Signal, Image and Video Processing
DOI 10.1007/s11760-014-0672-1

Instantaneous threat detection based on a semantic representation of activities, zones and trajectories

G.J. Burghouts · K. Schutte · R.J-M. ten Hove · S.P. van den Broek · J. Baan · O. Rajadell · J.R. van Huis · J. van Rest · P. Hanckmann · H. Bouma · G. Sanroma · M. Evans · J. Ferryman

Received: 12 December 2013 / Revised: 27 June 2014 / Published online: 17 August 2014

Abstract Threat detection is a challenging problem, because threats appear in many variations and differences to normal behaviour can be very subtle. In this paper, we consider threats on a parking lot, where theft of a truck's cargo occurs. The theft takes place in very different forms, in the midst of many people who pose no threat. The threats range from explicit, e.g., a person attacking the truck driver, to implicit, e.g., somebody loitering and then fiddling with the exterior of the truck in order to open it. Our goal is a system that is able to recognize a threat instantaneously as they develop. Typical observables of the threats are a person's activity, presence in a particular zone, and the trajectory. The novelty of this paper is an encoding of these threat observables in a semantic, intermediate-level representation, based on low-level visual features that have no intrinsic semantic meaning themselves. The semantic representation encodes the notions of trajectories, zones and activities. The aim of this representation is to bridge the semantic gap between the low-level tracks and motion and the higher-level notion of threats. In our experiments, we demonstrate that our semantic representation is more descriptive for threat detection than directly using low-level features. We find that a person's activities are the most important elements of this semantic representation, followed by the person's trajectory. The proposed threat detection system is very accurate: 96.6% of the tracks are correctly interpreted, when considering the temporal context.

Keywords: Threat detection · Human action recognition · Spatiotemporal features · Tracking of humans · Trajectories · Zones.

1. Introduction

In this paper, we consider the detection of threats. Threats may occur in many areas and applications [21,24-26], among others, security (e.g., stealing), safety (e.g., crowded area) and military (e.g., trespassing). Detecting threats is relevant, because it enables professionals to mitigate an unwanted situation at an early stage. It is an interesting research area, because threat detection is a challenging problem, for several reasons which are discussed below.

The first challenge is that threats appear in many variations. For some threats, the key characteristic is the walking pattern, e.g., loitering. For other threats, the cue is the presence in a particular zone, which is considered suspicious, e.g., being present in a place where other people usually do not be present. Another category of threats is characterized by the current activity by the person posing the threat, e.g., trying to open a door. In this paper, the objective is to recognize a wide range of threats, by representing a variety of aspects of human behaviour. The novelty is our intermediate-level representation including a person's trajectory, presence in particular zones, and activities.

The second challenge is that complex threats are a high-level semantic concept. A threat is an interaction between on the one hand the person or group of persons posing the threat, and the threatened person(s) or object on the other hand. The person posing the threat will try to limit exposure to a minimum. This leads to a complex interaction and the differences of the behaviour compared to other people, who pose no threat, may be very subtle. Together with the variations in which threats may occur, a thorough interpretation of the observed cues is required, beyond simple rules on simple cues. Yet, the popular

G. J. Burghouts · K. Schutte · R. J.-M. ten Hove · S. P. van den Broek · J. Baan · O. Rajadell · J. R. van Huis · J. van Rest · P. Hanckmann · H. Bouma: TNO, Intelligent Imaging, The Hague, The Netherlands. E-mail: gertjan.burghouts@tno.nl

G. Sanroma: University of North Carolina, Chapel Hill, NC, USA.

M. Evans · J. Ferryman: University of Reading, Reading, UK.

approach in computer vision for recognizing human behaviour is to start with low-level entities, the most common ones are trajectories resulting from tracking, e.g., [1], and hand-crafted features, e.g., STIP [2]. Such low-level features are very useful, because they capture essential details about trajectories, local shape, motion, and they are localized in space and/or time. However, they are not directly associated with persons, zones in the scene, a person's activities, and what happens during a person's trajectory. Such associations are not trivial: many of the well-performing methods consider the low-level features in the whole video [3], the whole scene [4], or in sub-volumes without making explicit associations [5]. Recent attempts for complex behaviours in complex scenes have not been successful yet [6], although reasonable performance have been reported for simple activities [7]. For threat detection, this is not sufficient: our aim is to identify who is posing the threat and when that happens. Clearly, there is a huge semantic gap between threat detection and low-level features. Our contribution is that we exploit the advantages of low-level features and bridge the semantic gap to threat detection by an intermediate-level representation of the person's trajectory and activities.

The third challenge is to recognize the threat as soon as possible, while in the midst of many other people who pose no threat. The cue for the threat will be more explicit and distinctive at a later stage, while the early cues may be less distinctive. At the beginning of a threat, the behaviour may look very similar to the behaviour of other people, e.g., just loitering is not really suspicious. Our objective is to distinguish between threats and normal behaviour, as soon as possible while the threat is building up, ideally from the moment that the person who is posing the threat starts to show the first cues with acceptable false alarm rate.

