PETS 2014: dataset and challenge

Conference or Workshop Item

Accepted Version

PETS 2014: Dataset and Challenge

Luis Patino and James Ferryman
University of Reading, Computational Vision Group
P.O. Box 225, Whiteknights, Reading RG6 6AY, United Kingdom
{j.l.patinovilchis, j.m.ferryman}@reading.ac.uk

Abstract

This paper describes the dataset and vision challenges that form part of the PETS 2014 workshop. The datasets are multisensor sequences containing different activities around a parked vehicle in a parking lot. The dataset scenarios were filmed from multiple cameras mounted on the vehicle itself and involve multiple actors. In PETS2014 workshop, 22 acted scenarios are provided of abnormal behaviour around the parked vehicle. The aim in PETS 2014 is to provide a standard benchmark that indicates how detection, tracking, abnormality and behaviour analysis systems perform against a common database. The dataset specifically addresses several vision challenges corresponding to different steps in a video understanding system: Low-Level Video Analysis (object detection and tracking), Mid-Level Video Analysis ('simple' event detection: the behaviour recognition of a single actor) and High-Level Video Analysis ('complex' event detection: the behaviour and interaction recognition of several actors).

1. Introduction

There is nowadays a significant amount of research achieved in the field of video surveillance. A large number of algorithms have been designed and tested for the tasks of object detection and tracking as well as for detection of events of interest, abnormalities or criminal behaviours. However it is still difficult to compare or evaluate such algorithms because of the lack of standard metrics and benchmarks that indicate how detection, tracking and threat analysis system perform against a common database. The goal of the PETS workshop has been to foster the emergence of computer vision technologies for detection and tracking by providing evaluation datasets and metrics that allow an accurate assessment and comparison of such methodologies. PETS 2014 is sponsored by the EU project ARENA[1]. ARENA addresses the design of a flexible surveillance system to enable situational awareness and determination of potential threats on mobile assets in transit. ARENA is making available for this workshop the 'ARENA Dataset': a series of multi-camera video recordings where the main subject is the detection and understanding of human behaviour around a parked vehicle; with a main focus on discriminating behaviour between normal, abnormal/rare behaviour, and real threats. The dataset presents different challenges covering object detection and tracking, abnormal event detection and behaviour understanding.

[1] www.arena-fp7.eu

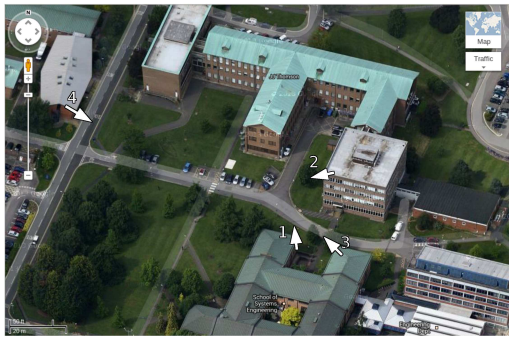

Figure 1. Recording site and environmental camera placement. There are four cameras available marked with a red arrow.

The remainder of the paper is organised as follows. Section 2 presents the PETS 2014 dataset in detail. The challenges addressed with this dataset are given in Section 3. For authors willing to investigate only abnormal event detection and behaviour understanding topics, we have tracking results made available for them. A brief explanation of such trajectories is given in Section 4. Some conclusions on the current dataset as well as some potential new challenges in forecoming PETS workshop are presented in Section 5.

2. Dataset

The datasets are multisensor sequences containing different activities around a parked vehicle in a parking lot.

In PETS2014 workshop, 22 acted scenarios are provided of abnormal behaviour around the parked vehicle. Although

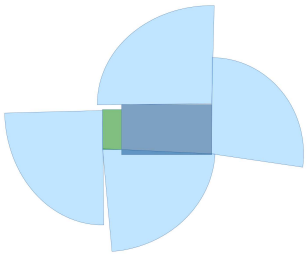

Figure 2. Truck on-board camera setup. Four non-overlapping visual cameras are mounted at each corner of the vehicle. They are represented with light blue colour.

