Long-term behavior understanding based on the expert-based combination of short-term observations in high-resolution CCTV

Klamer Schutte [1], Gertjan Burghouts, Nanda van der Stap, Victor Westerwoudt,
Henri Bouma, Maarten Kruithof, Jan Baan, Johan-Martijn ten Hove

TNO, Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands

ABSTRACT

The bottleneck in situation awareness is no longer in the sensing domain but rather in the data interpretation domain, since the number of sensors is rapidly increasing and it is not affordable to increase human data-analysis capacity at the same rate. Automatic image analysis can assist a human analyst by alerting when an event of interest occurs. However, common state-of-the-art image recognition systems learn representations in high-dimensional feature spaces, which makes them less suitable to generate a user-comprehensive message. Such data-driven approaches rely on large amounts of training data, which is often not available for quite rare but high-impact incidents in the security domain. The key contribution of this paper is that we present a novel real-time system for image understanding based on generic instantaneous low-level processing components (symbols) and flexible user-definable and user-understandable combinations of these components (sentences) at a higher level for the recognition of specific relevant events in the security domain. We show that the detection of an event of interest can be enhanced by utilizing recognition of multiple short-term preparatory actions.

Keywords: Surveillance, CCTV, security, image understanding.

1. INTRODUCTION

The current technological trend is that availability of sensor data is no longer the most important issue. More and more sensor are getting deployed with decreasing cost, and networking techniques allow access of this data at many places. Due to this development, the bottleneck in obtaining situation awareness (e.g., target acquisition and target validation) is transferred from the pure sensing domain to the data interpretation domain. Given the exponential growth in the amount of sensor data as driven by Moore's law it is not affordable to increase human data-analysis capacity at the same rate as more data becomes available. The specific problem addressed in this paper is to aid a human image analysist, up to the level that the majority of image interpretation is performed by the automated image-understanding system, and the human analyst is alerted when there occurs an event of his interest.

The current state of the art image recognition systems (e.g., deep learning systems [22] and bag-of-word approaches [5][6]) rely on a data-driven approach with very large amounts of representative training data to learn representations in feature space of concepts, such as objects [22], actions [2] or events [4][3]. Typically these feature-spaces are extremely high dimensional and thus their representations are not explainable to human users. In addition, due to their need for large amounts of training data they are not very flexible w.r.t. changes in scenes and they are not very robust to new situations. Furthermore, the relevant incidents in the security domain are rare events and a large collection of training material does not exist. Earlier work [20] relied on a recognition of complex behavior using stochastic grammar matching. The computational complexity of stochastic grammar matching does not allow to use it in combination with normal visual tracking as it will generate too much possible hypotheses in trying to fix tracking errors such as track fragmentation.

The objective of this paper is to have a system that has a dedicated high-level image understanding and reasoning capability built into it to separate interesting rare events from a common background. In this situation, the high-level knowledge can be explained to, as well as specified by, humans and it can be robust to changes in the situation as well as allow rather easy application in completely new scenarios. Our use of expert knowledge allows easy generalization with minimal training examples. In our experiments, we focused on scenarios that are relevant for defense and security, such

[1] klamer.schutte@tno.nl; phone +31 888 66 4054; http://www.tno.nl

K. Schutte, G. Burghouts, N. van der Stap, V. Westerwoudt, H. Bouma, M. Kruithof, J. Baan, J.M. ten Hove,
"Long-term behavior understanding based on the expert-based combination of short-term observations in high-resolution CCTV,"
Proc. SPIE, Vol. 9995, (2016). http://dx.doi.org/10.1117/12.2241180

as a guard that is followed by a terrorist, terrorists that are placing an improvised explosive device (IED) near the road, terrorists that try to climb the fence of a compound, or terrorists that are hiding to prepare an ambush.

The key contribution of this paper is that we present a novel real-time system for image understanding based on generic instantaneous low-level processing components (symbols) and flexible user-definable and user-understandable combinations of these components (sentences) at a higher level for the recognition of specific relevant events.

The outline of this paper is as follows. Section 2 describes the system for image understanding based on generic symbols and flexible user-definable sentences. Section 3 describes the experimental setup and it presents the results. Finally, Section 4 summarizes the conclusions.