In this paper, we consider the theft of cargo from a truck, when the truck is parked. This is an interesting case, because there is a wide variety of threats, and there are many other people present. The threats range from explicit, e.g., a person attacking the truck driver, to implicit, e.g., somebody loitering and then fiddling with the exterior of the truck in order to open it (see Fig. 1 for two illustrations).

A graphical outline of the proposed system is displayed in Fig. 2. In our experiments, we demonstrate that our intermediate representation (i.e., trajectories, zones, activities) is more distinctive for threat detection than directly using low-level features (i.e., tracks and STIP). We will show that each element in our representation contributes to the overall discriminative power. The proposed threat detection system is very accurate: 96.6% of the tracks are correctly interpreted, when temporal context is considered.

Fig. 1. Two examples of threats to a truck: somebody fiddles with the truck (*top*) and the truck driver is attacked (*bottom*). The goal of this paper is to detect such threats, in the midst of other people who pose no threat, as soon as possible. The novelty is that we detect a wide variety of threats based on a semantic, intermediate level representation that describes the state of a person (see *bounding boxes*) by the trajectory, presence in zones, and activities (see *text boxes*).

The paper is organized as follows. Section 2 discusses other research on threat detection. In Sect. 3, we introduce the low-level features. In Sect. 4, we propose the intermediate-level representation. Section 5 defines the experimental setup, followed by the threat detection results in Sect. 6. Section 7 concludes the paper with our findings.

2. Related work

Recently, researchers started targeting a wider variety of threats, such as unwanted behaviours inside a train [8]. Detection of multiple threat models is the focus of our paper, where threats range from explicit, e.g., a person attacking the truck driver, to implicit, e.g., somebody loitering and then fiddling with the exterior of the truck in order to open it.

Threats may have a short duration (e.g., an instant attack) or a long duration (e.g., loitering and fiddling with the exterior of the truck). Many approaches have investigated longer-term behaviours [9]. When longer-term behaviours are composed of several short-term actions in a particular sequence, the temporal structure can be exploited in sequential models such as hand-crafted grammars [10], or statistical, graphical models [11].

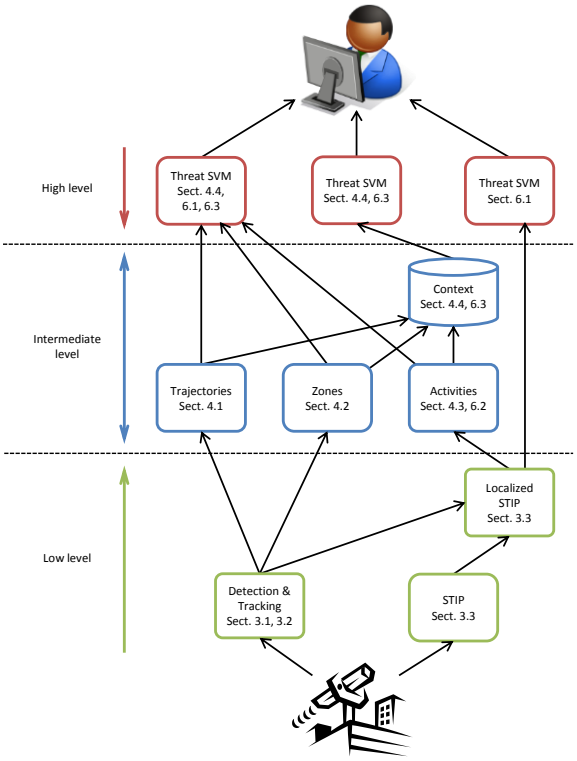


Fig. 2. Graphical outline of the proposed system.

The merit of temporal structure has been demonstrated for behaviours such as 'two people meet, then depart' [10] and 'having a snack' inside a small living room [11]. These behaviours have a small spatial extent. In more complex scenes, where behaviours spread to more spatial extent (e.g., loitering and then attacking), track breaks are common (e.g., obstructions cause non-visible areas resulting in temporary occlusions). There is yet no widely accepted methodology to recognize longer-term behaviours in complex scenes where obstructions and track breaks are common. The temporal structure can also be exploited by simpler approaches, where recent history is encoded in the representation itself [8,12]. We include this notion that recent history is an important queue as described in Section 4.4.

A threat may appear by a multitude of observable properties, ranging from a trajectory (e.g., a specific interest to the truck driver) to a particular activity (e.g., trying to open the truck). Many previous approaches have considered one particular observable property, such as trajectories [13], motion features [4], or group interactions [7]. Multimodal fusion has been an active research topic, in particular to detect specific video concepts [14]. Our approach will also comprise various observables, which can be derived from the visual source of a camera feed as this is the most commonly available sensor in surveillance. Our focus is on those observables that are indicative of threats: trajectories, activities and presence in particular zones in the scene.

Hierarchical models [11,13] and models with an intermediate level [8,15] have gained interest for automated recognition of complex behaviours. These models facilitate machine learning, by reusing the low-level actions [11]. Further, multi-level approaches decrease the semantic gap between low-level features and complex behaviours [8]. An intermediate-level representation with dedicated components for the complex behaviours of interest has been successful for highly semantic phenomena such as the TRECVID MED competition [14,16]. We follow this approach for threat detection and design a dedicated intermediate-level representation that captures semantic observables related to threats.