CCTV cameras allow for the surveillance of the area around the vehicle, the main objective is to understand the different behaviours and detect potential threats from four visual (RGB) cameras mounted on the vehicle itself.

2.1. Camera setup and characteristics

The recordings were carried out at the University of Reading, more precisely at the crossing path and car park in front of the School of Systems Engineering.

Environmental cameras. These are installed at the locations shown in Figure 1 to cover an approximate area of 100m x 30m. These cameras were active during the scenario recordings. Note that these cameras are employed to allow authors to obtain a global view of the employed space

On-board cameras. The on-board camera configuration during the recordings is shown in Figure 2. Four visual cameras are employed. Their characteristics are as follows: Model: BIP2-1300c-dn (http://www.baslerweb.com/products/Fixed-Box.html?model=178); Resolution: 1280 x 960 pixels; frame rate: 30 fps.

2.2. Scenarios Recorded

Scenarios have been divided into three different threat-level categories:

- 'Something is wrong': Abnormal behaviour that however cannot be considered as a real threat. Recorded scenarios in this category include, for instance, people standing a long time by the truck while phoning, or security guards walking around the truck, activity which is however innocent because the driver, the vehicle or any contained asset are not in danger. Some other scenarios include the driver accidentally falling by themselves but standing up again later on.
- 'Potentially criminal': The security of driver, the vehicle or any contained asset is in danger. Scenarios included in this category show, for instance, a potential thief walking around the vehicle and clearly trying to open the truck. Other potential dangerous situations are the driver falling down but this time pushed by someone else (although accidentally).
- 'Criminal behaviour': The security/safety of driver, the vehicle or any contained asset has been breached. In this category the scenarios include, for instance, people succeeding to access the truck and steal an object from it. Other scenarios include an attack to to the driver (physical agression), in order to steal something from them.

Sequence	Description	Difficulty
03_06	1 person stands near truck	1
06_01	1 security guard inspecting; walking around truck	1
08_02	Driver falls; someone comes to help	2
03_05	1 person stands near truck, then someone comes to ask directions	2
08_03	Driver falls; two persons come to help	3
06_04	Two security guards walk around the truck	3

Table 1. 'Something is wrong' Scenarios

Sequence	Description	Difficulty
10_04	1 person walks around truck and Attempt to open truck;	1
10_05	1 person walks around truck and Attempt to open truck;	1
11_04	Driver falls pushed by someone	2
10_03	1 person walks around truck and Attempt to open truck;	2
11_05	Driver falls pushed by someone	3
11_03	Driver falls pushed by someone	3

Table 2. 'Potentially criminal' Scenarios

Sequence	Description	Difficulty
14_01	1 person walking around truck and steals something	1
14_07	1 Person stealing; loitering; walking around truck	1
14_03	1 persons steal something	2
14_05	1 person loitering; 1 person stealing	2
14_06	Group of 3 come to steal; two of them loitering	3
15_02	Driver hitten by two and brought to floor. Attackers take possesion of truck	3
15_05	Driver trapped by asking directions. Driver hitten by two and brought to floor. Attackers take possesion of truck	3

Table 3. 'Criminal' Scenarios

Sequence	Description	Difficulty
15_06	Driver involved in a fight with someone. Two more people come to hit him and brought to floor.	3*
22_01	Driver hit and stolen from someone from a car	3*
23_01	Driver attacked from someone from a car and brought to floor	3*

Table 4. 'Extra Criminal' Scenarios

The sequences composing specifically each category are shown in the next tables (Tables 1-4). In these tables, the marked difficulty involves the following criteria: The number of main actors involved in the scene, the general density of people in the scene, and if the scenario needs analysis from multiple cameras or if possibly one view is sufficient.

Figure 3. The truck driver is hit by three subjects that eventually run after the driver falls to the ground.

Figure 4. Someone opens the truck door while the driver is away an steals something.

Typical combinations for the different difficulty levels are as follows:

- Difficulty 1: one actor involved; scene quite empty; one camera view may suffice.
- Difficulty 2: two main actors; scene quite empty; one/two camera views may suffice.
- Difficulty 3: three or more actors are involved; the scene is more frequented.
- Difficulty 3*: multiple actors are involved and their interaction is intricate (involving a fight or an attack) and in some cases a moving vehicle is also employed.

