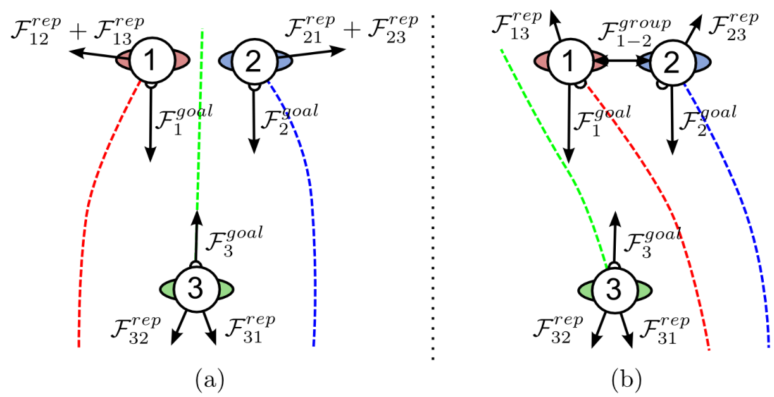


Figure 4: Depending on whether the individuals 1 and 2 (a) do not know each other or (b) know each other, the Social Force Model produces different sets of trajectories combining together repulsive ($\mathcal{F}^{rep}$), goal directed ($\mathcal{F}^{goal}$), and group ($\mathcal{F}^{group}$) forces influencing the individuals.

(2007) the groups are detected when two individuals keep close enough for a significant fraction of time over a given period. Experiments undertaken by the authors have shown that such simple measures are not sufficient for reliable group inference in complex scenes. In the proposed approach the calibrated SFM instead is relied upon. Similar measurements were used in Pellegrini et al. (2010) to improve tracking by jointly tracking and inferring the social groups.

Also based on distance, but including the difference in velocity as well as position, the method proposed in Ge et al. (2009) applies clustering to the (complete) person trajectories. The merging criterion takes into account the fraction of time in which the individuals are seen close to each other and allows the addition of a person to the group only if they have been close to at least half of its members. Figure 4 illustrates the Social Force Model. Full details of the approach are given in Sochman and Hogg (2011).

5. Threat Assessment

The threat assessment stage determines whether the inferred situation constitutes a threat, utilising the inferred knowledge of ownership and social relations described in Section 4. The mechanism adopted is sufficiently general to accommodate external information (e.g. the state of alert, time of day) alongside information on the observed scene in determining whether or not to raise an alarm.

Three increasingly sophisticated definitions are considered for what constitutes an abandoned bag. The first adopts the simple *baseline* definition that defined the PETS2006 challenge. In this, a threat (i.e. abandonment) is defined as follows:

- *Bag unattended if no person within 2 metres*
- *Bag abandoned if unattended for 30 seconds*

Here, the notions of ownership and social relationships are not used.

The second definition (*owner*) includes the notion of ownership (Section 4.1) and is defined as follows:

- *Bag unattended if owner is not within 2 metres*
- *Bag unattended if there is no assigned owner and if no person within 2 metres*
- *Bag abandoned if unattended for 30 seconds*

When there is no assigned owner, this is equivalent to the baseline definition, but where one or more possible owners have have been assigned, the condition for an alarm to be raised is less stringent since the behaviours of non-owners within the scene is ignored (unless there is no assigned owner).

The third definition (*owner+group*) includes both the notions of ownership (Section 4.1) and social relationships (Section 4.2). In this, a threat is defined as follows:

- *Bag unattended if owner or someone in the same social group as owner is not within 2 metres*
- *Bag unattended if there is no assigned owner and if no person within 2 metres*
- *Bag abandoned if unattended for 30 seconds*

This relaxes the *owner* definition in the direction of the *baseline* definition, since now the circle of people attending to a bag is widened to include people in the same group as the possible owner(s). The likelihood of raising an alarm is therefore reduced.

5.1. Implementation

The aim in threat assessment is to make it straightforward to encode the evolving state of the world and explore different behavioural patterns that constitute a potential threat. To achieve this, a simple logic-based inference system (Prolog) is adopted in which the current state of the world is represented by a set of facts and the behavioural patterns that constitute potential threats are encoded as rules.

The elements of this logic-based approach are:

- Facts (logical atoms), which are employed to describe situations. A fact is of the form R(A,B,...), where R indicates a type of relation between the elements inside the brackets.