An important design choice is whether the system is constructed by manual design [9,10,13] or trained [8,12,16]. The advantage of hand-crafted models is that expert knowledge can be included and that limited or no training data is required [9]. This has proven to be very effective for modelling trajectories through a scene [13]. Our representation constitutes more than trajectories only: we also represent activities and presence in particular zones. Our representation has a much higher dimensionality, and to the best of our knowledge, hand-crafted models have not been applied successfully to such representations. Therefore, we adopt the two-stage methodology by [8] which decomposed the learning problem from low-level features to high-level concepts into two steps: from low-level features to the intermediate level representation, and from this representation to threat detection.

In many approaches the scene is interpreted as a whole, for instance, for detection of actions [3] and video concepts [14]. A popular approach to have some spatial localization capacity, is to include a fixed segmentation of the scene, such as tiling by a so-called pyramid [17], or a weak segmentation incorporating many candidates [18]. For weak spatial and temporal localization, a spatio-temporal layout model was proposed in [4]. For threat detection, we will exploit the fact that we know the object of interest: a person and its track through the scene. Ultimately, we aim to identify who is posing the threat and where that person is in the scene and at which time.

The contribution of this paper is that we propose a method that is able to detect threats in early stages, in a complex scene with obstructions and many other people who pose no threat. The novelty is that we detect and localize a wide variety of threats based on a semantic, intermediate level representation that (a) describes the state of a person by the trajectory, presence in zones, and activities, and (b) can be constructed from realistic, imperfect tracks.

3. Low-level features

This section discusses the low-level features that are used in the proposed system, i.e., tracks and motion (STIP) features, see the bottom part of Fig. 2.

3.1 Object detection

Object detection is performed using a combination of motion and change detection. In a typical surveillance scene the objects of interest will generally be observed to move through the scene, and this motion allows objects to be detected either by explicitly detecting the motion using optical flow, or by learning the appearance of the static background and identifying when pixels change. Change detection often produces foreground masks that are highly accurate indications of the location of moving objects of interest, however the approach struggles to split foreground blobs for objects whose paths cross and will leave ghosts where objects linger long enough to become part of the background model. Optical flow determines the motion of pixels in the image in terms of both speed and direction, and foreground masks can be produced by thresholding based on the speed while ensuring the resulting detection regions have homogeneous direction. As such, optical flow based motion detection, can overcome the merging of objects that are travelling in different directions or at significantly different speeds and will not produce ghost regions as there is no background model. However, the resulting foreground mask will not generally provide accurate silhouettes of the objects. This makes the two approaches highly complementary and as such, in this work, change and motion detection are combined together. The fusion process consists of a logical OR of the two approaches' foreground masks and then the use of heuristic reasoning to permit the motion detector's regions to split change regions that have inhomogeneous motion, while the regions of the change detector are permitted to merge motion regions that show homogeneous motion. For this work, change detection is performed using the Adaptive Gaussian Mixture Model of [1], while real-time optical flow is provided by OpenCV's GPU implementation of [19].

3.2 Object tracking

Targets that are tracked are initialised from object detections. Each target maintains a model describing its RGB appearance and its spatial extents produced from a running average of the image and foreground mask pixels inside the target's bounding box. Each frame, the optical flow determined during detection is used to predict target motion between frames, and this motion used to initialise a search for the target's location. The search consists of minimising the difference between the target's RGB appearance and the image for a given bounding box location. The difference is computed as the sum of RGB pixel differences, weighted by the magnitude of the pixels in the extents model to minimise the impact of background pixels. Tracks are associated to the detections in the current frame using heuristics to handle common detection issues such as fragmented detections and merged detections. Each tracking target "claims" a portion of the foreground image based on its extent mask, resulting in "atomic" regions – unclaimed detections, claimed detections, undetected claims. These atomic regions are associated to existing tracking targets based on region overlap. Association can be many atoms to one target, allowing for partial or fragmented detections to be handled. If multiple atoms associate to one target, the target tracks each of the atoms independently and as a whole – if they continue to have motion consistent with the whole, they are merged back into the target (basically, ignored), however, if they move away from the target, a splitting event occurs and two new targets are created from the original. Unclaimed detections become new targets.

3.3 Localized motion features

To capture the motion patterns of human actions, STIP features [2] proved to be very effective. They were found to be superior to track- and object-based features [15]. For each track and each bounding box, we associate the STIP features that are within the box, by comparison of (x,y) location and frame number and checking whether the location is inside the box at that frame [7]. An illustration is shown in Figure 3. The tracks and their STIP features are our starting point to represent human activities, trajectories and presences in zones (see Sect. 4).

4. Threat detection based on a semantic representation

In this section we propose the intermediate-level representation and high-level threat detection, see the middle and top parts of Fig. 2.

4.1 Trajectory

The intermediate-level representation includes a person's trajectory, comprising the positions, kinematics and travelled distance. The kinematics are described by speed, orientation and travelled distance. The positions, travelled distance and kinematics are encoded in the feature vector, in image as well as world coordinates. Both have their advantages and disadvantages. Image coordinates are more robust but kinematics depend on the projection on the image plane, which has the disadvantage that it varies with distance. World coordinates are more or less invariant with respect to the projection, but are less

reliable. We represent these properties in both coordinate systems to be able to exploit the respective advantages.