2. METHOD

2.1 Framework overview

In this paper, we present a system that performs real-time image analysis to recognize events of interest in the scene. Input of the system consists of video streams of surveillance cameras and output of the system is a user-understandable summary of interesting events in the form of short messages (e.g., tweets). The system consists of four main components, as shown in Figure 1. The first component performs preprocessing, which consists of person detection, person tracking and localizing these to the contextual patterns, such as roads and exit regions. The second component performs direct instantaneous processing of short fragments. We will call these fragments 'symbols'. The third component performs long-term behavior analysis based on expert-defined knowledge. The generic symbols are used and combined as building blocks to recognize specific events of interest. We will call the user-definable combinations of symbols 'sentences'. Finally, in the fourth component describes the recognized events and generates a user comprehensive summary. Each of the components is described in more detail in the following subsections.

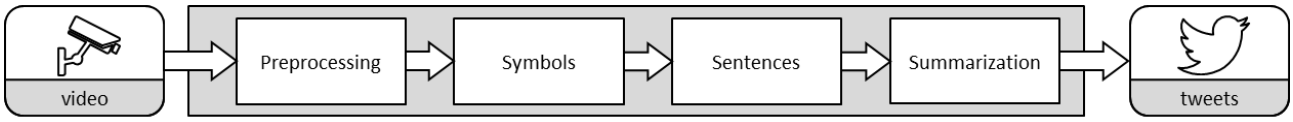

Figure 1: Overview of the system.

2.2 Preprocessing: preparation and context

The person tracking system performs person detection and tracking of every pedestrian in a camera [1][9]. The tracking module aims to avoid fragmentation and id-switches and it aims to handle occlusions, lighting variations and pedestrian size variations (expected deterioration below 60 pixels in height). The module can handle high-definition video streams of a resolution of 1920x1080 pixels at 30 FPS, although internally it performs some steps at a lower resolution and frame rate. Note that parts of the experiments had been performed on a larger image with 5120x3480 pixels which cannot be processed in real-time. The generated tracks contain information about position, time and bounding boxes and they are the basis for further behavior analysis.

For the analysis of normal and deviant behavior in static surveillance cameras it is important to obtain context information. The context information can be added by a human, by manually indicating regions in an image, or by a machine, by automatically recognizing patterns in historically recorded video and tracks. In our approach, we focus on three types of regions with contextual information: visible roads, tunnels and entry/exit regions. The visible roads are defined as clearly visible busy regions in the environment. The tunnels are defined as invisible roads connecting visible busy regions. The entry/exit regions are regions where people enter or leave the visible environment. Each of these three is explained below in more detail.

The visible roads are busy regions with many tracks and they are identified by generating a low-resolution heat map of the environment. In a large environment, visibility of the roads can be hindered by trees, bushes or buildings. When people travel from one visible region to another visible region, they may be invisible during the transition from one to the other region. This may hinder long-term behavior analysis. Recently, several solutions have been presented for the automatic inference of inter-camera topology [13]. These methods derive the most probable camera transitions and an

estimation of the travel time. Others have used topology inference for multiple cameras and we used it to infer the topology between multiple regions within the same large 5120x3480 camera view. We will call the invisible connections 'tunnels' and they are found by correlating the tracks leaving one visible region with the tracks entering another visible region. In our experiments tunnels typically are paths occluded by trees. This roads and tunnels model the motion paths of people within the environment under investigation. Finally, there are also regions at the boundary of the environment where people leave the environment (without reappearance) or enter the environment (without earlier visibility). The visible roads, tunnels and entry/exit regions are used to describe the context and enhance long-term behavior analysis.

2.3 Symbols

In order to generate meaningful descriptive information from a track, so-called symbols are generated. Each symbol describes the current state of a track for a fixed period of time (e.g., 1.0 second). This symbol description is the result of a set of filter conditions that are applied to the fundamental information in a track, e.g. velocity, location and size of the bounding box. The use of such filter-based conditions allows the generation of a description that is human understandable. An example of this is the symbol 'walks on road', of which the two most important filter conditions are: distance to the road has to be less than 2 meters, and the track speed is between 3 and 10 km/h.