- Rules, which are employed to infer new facts from existing ones.

Given these elements, the threat assessment proceeds in two steps:

1. Tracking and detection data are converted into a set of facts;
2. A set of pre-defined rules is invoked to infer additional facts.

The position of an object in each frame is represented by a unique ID for the object, it's class (person or bag), it's x,y position on the ground-plane and the frame number:

track(id,class,x,y,frame).

The social relationships between individuals are represented by a single predicate that records a unique group ID for each person. This partitions the set of people into social groups. Any person not assigned to a social group is assumed to be outside any group. This is represented simply by facts of the form:

group(id,group_id).

For convenience, a 'class' predicate is used (as in *class(id, person).*) to record the class of each object independently of the 'track' facts.

The ownership of bags is inferred next by a set of Prolog rules that embody the criteria described in Section 4.1. The result is a new set of facts, each representing the ownership of a bag (b) by a person (p):

owner(p, b).

Finally, the alarm condition for the chosen threat definition is posed as a Prolog query. As part of this, for the baseline definition, the condition that a bag is attended translates into the rule:

attended(B, T) :- class(P, person), nearby(P,T,B,T,2).

Here the rule states that a bag is attended at time T (shown on the left of the ':-') if it is owned by someone (call them P), and the position of P at time T is within 2 metres (i.e. nearby) of the position of B at time T (shown on the right of the ':-'). Upper case arguments are used to signify that these are variables.

The equivalent set of rules for the *owner+group* definition, incorporating the notions of ownership and social relationships, is as follows:

attended(B,T) :- owner(P,B), nearby(P,T,B,T,2), !.

attended(B,T) :-

\+owner(_,B), track(P,person,_,_,T), nearby(P,T,B,T,2).

attended(B,T) :- owner(P,B), knows(P,Q), nearby(Q,T,B,T,2), !.

knows(P,Q) :- group(P,G),group(Q,G).

The first rule states that a bag B is attended at time T if there is an owner P for the bag and this person is nearby. The second rule invokes the baseline notion of being attended when there is no owner - the meaning of '\+' before the owner predicate means that this isn't present in the database. The third rule states that a bag is attended (at time T) if there is a second person Q who is nearby the bag and P and Q know one another. The fourth rule implements the notion of two people knowing one another in terms of their group membership - i.e. they know one another if they are from the same social group. The *owner* definition, incorporating only the notion of ownership, is defined by the first two of the rules above.

Finally the condition for an alarm to be raised is the same for all three definitions - a bag must be unattended for a fixed period of time. The definition of 'unattended' is expressed in terms of the different definitions of attended, as follows:

```
unattended(B,T) :- class(B,bag), track(B,bag,_,_,T),
                   \+attended(B,T).
```

This states that an object is unattended at time T if it is a bag, it is in existence at time T, and there is no 'attended' fact in the database for that bag at time T.

Thus, only the definition of 'attended' varies between the three definitions of what constitutes an alarm.

Generally, Prolog was found to be a convenient way to represent definitions in a readily understood fashion, facilitating extension and experimentation. On the other hand, there are aspects of the inference mechanism in Prolog that require care - for example the use of the cut ('!') in two of the rules above is necessary to avoid the same alarm being raised multiple times.

6. Results

6.1. Datasets

Two different datasets are used to test the performance of the proposed algorithms, the publicly available PETS2006 (PETS2006) and the second produced during the SUBITO project specifically for this study. The PETS2006 dataset consists of ten sequences with increasing complexity of a staged aban-