4.2 Meaningful zones in the scene

The scene that will be considered in the experiments (Sect. 5) is shown in Fig. 3. The four zones are identified manually. The truck park area is the area of interest (middle right): threats involve the truck, its contents, or its driver. Typically, the person posing the threat arrives by car (middle left) or walks into the scene through the bus stop area (lower right). The driver often goes to the cafeteria (lower left) or stays close to the truck (e.g., to make a phone call or smoke a cigarette). These zones are important for an interpretation of the observed behaviour. In the representation, we store for each second a boolean indicating a person's presence for each of the four zones. Presence in a zone is defined as at least one pixel overlap between a person's bounding box and the zone area.

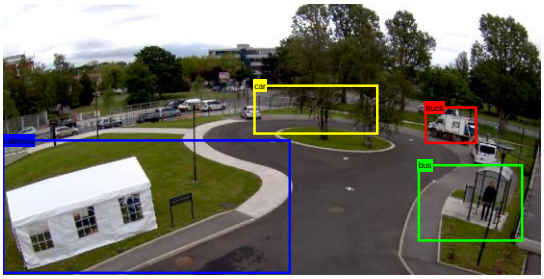


Fig. 3. The four relevant zones in the scene, i.e., car park area (see *yellow box*), truck park area (*red*), bus stop area (*green*) and cafeteria (*blue*).

4.3 Set of human activities

The threats considered in the experiments (Sect. 5) range from explicit, e.g., a person attacking the truck driver, to implicit, e.g., somebody loitering and then fiddling with the exterior of the truck in order to open it. The threats generally involve particular activities. Clearly, the representation of human activities is valuable to interpret complex behaviours and to detect threats. We define a set of human activities that span both threat scenarios as well as normal behaviours that pose no threat, in order to disambiguate the two situations. The set of activities used is: {walk, run, loiter, turn, enter/exit vehicle, fiddle/check the vehicle, fight}.

The model for each activity is obtained by a bag-of-features approach, which we describe in detail in our recent work [15]: For each one-second interval of the track, it transforms the STIP features into visual words and represents the track fragment as a frequency count of the words. The quantizer of choice is a random-forest [20]. It has a high discriminative power because it exploits the labels for each activity (further described in Section 5.2) during the training phase. This way of quantization led to good performance in our recent experiments [4]. The final step is the SVM classifier with a χ^2 kernel. The classifier serves as the detector for each activity. For each one-second fragment of the track, we obtain a posterior probability for each of the activities.

4.4 High-level classification: threat detection

In Sections 4.1, 4.2 and 4.3 we have described the elements of our intermediate-level representation. For each one-second fragment of each track, we obtain features capturing trajectory, kinematics, presence in zones, and a set of activities.

- The trajectory level provides information about the kinematics of the object, including the direction of movement, which is important for detecting threatening situations. For example, the action consisting of a person walking towards the truck could be related to a threatening situation depending on the direction of movement. That is, if the person is coming from the service area, it could well be the truck driver returning from having a meal (i.e., normal behavior), whereas if the person is coming from the car parking area it could well be a potential thief aimed at checking the truck (i.e., suspicious behavior).
- The zone information helps determine where the action is taking place. This provides contextual information essential to determine threatening level in some situations. For example, loitering in the bus stop area can be considered normal behavior, whereas loitering in the truck parking area could be indicative of a threat.
- Human activities provide the semantics about the instantaneous actions that are taking place, which complements both zone and trajectory information in determining the threats, as already discussed.

Together the trajectory, kinematics, presence in zones and activities yield a feature vector, which together with the threat labels (further described in Sect. 5.2), are used to train another SVM (also with the χ^2 kernel which showed best performance compared to radial basis function). The result of this classification is for each track a per-second assessment of the threat level.

We present a causal system that detects threats instantaneously based on one-second segments. This is a causal system, as no information about the future is used. As an alternative, we also present a variation of the method that assesses the complete track. This is a method that has a slightly delayed response, as it requires the track to be finished before the threat assessment can be done. We call this variation 'fragments with context', where the term context refers to temporal context. Contextual information is represented by the maximum and mean values of the included elements during the complete track. Both implementations will be compared in the experiments.

5. Dataset and set-up

5.1 Videos of threat and normal scenarios

We perform threat recognition experiments on a dataset recorded for EU project ARENA[†]. The ARENA project aims to detect threats to mobile assets from multiple affordable sensors. This dataset contains 23 videos that include many tracks that pose a wide range of threats, and almost thousand other tracks of normal behaviour in a parking lot. The total duration is 77.5 minutes. The average video has a duration of 3.4 ± 1.2 minutes.

To the best of our knowledge, our dataset is the first one about detection of complex threats in video that involve (combinations of) various human behaviours, and which is publically available to the research community. To enable other researchers to compare their methods to our method, we make our dataset available in the international renowned PETS2014 benchmark as the "ARENA dataset" to enable direct comparisons. We refer to the PETS2014 website: www.pets2014.net.