The symbols are generated from the tracks. Some other examples are: walk off road, loiter, dig, etc. The symbols have a description that is human understandable and it is a combination of multiple conditions. A naive user shall easily make an additional symbol by combination of some conditions.

Each track can generate one or multiple symbols for each moment, when multiple symbols are simultaneously true. When the track at a certain moment does not fit to any of the conditions, an *Empty* symbol is generated. These *Empty* symbols are used in the high-level sentences as a keep-alive symbol to follow the track of a person and enhance long-term behavior analysis, so the sentence can be continued and connected to the track, also in cases of meaningless or empty symbols.

Specific scenarios may require specific symbols. For example, the 'Follow' scenario requires a symbol that estimates the distance between a follower and a followee (Figure 2). A real-time system leads to a time constraint that prevents the use of data in the future. Only the data that was recorded up to the current point in time shall be taken into account. This has specific consequences for the follow-scenario. Only the followers can instantaneously be detected, since a followee often is in the field-of-view before the follower is, and no visible interaction can be identified between them at that moment. When a follower fulfills the requirements, the track-ID of the most likely followee was added as a symbol to the follower-track to identify the relation between a follower and the followee. The Follow scenario requires interaction between tracks during a longer period of time, which requires memorization in the symbols of the distance to the nearest track and the start time of the follow action. The key conditions that an expert uses are based on path and direction correlation with time delay, and speed correlation without time delay. For the Follow scenario, too short tracks (< 1 second) and tracks that do not show any displacement (e.g. erroneous tracking of trees) were excluded. Furthermore, tracks within 10 seconds of the same location were classified as people walking together instead of people following each other. Other track filter features are: the distance between the current track and the closest track, the followed track ID (since the follower could only be detected), the distance over which this track has been followed and the timestamp of the beginning of the follow action (to compute the duration).

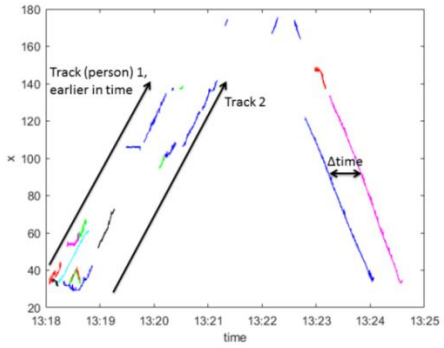

Figure 2: Example of person tracks, of which one is following the other. Note the relatively constant time delay between the two major paths and the track fragmentation. The three arrows are overlaid to indicate the tracks of the first person, the tracks of the second person and the time delay.

The ‘IED-dig’ scenario aims to retrieve events where people hide an explosive near the road and it involves three symbols: two persons walking on a road, loitering for some time close to the road, and digging. The first symbol represents ‘walking on the road’ and is encoded as: distance to road equals 1 meter or less, track speed equals between 3 and 10 km/h, circular standard deviation of the track is between 0 and 0.1 meter, the absolute time-derivative of the distance to the road ranges between 0 and 0.5 meter. The fact that there is a second person is not taken into account for this symbol, because due to the persons walking side-by-side the other person is often occluded. The second symbol of IED-dig represents ‘two persons loitering close to the road’ and is encoded as: distance to road equals 0.5 meter or less, track speed is less than 3 km/h, the absolute time-derivative of the distance to the road ranges between 0 and 0.25 meter per second, and the distance to another person is less than 3 meter. The third symbol of IED-dig represents ‘Digging’ and is encoded as: distance to road equals 0.5 meter or less, track speed is less than 1.5 km/h, the absolute time-derivative of the distance to the road ranges between 0 and 0.25 m/s, and the vertical variation is larger than 0.1 meter. The latter variable captures the typical upper-body up-and-down activity during the digging. An obvious way to improve the discriminative power of this variable is to replaced it by an action classifier [2][5][6] but that is not the goal of this paper.