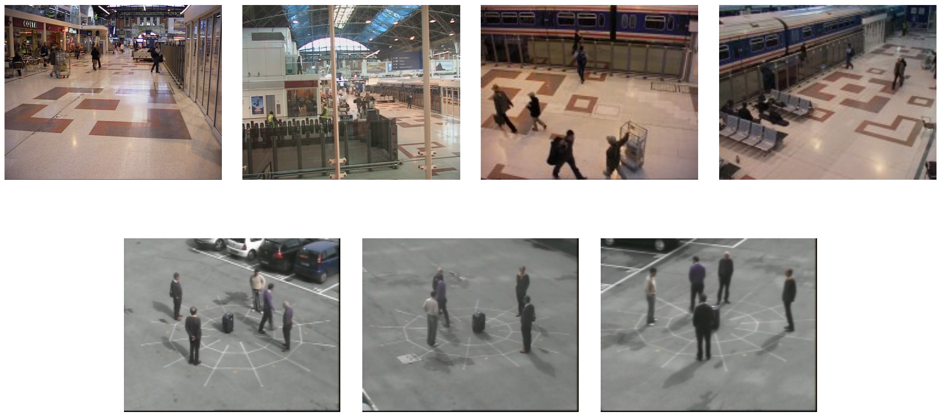


Figure 5: Datasets used. Top row: Four views from PETS2006 which contains scenarios with abandoned luggage. Bottom row: Three views from the SUBITO dataset describes scenarios where luggage owner enters the scene, sometimes interacts with other individuals and leaves the scene with/without the luggage.

doned bag scenario at a train station. All four camera views in the dataset were used in turn for the first four sequences used (PETS-S1-1, PETS-S1-2, PETS-S1-3 and PETS-S1-4), and camera view 3 used only for the other sequences (PETS-S2-3, PETS-S3-3, PETS-S4-3, PETS-S5-3, PETS-S6-3 and PETS-S7-3). The SUBITO dataset was recorded specifically for the SUBITO project. It contains thirteen sequences (19-22, 24-29, 31, 36, 37) each recorded from four synchronised cameras placed around the scene. In sequences 19-22 a single person brings a bag to a marked position and loiters around the bag (sequence 19), abandons the bag (sequence 20), or leaves the bag unattended for a while and then comes back (sequences 21, 22). Sequences 24-29, 31, 36 and 37 contain more challenging variants in terms of number of people and the group relationships. Each action is recorded 12

Table 1: Aggregate results across all SUBITO sequences comparing predicted alarms with corresponding baseline/owner/group ground truth.

Ruleset	TP	GTalarms	Alarms	Recall	Precision
baseline	16	71	35	0.23	0.46
owner	48	143	75	0.34	0.64
group	39	107	66	0.36	0.59

times for different entrance/exit directions. Depending on different threat definitions, the same action may or may not raise an alarm. Each sequence therefore should either correspond to 12 alarms (except for sequence 36 which only corresponds to 11 alarms), or none. The ground-truth alarms were obtained manually for all three threat definitions. The alarm time is determined by first visually deciding the very frame when the owner is just outside the prescribed distance from the bag, then adding a fixed time interval before the alarm is raised. Within the SUBITO dataset, the critical distance around a bag is assumed to be 2.5 metres (as opposed to 2 metres used in the PETS2006 challenge)- this assumption is therefore used in the three threat definitions. The time a bag must remain unattended to raise an alarm is reduced to 4 seconds.

6.2. Preliminary experiments on PETS2006 data

In the first experiments, the baseline functionality of (PETS2006) was implemented and evaluated. These experiments were carried out using an earlier version of the threat assessment logic implemented in C++. This was subsequently re-implemented in Prolog as part of the real-time system. To achieve this, the Prolog is queried for an alarm on every frame, based on

Table 2: Aggregate results across all SUBITO sequences comparing the use of all three threat definitions with the ground truth for the *owner+group* definition.

Ruleset	TP	GTalarms	Alarms	Recall	Precision
baseline	16	107	35	0.15	0.46
owner	42	107	75	0.39	0.56
group	39	107	66	0.36	0.59

Table 3: Aggregate results across all SUBITO sequences comparing the use of all three threat definitions with the ground truth for the *owner+group* definition with *stiched-together tracks*.

Ruleset	TP	GTalarms	Alarms	Recall	Precision
baseline	15	107	36	0.14	0.42
owner	43	107	94	0.40	0.46
group	41	107	88	0.38	0.47

the current state of the world and pertinent facts from the recent past. This world model is continually refreshed with the current location of each tracked object.

For the threat assessment to be correct, the system is required to raise an alarm following a potential threat, and to correctly identify the ID of the abandoned bag. Specifically, an alarm must be raised within 50 frames of a ground-truth alarm for it to be successful detected. The results on the PETS2006 dataset employ automatic tracking using an implementation of Breitenstein et al. (2011) and bag detection using Porikli et al. (2008). Alarms were raised correctly on all tested sequences except PETS-S4-3 and

PETS-S7-3. The failures on these two sequences were caused by individuals, having nothing to do with the abandoned bag, nevertheless being close enough to prevent the bag being classified as unattended. This result motivates the concept of ownership considered in the main set of experiments.