The threat scenarios are 'aggression towards the truck driver' (3 videos), 'hostile take-over of the truck' (3 videos), 'stealing from the truck cabin' (1 video), 'inspecting the truck exterior' (3 videos), 'touching the truck' (1 video), 'trying to open the truck' (1 video), 'normal behaviours' (11 videos). The main activities that may be cues for a threat are 'attack person', 'follow person', 'try to open truck', 'enter/exit truck', 'take over truck'. The main cues that can be derived from trajectories are 'loiter', 'approach truck or person', 'stay in car park area' (from which some threats originate), while going to the cafeteria or bus stop may be cues for normal behaviour.

In total, 11 volunteers were involved in the experiments. They entered the scene multiple times. The maximum number of people visible in the video was 7. For number of persons during threats ranged from one (one person posing a threat to the truck) to five (one person attacking the driver and the other person trying to enter the truck on one side of the scene and three neutral persons passing by on the other side of the scene).

5.2 Annotations

Threats. For each video, for each track, an annotation threat vs. normal is obtained. There are in total 998 tracks found by processing the videos, with a total time span of 6,842 s (on average 6.9 s per track), of which 86 tracks are annotated as threats, with a total time span of 1,170 s (on average 13.6 s per track). Tracks that are associated with threats have a longer duration, because typically such persons wait until the right moment to approach, attack, break in, etc. In Fig. 1, two example threats are shown.

[†] www.arena-fp7.org

Human activities. For each video, for each track, per one-second interval, a human activity label is annotated: walk (478 annotations), run (14), loiter (100), turn (133), enter/exit vehicle (123), fiddle/check the vehicle (26), fight (31). The one-second track intervals that span less than 10 frames, that have less than five bounding boxes, or that have less than seven STIP features, are discarded for training. Results of the annotations are illustrated in Fig. 4. It shows the difficulty of recognizing human activities: low contrast (people do not wear colourful clothes), large viewing angle variations (caused by the large field of view), and partial tracks (due to occlusions and track breaks). Other issues complicating automated analysis are as follows: due to low resolution (when the person is far away from the camera) and large variations of scale (persons can be close to or far from the camera). The most frequent activities are walking, loitering, and standing somewhere and turning around. The interesting activities occur only several times, which makes this dataset very challenging and relevant for surveillance.

Fig. 4. Examples of annotations of human activities, in one second windows of the tracks on people (shown by the *masks*).

5.3 Performance measure and cross-validation

As a test framework, we consider a leave-one-video-out setup for cross-validation. We exclude 1 video for testing, and learn the activity models and threat model on the remaining 22 videos. This is repeated for all 23 videos. Leave-one-out cross-validation is commonly accepted in action recognition from video, see e.g. [22] and the ICPR'10 action detection benchmark [23].

Since we aim at instantaneous, early threat detection, we evaluate the performance per one-second fragment of all tracks in all videos. Note that this does include assessment of track fragments where up to its time no observable threat evidence has been seen, i.e. this performance measure will for causal systems never reach a perfect score. The performance measure is the average

classification accuracy, i.e., the average of the diagonal of the confusion matrix. Each video is a new recording with a different threat or normal behaviour, and also the people in the scene are varied. Therefore we believe that the leave-one-video-out is a sound evaluation of discriminative power and generalization capabilities or the intermediate-level representation and threat detection method.

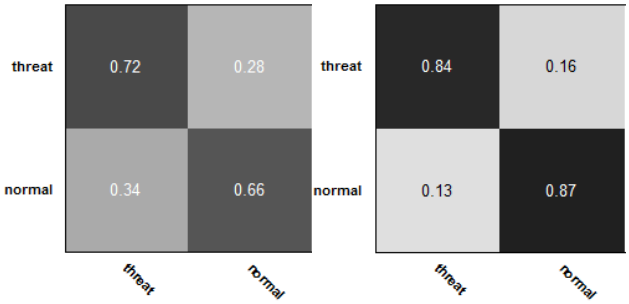
6. Threat detection results

The experiments include three assessments: a comparison between the semantic intermediate-level representation and low-level features (Sect. 6.1); the quality of the representation of human activities (Sect. 6.2); the merit of each element of the representation, i.e., trajectories, zones and activities (Sect. 6.3).

6.1 Semantic representation vs. low-level features

To evaluate the merit of our semantic, intermediate-level representation, we compare against an approach based only on the low-level features (i.e., the STIP features in the one-second fragments of tracks). In both cases, we consider the high-level classification, i.e., threat vs. normal, but with different input features. For intermediate-level features, we perform this classification using the features that encode trajectories, zones and activities, using the classification as described in Sect. 4.4. For low-level features, we perform this classification by using the track-localized STIP features in the same bag-of-features approach as described in Sect. 4.3. As this approach is a common practice in action recognition, e.g., [3,4,7], this is our baseline.

The results are shown in Fig. 5. The intermediate-level representation improves the threat detection significantly. With low-level features only, an accuracy of 69.0% is achieved. The accuracy is better when the intermediate-level representation is used: 85.1%. With our representation, the threat detection has less false negatives (16% vs. 28%) and less false positives (13% vs. 34%).



Low-level features (69.0%) Semantic representation (85.1%)

Fig. 5. Confusion matrices of the threat detection accuracy for the low-level features (*left*) vs. semantic representation (*right*).