The ‘Ambush’ scenario aims to detect suspicious disappearances of persons in the scene, which is technically similar to a ‘Hide’ scenario. The symbols are not only used to measure the behavior during the life-time of a track, but also after termination of the track, which is essential for disappearing persons. A disappearing person may relate to people permanently leaving the environment in an exit region, people temporarily leaving the visible regions by entering a tunnel, or people that are actually hiding or ambushing. For this scenario it was chosen to focus on encoding the interaction of a subject and its track with specific zones in the environment. With respect to using a symbol-notation this means that the Ambush scenario consists of the following two symbols: ‘Walk’ and ‘Disappear’. The ‘Walk’ symbol requires that the subject is walking. This symbol filters the tracks on speed only. Any track speed between 2 and 10 km/h is given symbol Walk, effectively eliminating false detections of stationary objects as well as fast moving false detections. The ‘Disappear’ symbol requires that the subject disappeared. At the end of each track a symbol is generated at the location of termination. Based on the location of this symbol, and thereby the interaction it has with specific zone information, a specific maximum timespan is given to this symbol before it is translated into a warning message on a sentence-level. For example, when the ‘Disappear’ symbol occurs in an exit-zone (i.e. a zone where a road leaves the environment) the symbol will not trigger an alarm. In another case, where the subject disappears at a known location of a tunnel or occluded road, a maximum disappearance time of 20 seconds is assigned to the symbol before triggering an alarm. In the case that a track disappears in a location without special zone information – which could be a potential ambush location – a maximum disappearance time of 5 seconds is assigned to the symbol. This approach of assigning symbols based on zone information means that on a sentence level a subject that disappears from the scene and reappears before the maximum disappearance time has passed will be attributed to the same sentence, re-allocating the first symbol to the sentence. Also it means that a sentence that ends with the second symbol outside any of the exit-zones, signifies a suspicious disappearance and therefore triggers an alarm.

2.4 Sentences

The symbols are connected to sentences, using a grammar which describes the symbol order and duration in the sentence. The connection of the symbols to sentences is described by the grammar by specifying which transitions are allowed between states. The grammar describes the order of the expected symbols. A sentence can only be started by one or more specific symbols. A symbol can be connected to an existing sentence when it fits to the grammar and the spatial-temporal limits as expressed for the transitions are not exceeded. Each symbol describes a short-time behavior, where the sentence describes the long-time behavior by combination of the proper short-time behaviors. For a specific scenario, the user has to specify the expected behavior. For example: a suspect walks on the road during more than 10 seconds (symbol: “walk on road”), after that the suspect will stop an looking around for 3-10 seconds (symbol “loiter on road”), then the suspect will leave the road an run off-road (symbol “run of road”). The sentences have two major advantages over detecting only single short-time behavior: (1) the user is able to specify complex long-duration actions and (2) false alarms of single short-time behavior detection are reduced by using long-term behavior. Normally a short-term event will be preceded by preparatory behavior. With the sentences, the preparation can be taken into account and thereby will lead to a reduction of false alarms.

When a suspect is occluded and not visible, the spatial position of the sentence over time will be predicted by the speed and direction of the last symbol. When more objects appear after an occlusion, a copy of the sentence is made, so all of them are a potential suspect. This approach is similar to multi-hypothesis tracking and it allows handling fragmentation due to track breaks or tunnels. Only the sentence which reaches the end of a grammar will create an alarm. The number

of sentences can explode due to the copying of sentences. The number of sentences can be reduced by selecting of only the most probable sentences and remove sentences with low probability.

The grammar intends to detect scenarios by connecting the symbols that represent each respective part of the scenario. For example, IED-dig connects the three symbols: walking on road, loitering with two people, and digging. The objective is to increase the discriminative power of digging detections, such that an operator of the system has to validate less detections and can perform the surveillance task more efficiently and possibly more effectively. The grammar is proposed to reduce false positives of digging while maintaining the true positives, by requiring the specific order of the three symbols, each observed at a specific minimum and maximum duration. Each of the individual symbols may not be very discriminative, e.g., many people traverse the scene by walking on the road, and there will be many detections of digging as we do not employ an advanced detector (recall that we measure it simply by vertical variation). By searching for sentences of symbols 1 to 3 in a sequence with specified respective durations, the grammar aims to reduce the number of digging detections significantly. The durations for each state are respectively: minimally 3 to maximally 20 seconds; 0 to 5 seconds; and 10 to 20 seconds. An example of a valid symbol sequence is shown in Figure 3.