6.3. Experiments on SUBITO data

The main set of experiments were carried out on the challenging SUBITO dataset. The inverse SFM system is run in batch mode so that it has access to an entire sequence in predicting social groups rather than only the history up until the current time. The entire sequence is therefore used in inferring the set of alarms. This enabled evaluation of the interaction of the detection and tracking sub-system and the threat assessment sub-system, giving the inverted SFM the best chance of assigning correct social groups within relatively short scenarios. A single threshold in the inverse SFM system controls the propensity of pairs of individuals to be combined into the same group; a lower threshold results in larger social groups. For the SUBITO data, we found that both precision and recall reach their highest values within a small range of this threshold and the results we present are for a choice of threshold in this range.

The aggregate results across all SUBITO sequences are shown in Table 1, comparing predicted alarms with the corresponding ground-truth - that is baseline results are compared with the baseline ground-truth, etc. The aggregate results comparing the use of all three threat definitions with the ground-truth for the *owner+group* definition are shown in Table 2. As expected, the precision and recall for the *baseline* definition are lower in this case since the ground-truth reflects a more sophisticated notion of threat, incorporating concepts that are not present in the *baseline* definition. The evaluation reported here attended only to the time an alarm is raised and ignored the ID for the person and bag involved. Where there is more than one true positive alarm for a ground-truth alarm, this is counted once in computing recall and does not contribute to loss of precision. In other words, multiple predicted alarms for the same ground-truth alarm are counted only once. In general, there were few instances of this occurring in the experiment.

Within Table 2, there is a clear improvement in precision and recall between *baseline* and *owner* definitions. However, the comparison of performance between *owner* and *owner+group* definitions is less decisive. Here the recall has reduced slightly with the introduction of the social relationships, but there is a comparable improvement in precision. Looking in more detail at the results on individual sequences and alarms, several alarms have been surpressed by correct assignment of an owner and partner to the same social group. This is illustrated in Figure 6 showing a set of frames from SUBITO sequence 36. Two individuals (d:211, d:212) entering the scene (Figure 6 (top)) are assigned to the same social group (indicated by blue line between them), and one is detected as the owner of a bag (d:212) that appears within the scene (Figure 6 (middle). The owner subsequently goes away from the bag and outside the prescribed distance (shown as a green circle around the bag), leaving their partner attending to the bag (Figure 6 (bottom)). No alarm is raised.

In general the recall and precision are below acceptable performance for a deployed threat assessment system. The principal source of error arises from the highly challenging video sequences containing multiple overlapping