6.2 Quality of human activities recognition

The proposed representation includes human activities as one of the most important properties of a threat. Here we evaluate the classification of the seven activities from Sect. 4.3. The classification of these activities are described in more detail in recent work [7]; here we summarize these results. Fig. 6 shows the confusion between the activities. The average classification accuracy is 70.8%. For this seven-class problem the average performance by chance is 14%, so our performance is reasonable, especially since our setup for early threat detection is to estimate the activities from very short track fragments of one second only. The good performance of 'walk', 'loiter' and 'enter/exit vehicle' (around 80%) can be explained from their prominent spatiotemporal appearance, whereas other activities are more subtle, and their duration is shorter. Hence it is harder to distinguish between 'turn', 'run', 'check vehicle', and 'fight' (around 65%). The challenges are: tracks compromised due to the speed ('run'), interaction with the vehicle ('check vehicle'), or with another person ('fight'). These results in misaligned bounding boxes which has a negative impact on the activity classification.

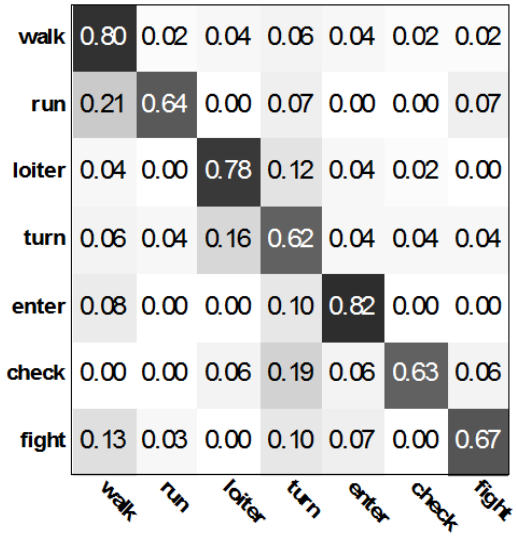


Fig. 6. Confusion matrix for the human activities that are part of the semantic representation (average classification accuracy: 70.8%).

6.3 Merit of each element of the semantic representation

The final experiment is to assess the merit of representing trajectories, presence in zones, and activities. We evaluate the threat detection accuracy for each element, and also when all elements are combined into our representation. Table 1 shows the results, where the columns indicate the elements included in the representation (left), the obtained accuracies when using only the track fragment itself

(middle), and when adding contextual information from the complete track (right). Context information is represented by the max and mean values of the included elements during the complete track. This gives an indication how well the method performs if we would allow the whole track to be analysed. In actual applications on streaming video, the system needs to be causal, and in case of temporal context the system would need to wait with the potential alert until the track is finished.

The findings are as follows. The full representation, on the one-second track fragments results in a 85.1% accuracy (this is the accuracy from the confusion matrix from Sect. 6.1 and Fig. 5). The activities are the most important element in the representation, when considering the fragments (78.2%). The zones and trajectories are less important (they perform approximately 15% less). Interestingly, the accuracy of threat detection based on human activities is better (78.2%) than the actual recognition of human activities themselves (70.8%). This implies that for threat detection, the selected activities are probably redundant. From this result we learn that the chosen set of activities may be too fine-grained for the purpose of threat detection.

The accuracy can be increased by considering the context of the complete track. Interestingly, for fragments combined with context, the trajectories are the most important element of the representation. This implies that when observing a complete track, its trajectory and kinematics are a very rich description already. Adding activities and zones to the trajectory and kinematics makes the representation much more powerful: the threat detection accuracy increases to 96.6%. We expect that a large portion of this increase is due to track fragments at the beginning of the video with up to that moment no observable threat evidence being correctly classified as threat when adding complete track context, i.e. this change is mainly due to the available granularity of the treat annotations, at the whole track only.

We have also evaluated the pairs of representations, see the middle part of Table 1. Interestingly, for the track fragments of one second, the pairs perform less than the best of the constituting representations. When including temporal context, the results for the combination trajectories and activities and the combination zones and activities are very good. These combinations perform almost as good as the combination of all three representations. The combination trajectories and zones does not perform well; the performance is less than either trajectories or zones. The least valuable representation, in case of including context, are the zones. There are not much zone transitions, so in a larger temporal context, this representation by itself is not very valuable. Yet, together with the complementary representation of activities, it has a strong additional value. Zones and activities are the best pair of representations and perform just 0.4% less than the combination of all three representations.

Table 1. The threat detection accuracy (%) for each element of the semantic representation, for detection based on a track fragment of one second, and when adding temporal context.

Representation	Track fragment	Track fragment + Context
Trajectories	62.8	91.6
Zones	64.7	65.1
Activities	78.2	85.7
Trajectories + Zones	67.8	66.9
Trajectories + Activities	75.3	95.8
Zones + Activities	76.7	96.6
All	85.1	96.6

6.4 Generalization to a new environment and other situations

In order to demonstrate that our system generalizes well to unseen data, recently, our system was used to perform a live demonstration using completely unseen data from another parking lot in Paris. This is a different environment with different cameras, different viewpoints, different actors and different threats. The zones were re-defined for this new environment; the other parameters of the method were not changed (no re-learning). Due to privacy reasons, it is not allowed to make the new demonstration data publicly available. Below we show the quantitative as well as qualitative results.