2.3. Acted Behaviours

Each scenario contains acted behaviours, which may be taken as indications or pieces of information allowing inferring if the scenario activity corresponds to innocent abnormal activity ('something is wrong') or if the activity can be considered as 'criminal behaviour' or 'criminal behaviour'.

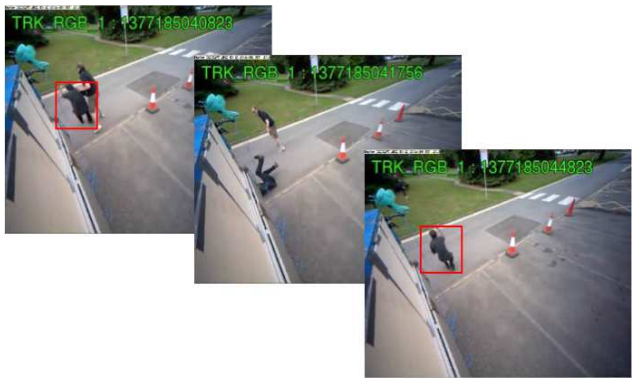

Figure 5. A person falls to ground after accidentally colliding with another person.

Figure 6. A person stays long in the vicinity of the vehicle and eventually walks all around the truck.

Acted behaviours are divided as 'normal', 'abnormal' or 'threats'. They are defined as follows:

Normal behaviours. This corresponds to people simply walking along the pathways around the parked truck (see Figure 7).

Abnormal behaviours. There are a large number of abnormalities recorded in the dataset. However, only some of them are annotated. Behaviours that can be clear indications of abnormalities are:

- Falling: Person losing balance and falling to ground. Can be caused by themselves or by a third person (see Figure 5).
- Staying long in the vicinity of the vehicle (Loitering): Person stands/moves slowly in the same area (see Figure 6).
- Walking around the vehicle: Walking at least two sides of the vehicle (see Figure 6)

Threats. Behaviours that are clearly criminal behaviour:

Figure 7. A group of people detected and tracked walking by the Truck.

- Attack to driver: Physical and intentional aggression to driver where they are hit or menaced with an arm, and possibly brought to the ground (see Figure 3).
- Stealing from vehicle: Someone penetrates the vehicle completely or partially and departs with an object removed from the vehicle (see Figure 4).

3. Challenges

PETS 2014 dataset is designed to address vision challenges corresponding to different steps in a video understanding system: Low-Level Video Analysis, Mid-Level Video Analysis and High-Level Video Analysis.

3.1. Low-Level Video Analysis

The task is to detect and track objects in all frames from video sequences and report detected/tracked object bounding boxes for each object at each frame.

Sequences that can be processed in this category are 01_02 and 22_02 (only cameras ENV_RGB_3 and TRK_RGB_2 are evaluated in this category). Sequence 01_02 has a lower degree of difficulty as the focus is mainly on tracking a group of three people walking along the truck. Sequence 22_02 is a very complex sequence containing a large number of actors (sometimes in groups) walking on the different paths around the truck. Typical problems of partial and total occlusion appear often in the sequence.

3.2. Mid-Level Video Analysis

The task is to detect any of the abnormal behaviours stated in Section 2.3 and report this as an event with the frame and bounding box of the involved mobile at the start of the event and the frame and bounding box of the involved mobile at the end of the event.

- 'simple' abnormal events considered are:
  - people falling.
  - staying long in the vicinity of the vehicle (Loitering).
  - walking around vehicle.
- The learnt/modelled events should comply with the definition of the behaviour given in Section 2.3.
- All dataset sequences can be processed in this category except those destinated for evaluation of object detection and/or tracking (01_02 and 22_02).

3.3. High-Level Video Analysis

At this level, two different challenges can be addressed. The first is to detect events of high complexity given the interaction between actors and/or with the vehicle itself, namely the behavioural events indicating a threat as defined in Section 2.3. The second challenge at this level is on scene classification.