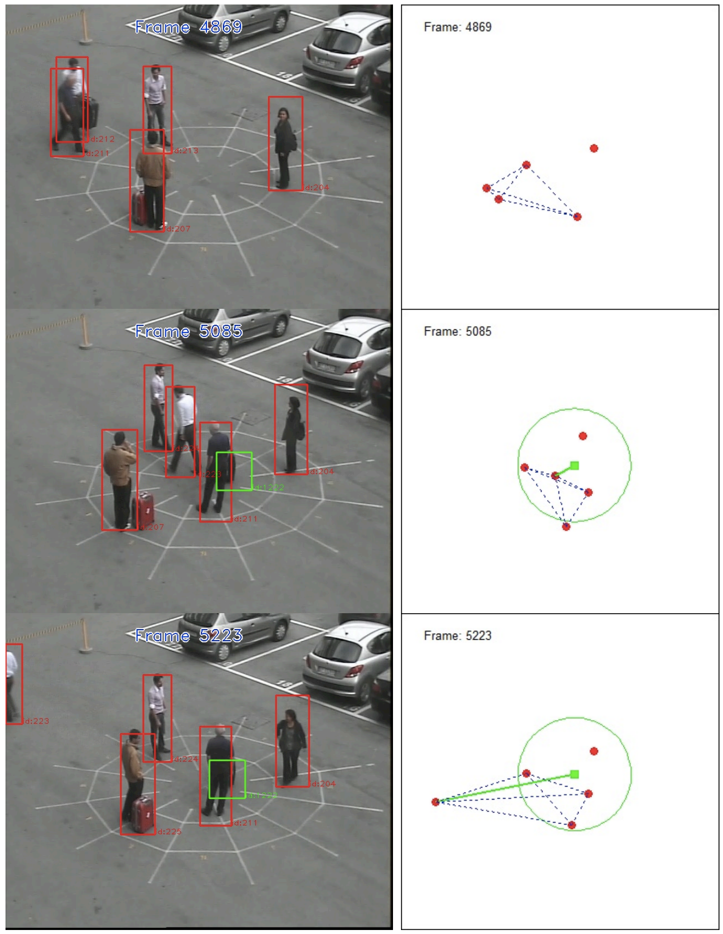


Figure 6: Social group analysis applied to SUBITO sequence 37 resulting in correct suppression of false alarm.

actors at any time. The consequential limitations in detection and tracking performance are translated directly into the threat assessments that can be achieved using the logic described above. Some improvement in performance was achieved by automatically stitching together tracks for which there is sufficient evidence that they belong to the same objects at different periods of time - specifically, one track (of more than 10 frames duration) ends within

4 seconds and 1 metre of another track (of more than 10 frames duration) beginning. The precision and recall for the equivalent evaluation to that in Table 2 is shown in Table 3. Finally, a real-time system that incorporates all stages of the pipeline, including on-line estimation of social groups up to the current frame, has also been implemented to demonstrate the practical viability of the method.

7. Conclusions and Future Work

This paper has described a video surveillance framework that detects abandoned objects in surveillance scenes containing multiple interacting individuals, extending the state of the art. Future work will address methods to further improve the underpinning object (person and bag) detection and tracking accuracy, as well as introduction of goal-directed and intentionality modelling strategies in the behavioural analysis.

There is scope to perform a more rigorous analysis of ownership through detecting bags being carried into the scene and hence identifying the owner more reliably. Similarly, confidence that a bag has been removed from the scene would be raised if it could be detected as it was carried out. There is prior work on this problem that should in principle be directly applicable to sequences such as those in the SUBITO dataset (e.g. Damen and Hogg (2008)).

Finally, expressing the the conditions of a threat in terms of logic, suggests that it may be possible to induce such conditions automatically from examples, thereby providing a way to incorporate different kinds of information about the scene without having to provide the logical rules by hand.

Earlier work on the use of inductive logic programming in video analysis indicates how this might be achieved in principle (Dubba (2010)).

Acknowledgements

This work was supported by EC projects SUBITO Grant Agreement No. 218004 and FP7-ICT-247022 MASH. Any opinions expressed in this paper do not necessarily reflect the views of the European Community. The Community is not liable for any use that may be made of the information contained herein.

H. Ardö and K. Aström., 2007. Multi Sensor Loitering Detection Using Online Viterbi. In Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1423-9.

D. Arsić, M. Hofmann, B. Schuller and G. Rigoll., 2007. Multi-Camera Person Tracking And Left Luggage Detection Applying Homographic Transformation. In Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1423-9.

E. Auvinet, E. Grossmann C. Rougier, M. Dahmane and J. Meunier., 2006. Left-Luggage Detection Using Homographies and Simple Heuristics. In Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1422-0.

A. F. Aveni., 1977. The Not-So-Lonely Crowd: Friendship Groups in Collective Behavior. Sociometry, 40(1):9699.

A. Bayona, J. C. SanMiguel, and J. M. Martínez., 2009. Comparative Evaluation of Stationary Foreground Object Detection Algorithms based on Background Subtraction Techniques. In Proc. of the 6th IEEE Int. Conference on Advanced Video and Signal Based Surveillance (AVSS 09), DOI:10.1109/AVSS.2009.35, pp. 2530.

J. Berclaz, A. Shahrokni, F. Fleuret, J. Ferryman and P. Fua., 2009. Evaluation of Probabilistic Occupancy Map People Detection for Surveillance Systems. In Proc. of the IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), pp 55-62, ISBN 978-07049-1501-4.

P. Birch, W. Hassan, N. Bangalore, R. Young and C. Chatwin., 2011. Stationary Traffic Monitor. In Proc. 4th Int. Conf. on Imaging for Crime Detection and Prevention (ICDP-11), DOI:10.1049/ic.2011.0128 , pp. 1-6.

S. S. Blackman., 2004. Multiple Hypothesis Tracking for Multiple Target Tracking. In IEEE Aerospace and Electronic Systems Magazine, 19(1), pp. 5-18.