Fifteen scenarios were recorded, respectively: five normal, five posing a threat to the truck, five posing a threat to the driver. The maximum five threat confidences for each of the fifteen scenarios are shown in the chart in Fig. 7. It shows the scenario names horizontally, where prefix 01 refers to normal scenarios and prefixes 02 and 03 refer to the threatening scenarios. Vertically the chart shows the threat confidences, where for each scenario the maximum 5 are shown by the blocks of coloured bars. Clearly the threatening scenarios have in general larger threat confidences than the normal scenarios. With a threshold of 0.15, no threats are missed, at the cost of one false positive (scenario 01_04).

To provide some insights in the video data and the threat detections, we visualize several outcomes in Fig. 8, at the top a normal scenario, and in the middle and at the bottom, a threat to the truck and a threat to the driver. In our visualization, the tracks are summarized in one image by means of insets of the tracks, where each track is shown by five boxes including the image content inside that particular box and frame. The normal scenarios include friendly interaction with the driver. The system was able to identify that friendly interactions such as walking towards, or walking next to the driver, or talking to him, were not a threat (Fig. 8, top).

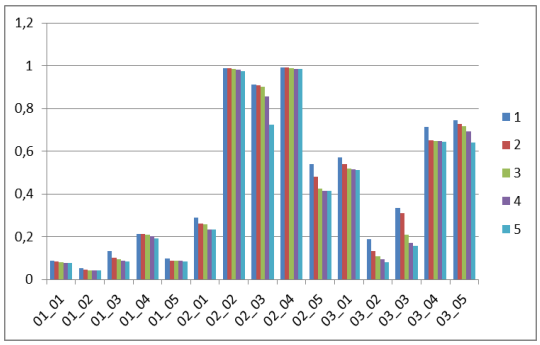


Fig. 7. The five maximum threat confidences for fifteen scenarios, of which the first five scenarios are normal and the last ten scenarios are threatening. For threats, i.e., scenarios with prefixes 02 and 03, the confidences are in general much higher.

The reason that such a scenario is well interpreted by our method, is that there are many friendly interactions between people (although not with the driver) in the learning set. The person with the pink shirt walks up to the driver (with yellow jacket) and a third person joins the conversation (with white shirt). Figure 8, in the middle and at the bottom, shows two detections of threats posed against the truck and the driver.

7. Conclusions

In this paper, we have proposed a threat detection system based on an intermediate-level representation that captures semantic descriptions including a person's trajectory, activities and presence in particular zones. We have described how this representation can be constructed from simple, low-level features, such as automated, imperfect tracks and common, localized motion features. We have shown how the intermediate-level representation outperforms the low-level features for threat detection. The activities and trajectories are very important elements of this representation. We envision that the proposed intermediate-level representation is also beneficial for explaining why the system has detected a threat, which is an important component of a surveillance system [21]. The representation consists of descriptions that have a meaning, like 'a person was present in the car park area' and then 'loiters around the truck area'. Such descriptions may help the operator to assess the threat. Further, it helps to get insight into system's decisions and errors. The latter is important for fine-tuning the system for optimal performance. The proposed system reliably detects threats. In 23 challenging videos, the average accuracy is 85.1% with just one-second track intervals. This is a very reasonable performance, given the small temporal extent of the analysis. When the total track is available for analysis, the performance is very good: the threat detection accuracy increases to 96.6%.