- 'complex' Threat event detection
  - The task is to detect any of the threat events in this category and report the frame and bounding box of the individual under attack or stealing from vehicle, at the start of the event, and the frame and bounding box of the individual under attack or stealing from vehicle, at the end of the event.
  - Threat events considered are thus:
    - attack to person.
    - stealing from vehicle.
  - The learnt/modelled events should comply with the definition of the behaviour given in Section 2.3.
  - Sequences that can be processed in this category are those included in the 'Criminal' and Extra 'Criminal' scenarios tables (Table 3 and Table 4).
- Sequence classification

– The task is to analyse all sequences in this category and label them as ‘Something is wrong’, ‘Potentially criminal’, or ‘Criminal behaviour’.
– Categories:
  * Something is wrong: Abnormal behaviour that however cannot be considered as a real threat.
  * Potentially criminal: The security of driver, the vehicle or any contained asset is in danger.
  * Criminal behaviour: The security/safety of driver, the vehicle or any contained asset has been breached.
– The classifier should comply with the scene classification given above so that the sequences can be grouped as shown in the Tables 1-4 in the ‘Dataset’ description.
– All dataset sequences can be processed in this category except those destinated for evaluation of object detection and/or tracking (sequences 01_02 and 22_02).

4. Tracking made available

For authors addressing only the Mid-level or High-level video analysis challenge, the workshop is providing tracking results for the corresponding sequences. This section provides a brief description of the tracker employed.

The tracking system developed by the University of Reading is designed to run in real-time, or near real-time, at upwards of 5 frames per second. It uses input from a combination of change and motion detection, and performs reasoning regarding the current state of the scene, how existing tracking targets should be associated to current detections, and how to update the location of current tracking targets. In outline, the tracking system can be described using the diagram in Figure 8 below.

Images are fed into the motion/change detectors, and the result fused and optionally filtered. The resulting detections are directed to the tracker. Before making use of these detections, the tracker exploits information in the image, including motion information from the optical flow field computed by the motion detector, to track and optimise the location of known targets to suit the current image. The tracker now breaks down all of the detections and existing tracking objects into what are termed atomic regions or simply atoms. These atoms are then used to determine the association between detections and existing targets. Finally, new targets are created as appropriate and appearance models of existing targets updated ready for subsequent frames. Reasoning is also carried out to determine which targets are no longer being tracked, or no longer exist.

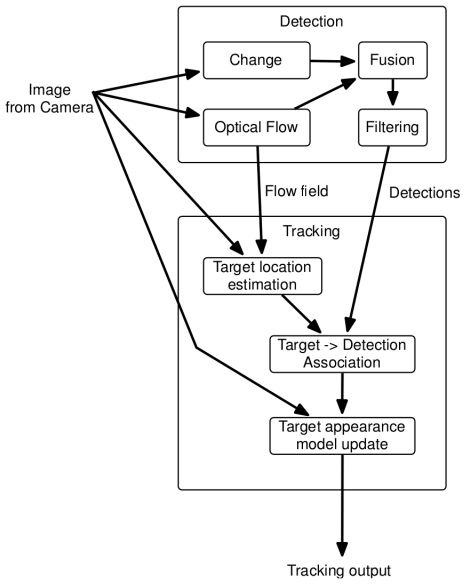

Figure 8. Processing chain employed on the tracking results made available.

4.1. Tracking Target Creation and Tracking

The detector provides a foreground mask indicating pixels of the image where objects are believed to be, as well as an optical flow field indicating object motion in the images. It reports a set of detections as bounding boxes interpreted from the foreground mask, along with an associated label map linking foreground pixels to specific detections. The specific algorithms chosen for this task are the Adaptive Gaussian Mixture model of Zivkovic2004 [2] and the optical flow estimation of Brox2004 [1]. Detections that are not associated to existing tracking targets are upgraded to new tracking targets. A tracking target consists of a bounding box (directly taken from the detection bounding box), a colour appearance model (an image the same size as the bounding box initialised to the pixels inside of the detection bounding box), and an extents mask (a greyscale image the same size as the bounding box that is initialised to white pixels for the foreground mask of the detection).