M. Brand, N. Oliver and A. Pentland., 1997. Coupled Hidden Markov Models for Complex Action Recognition. In Proc. Computer Vision and Pattern Recognition, pp. 994-999.

M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier and L. Van Gool., 2011. Online Multiperson Tracking-by-Detection from a Single, Uncalibrated Camera. In IEEE Trans. on Pattern Analysis and Machine Intelligence, 33(9), pp. 1820-1833.

R. Challenger, C. W. Clegg, M. A. Robinson, and M. Leigh., 2009. Understanding Crowd Behaviours: Simulation Tools. Technical Report, University of Leeds.

G. Dalley, X. Wang, and W.E.L. Grimson., 2007. Event Detection Using an Attention-Based Tracker. In Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1423-9.

D. Damen and D. Hogg., 2008. Detecting Carried Bags in Short Video Sequences. In Proc. 10th European Conf. on Computer Vision, 5304:154-167.

D. Damen and D. Hogg., 2012. Explaining Activities as Consistent Groups of Events: A Bayesian Framework Using Attribute Multiset Grammars. In Int. Journal of Computer Vision, 98(1), pp. 83-102, DOI:10.1007/s11263-011-0497-0.

K. S. R. Dubba, A. G. Cohn, and D. C. Hogg., 2010. Even Model Learning From Complex Videos using ILP. In Proceedings ECAI, 215, pp. 93-98.

R. Evangelio and T. Sikora., 2011. Static Object Detection Based on a Dual Background Model and a Finite-State Machine. EURASIP Journal on Image and Video Processing, Article ID 858502, DOI:10.1155/2011/858502.

W. Ge, R. Collins, and B. Ruback., 2009. Automatically Detecting the Small Group Structure of a Crowd. In Workshop on Applications of Computer Vision, pp. 18.

S. Guler and M. K. Farrow., 2006. Abandoned Object Detection in Crowded Places. In Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1422-0.

D. Helbing and P. Molnar., 1995. Social Force Model for Pedestrian Dynamics. Physical Review E, 51:4282.

S. Hongeng, R. Nevatia and F. Brémond., 2004. Video-Based Event Recognition: Activity Representation and Probabilistic Recognition Methods. Computer Vision and Image Understanding, 96:129-162.

Imagery Library for Intelligent Detection Systems, *http://www.ilids.co.uk*. Last accessed: 29 January 2012.

Y. Ivanov and A. Bobick., 2000. Recognition of Visual Activities and Interactions by Stochastic Parsing. In IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(8), pp. 852-872.

J. Jacques, A. Braun, J. Soldera, S. Musse, and C. Jung., 2007. Understanding People Motion in Video Sequences using Voronoi Diagrams. Pattern Analysis and Applications, 10:321332.

S-W. Joo and R. Chellappa., 2006. Attribute Grammar-Based Event Recognition and Anomaly Detection. Proceedings Int. Workshop on Semantic Learning Applications in Multimedia, New York , NY June.

N. Krahnstoever, P. Tu, T. Sebastian, A. Perera and R. Collins., 2006. Multi-View Detection and Tracking of Travelers and Luggage in Mass Transit Environments. In Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1422-0.

V. Leung and S. Herbin., 2011. Flexible Tracklet Association for Complex Scenarios using a Markov Logic Network. In Proc. 11th Int. Workshop on Visual Surveillance, pp. 1870-1875.

L. Li, R. Luo, R. Ma, W. Huang and K. Leman., 2006. Evaluation of an IVS System for Abandoned Object Detection on PETS 2006 Datasets. In Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1422-0.

J. Li, S. Gong and T. Xiang., 2008. Global Behaviour Inference using Probabilistic Latent Semantic Analysis. In Proc. British Machine Vision Conf., DOI:10.5244/C.22.20.

S. Lu, J. Zhang and D. Feng., 2007. Detecting Unattended Packages through Human Activity Recognition and Object Association. Pattern Recognition, 8:2173-2184.

F. Lv, X. Song, B. Wu, V. K. Singh and R. Nevatia., 2006. Left-Luggage Detection using Bayesian Inference. In Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1422-0.

J. Martínez-del-Rincón, J. Herrero-Jaraba, J. Ral Gómez and C. Orrite-Uruñuela., 2006. Automatic Left Luggage Detection and Tracking Using Multi-Camera UKF. Proceedings IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1422-0.