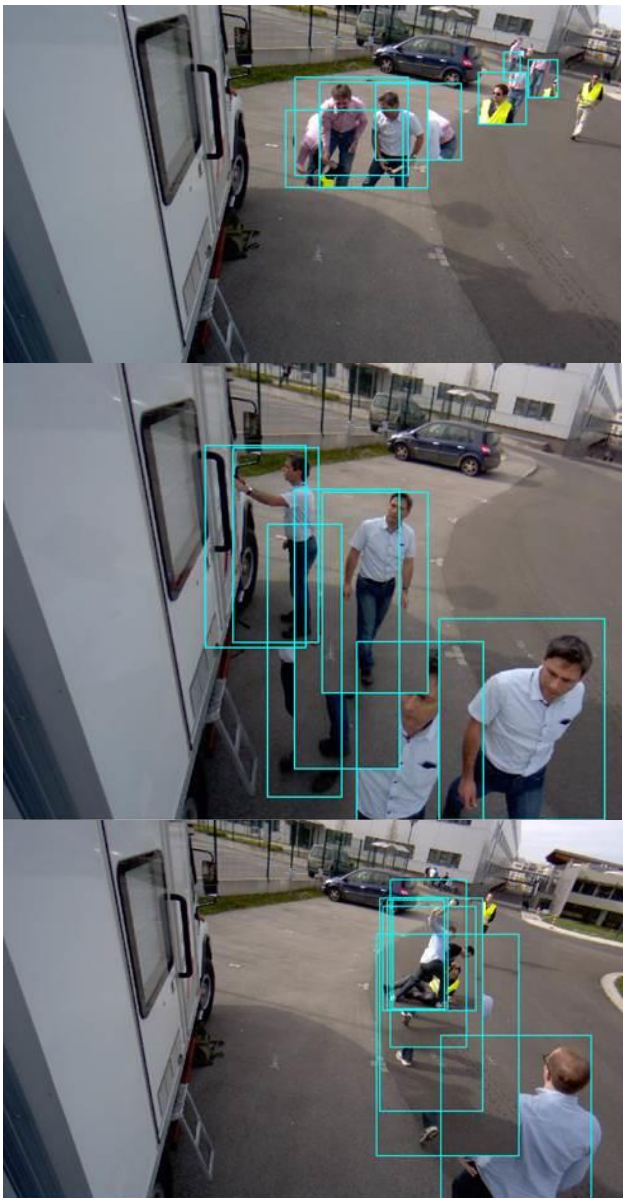


Fig. 8. *(Top)* Friendly interaction between a person and the driver, as part of an independent test set. This scenario was not in the training dataset. It resulted in a correct detection 'normal'. *(Middle and Bottom)*These threat scenarios were similar to samples in the training dataset, but with different cameras, viewpoint and actors. These two and the remaining threats all resulted in a correct detection 'threat'.

Acknowledgements The present work has been carried out in the framework of the EU FP7 project ARENA (grant ref. 261658). The authors want to thank the project partners for their contributions. 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] Zivkovic, Z: Improved adaptive Gaussian mixture model for background subtraction. ICPR **2**, 28-31 (2004).

[2] Laptev, I.: On Space-Time Interest Points. IJCV **64**(2-3), 107-123 (2005).

[3] Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". CVPR, 1996-2003 (2009).

[4] Burghouts, G., Schutte, K.: Spatio-temporal layout of human actions for improved bag-of-words action detection. Pattern Recognition Letters **34**(15), 1861-1869 (2013).

[5] Sadanand, S., Corso, J.: Action bank: a high-level representation of activity in video. CVPR, 1234-1241 (2012).

[6] Bouma, H., Burghouts, G.J., de Penning, L., et al.: Recognition and localization of relevant human behavior in videos. Proc. SPIE **8711**, (2013).

[7] Andersson, M., Patino, L., Burghouts, G.J., et al.: Activity recognition and localization on a truck parking lot. IEEE AVSS, 263-269 (2013).

[8] Lefter, I., Rothkrantz, L., Burghouts, G.J.: A comparative study on automatic audio-visual fusion for aggression detection using meta-information. Pattern Recognition Letters **34**, 1953-1963 (2013).

[9] Aggarwal, J., Ryoo, M.: Human activity analysis: a review. ACM Comput. Surv. **43**(3), 1-43 (2011).

[10] Zhang, Z., Tan, T., Huang, K., An extended grammar system for learning and recognizing complex visual events. IEEE TPAMI, **33**(2), 240–255 (2011).

[11] Nguyen, N., Phung, D., Venkatesh, S., Bui, H.: Learning and detecting activities from movement trajectories using the hierarchical hidden Markov models. CVPR **2**, 955–960 (2005).

[12] Hamid, R., Maddi, S., Johnson, A., et al. : A novel sequence representation for unsupervised analysis of human activities. Artif. Intell. **173**(14), 1221–1244 (2009).

[13] Sanroma, G., Burghouts, G.J., Schutte, K., Recognition of long-term behaviors by parsing sequences of short-term actions with a stochastic regular grammar. Struct. Syntactic Pattern Recognition, 225-233 (2012).

[14] Myers, G., Nallapati, R., van Hout, J., et al.: Evaluating multimedia features and fusion for example-based event detection. Mach. Vision Appl. **25**, 17-32 (2013).

[15] Burghouts, G.J., Schutte, K., Bouma, H., Hollander, R. den: Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos. Mach. Vision Appl. **25**, 85-98 (2013).

[16] Bouma, H., Azzopardi, G., Spitters, M., et al.: TNO at TRECVID 2013: Multimedia event detection and instance search, Proc. TRECVID, (2013).

[17] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, IEEE CVPR **2**, 2169-2178 (2006).

[18] Uijlings, J., Sande, K. van de, Gevers, T., Smeulders, A.: Selective search for object recognition. Int. J. Computer Vision **104**(2), 154-171 (2013).

[19] Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. ECCV **3024**, 25-36 (2004).

[20] Breiman, L.: Random forests. Mach. Learn. **45**(1), 5-32 (2001).

[21] Rest, J., van, Grootjen, F., et al.: Requirements for multimedia metadata schemes in surveillance applications for security. Multimedia Tools Appl. **70**(1), 1-26 (2013).

[22] Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. CVIU **104**, 249-257 (2006).

[23] Ryoo, M., Chen, C., Aggarwal, J., Roy-Chowdhury, A.: An overview of contest on semantic description of human activities. ICPR, 270-285 (2010).

[24] Jan, T.: Neural network based threat assessment for automated visual surveillance, IEEE Int. Joint Conf. Neural Networks, (2004).

[25] Sanromà, G., Patino, L., Burghouts, G., Schutte, K., Ferryman, J.: A unified approach to the recognition of complex actions from sequences of zone-crossings. Image Vis. Comp. **32**(5), 363-378 (2014).

[26] Ko, T.: A survey on behavior analysis in video surveillance for homeland security applications. IEEE Appl. Imagery Pattern Recognition Workshop, (2008).