Once a tracking target exists, it must be tracked into a new frame, and for this the optical flow field calculated by the detector is used. Given the previous location of the tracking target, the pixels inside the targets bounding box can have their motions accumulated giving a good indication of how the target has moved between image frames. Once this initial motion estimate has been computed, the appearance model can be used to further optimise the location of the target in the image and verify its continued presence. This is achieved by computing the difference between the appearance model of the target and the current image for

a given location in the image, weighting the significance of the pixels using the extents mask of the target. Using this difference error, a search can be undertaken to determine the location in the image that minimises the difference between the template and the image.

4.2. Detection Association and the Atomisation Process

Once targets have been tracked to an estimated position in the current image, the new set of detections must be associated to the existing targets to determine if targets are still being detected, or if new objects of interest have entered the scene. Often the detector will produce some errors which the tracker will need to identify and compensate for. These include the merging of multiple objects into a single detection region, as well as the partial detection of objects, or fragmented detection of objects. To this end, a process of atomisation is undertaken. The atomisation process consists of breaking existing targets and foreground regions into segments, with existing targets claiming sections of foreground regions they overlap with. This results in a set of atomic regions that can be described as either an undetected claim (a target region that does not overlap with a foreground region), a detected claim (a target region that does overlap with a foreground region), or an unclaimed detection (a foreground region which is unclaimed by a target).

Should the detector produce a large foreground region corresponding to multiple objects, then existing tracking targets will be able to claim their portion of that foreground region, and recover detections of the individual objects. Similarly if a single object produces a fragmented detection, then the existing tracking target will be able to claim multiple detections, and merge those fragments to a unified object. The association process consists of building a table denoting the overlap between each atomic region and each tracking target. Atoms are now associated uniquely to a single tracking target, iterating through the table from most certain association to weakest. This results, potentially, in a many-to-one detection to target association result, and does not permit a one-to-many situation. In the event of a many-to-one association, consideration must be given to whether a tracking target actually consists of multiple objects producing those multiple detections.

4.3. Target Updating

Once detections have been associated to a tracking target, the final stage of tracking is to update the information of the tracking target, that is to say, update the appearance and extent models and the size of the bounding box. The bounding box is resized to ensure that all associated atoms fit inside. The appearance and extents models are then updated as a running average, updated using the values of the pixels in the current image beneath the updated location of the tracking target. The running average must be performed with some care to compensate for any resizing of the bounding box.

5. Conclusions

We have presented in this paper the dataset and vision challenges that form part of the PETS 2014 workshop. The recorded videos contained in the dataset, correspond to multisensor sequences containing different abnormal activities around a parked vehicle in a parking lot. Recorded scenarios have been divided into three different threat-level categories: 'something is wrong', 'potentially criminal behaviour' and 'criminal behaviour'. The vision challenge comes down to deciding if the detected activities in the video are part of innocent abnormalities or if they constitute a real threat to the vehicle, its driver or any asset contained inside. For this, behavioural cues and the temporal history of the scenario must be analysed. Behaviours of interest included in the dataset are: people loitering; walking around the parked vehicle; fall on floor; fight and stealing. The dataset gives the opportunity to evaluate different steps in a video understanding system: Low-Level Video Analysis (object detection and tracking), Mid-Level Video Analysis ('simple' event detection: the behaviour recognition of a single actor) and High-Level Video Analysis ('complex' event detection: the behaviour and interaction recognition of several actors). Future challenges could include tracking with overlapping cameras and/or driver face analysis. The current dataset is publically available for the purposes of the PETS workshops and academic and industrial research (see download instructions at www.pets2014.net/a.html). Where the data is disseminated (e.g. publications, presentations) this source should be acknowledged.

Acknowledgements

This work was supported by the EU ARENA project under grant agreement 261658. 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.

References

[1] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision (ECCV)*, volume 3024 of *Lecture Notes in Computer Science*, pages 25–36. Springer, May 2004.

[2] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2 - Volume 02*, ICPR '04, pages 28–31, Washington, DC, USA, 2004. IEEE Computer Society.