M. Moussaid, D. Helbing, S. Garnier, A. Johansson, M. Combe, and G. Theraulaz., 2009. Experimental Study of The Behavioural Mechanisms Underlying Self-Organization in Human Crowds. In Proc. of the Royal Society B: Biological Sciences, 276(1668), pp. 27552762.

M. Moussaid, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz., 2010. The Walking Behaviour of Pedestrian Social Groups and Its Impact on Crowd Dynamics. PLoS ONE, 5(4):e10047, 04.

S. Pellegrini, A. Ess, and L. Van Gool., 2010. Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings. In Proc. European Conference on Computer Vision. 6311:452465, Springer Berlin / Heidelberg.

PETS2006. 2006. Ninth IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), New York, USA, June 18, ISBN 0-7049-1422-0, ISBN 0-7049-1422-0.

PETS2007. 2007. Tenth IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), Rio De Janeiro, October 14, ISBN 0-7049-1423-9.

F. Porikli and Z. Yin., 2006. Temporally Static Region Detection in Multi-Camera Systems. In Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1422-0, ISBN 0-7049-1422-0.

F. Porikli, Y. Ivanov and T. Haga., 2008. Robust Abandoned Object Detection using Dual Foregrounds, EURASIP Journal on Advances in Signal Processing, Article ID 197875.

C. W. Reynolds. 1987. Flocks, Herds and Schools: A Distributed Behavioral Model. In Proc. of the 14th Annual Conference on Computer Graphics and Interactive Techniques, pp. 2534, DOI:10.1145/37401.37406.

P. C. Ribeiro, P. Moreno and J. Santos-Victor., 2007. Detecting Luggage Related Behaviors Using a New Temporal Boost Algorithm. In Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1423-9.

M. Richardson and P. Domingos., 2006. Markov Logic Networks, Machine Learning, 62:107136.

C. Rudloff, T.Matyus, and S. See and D. Bauer., 2011. Can Walking Behaviour Be Predicted? An Analysis of the Calibration and Fit of Pedestrian Models. 90th Annual Meeting of the Transportation Research Board, January.

V. D. Shet, D. Harwood and L. S. David., 2005. Vidmap: Video Monitoring of Activity with Prolog. In Proc. IEEE Conference on Advance Video and Signal Based Surveillance, pp. 224-229.

Y. Shi, Y. Huang, D. Minnen, A. Bobick and I. Essa., 2004. Propagation Networks for Recognition of Partially Ordered Sequential Action. In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2:862-869, DOI:10.1109/CVPR.2004.1315255.

H. Singh, R. Arter, L. Dodd, P. Langston, E. Lester and J. Drury., 2009. Modelling Subgroup Behaviour in Crowd Dynamics Dem Simulation. Applied Mathematical Modelling, 33(12):44084423.

K. Smith, P. Quelhas and D. Gatica-Perez., 2006. Detecting Abandoned Luggage Items in a Public Space. In Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS), ISBN 0-7049-1422-0.

J. Sochman and D. C. Hogg., 2011. Who Knows Who - Inverting the Social Force Model for Finding Groups. In Proc. IEEE Intelligent Workshop on Socially Intelligent Surveillance and Monitoring.

Y. Tian, R. Feris, H. Liu, A. Humpapur, and M-T. Sun., 2010. Robust Detection of Abandoned and Removed Objects in Complex Surveillance Videos, In Proc. IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, pp. 112.

S. Tran and L. Davis., 2008. Visual Event Modelling and Recognition Using Markov Logic Networks. In Proc. European Conf. on Computer Vision, pp. 610-623.

X. Wang, X. Ma and E. Grimson., 2009. Unsupervised Activity Perception in Crowded and Complicated Scenes Using Hierarchical Bayesian Models. IEEE Trans. on Pattern Analysis and Machine Intelligence, 31(3):539-555.

A. Yildiz and Y. S. Akgul., 2010. A Fast Method for Tracking People with Multiple Cameras, Technical Report, Vision Lab, Gebze Institute of Technology.

Z. Zivkovic. 2004. Improved Adaptive Gaussian Mixture Model for Background Subtraction, In Proc. Int. Conf. on Pattern Recognition, 2, pp. 28-31.