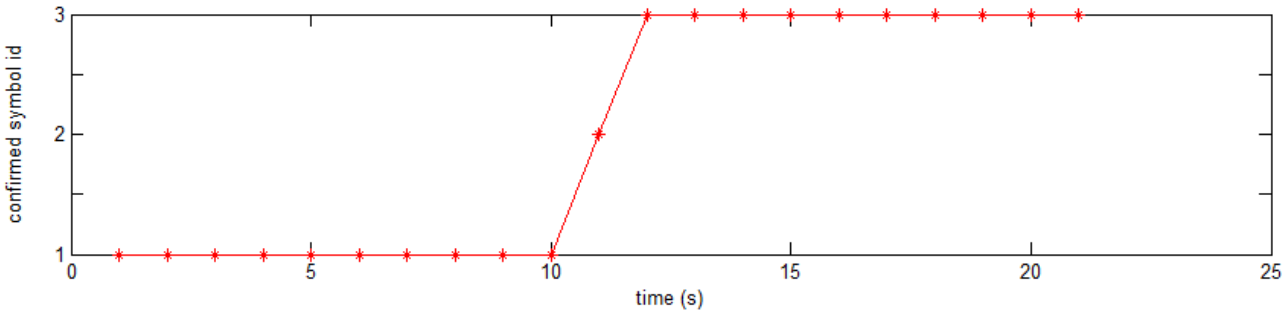

Figure 3: Example of a valid symbol sequence for the IED-dig scenario. The figure shows a transition from symbol 1 (‘walk on road’ on the left), to symbol 2 (‘loiter’ in the middle), to symbol 3 (‘digging’ on the right).

In essence, the sentence generation performs for every time step an association of new symbols to existing sentences. This is performed when it is allowed given the current state of the sentence and when the spatio-temporal constraints allow so. Note that track ID is not used in this association, to allow to correct for earlier track fragmentation errors. Use of *Empty* symbols allows to follow a person (expressed as set on fragmented tracks) over a longer period.

2.5 Summarization and explanation

The ‘Summarization’ component reports and clarifies the events that were encountered in the video. The output of our system is a user-understandable summary of interesting events in the form of a Twitter message that contains textual and visual information. The textual information describes the event in words, for example, by reporting that “Track <trackid> has been following track <followed_trackid> for n seconds at an average distance of <x> meters.” Multiple detections within a short period on the same track for the same event were suppressed to avoid an information overload. The visual information summarizes the event with an image containing the relevant actors and metadata information, such as bounding boxes and track-ID’s. The textual information allows a user to quickly search in historic data and the visual information allows rapid assessment and validation of the current situation.

3. EXPERIMENTS AND RESULTS

For the evaluation of our system, we used a recorded dataset (Sec. 3.1 and 3.2) and live data (Sec. 3.3).

3.1 Offline experimental setup and use cases

The “TwitCam Dunes Dataset” was recorded in 2015 in by TNO in The Hague. The dataset contains high-resolution video with a resolution of 5120x3840 pixels. The duration is almost 3 hours and it includes multiple people acting multiple scenarios at multiple locations. Most of the scenarios are recorded with five repetitions with positive and near-positive (hard negative) examples and varying volunteers.

The dataset focusses on use cases and scenarios that may be relevant in the defense and security domain. The following seven scenarios were recorded:

- Weapon: A 'terrorist' walks around carrying a weapon.
- Follow: A guard is followed by a terrorist.
- IED-dig: Two terrorists place an IED near the road.
- IED-hide: Three terrorists place an IED and hide when a patrol comes by.
- Fence: Two terrorists try to cross the fence
- Observation: A terrorist makes pictures of the compound for an attack
- Ambush: Terrorists that leave the road and suddenly hide in the bushes.

Within the work we concentrate on the Follow, IED and Ambush scenarios as the other scenarios are better suited for instantaneous event recognition rather than long term recognition. For the IED scenarios we for now concentrated on the IED-dig.

3.2 Results on the dataset

An overview of the scene with colored tracks is shown in Figure 4. The figure shows that people are clearly visible on many of the roads and that the trees are occluding several parts of the roads.

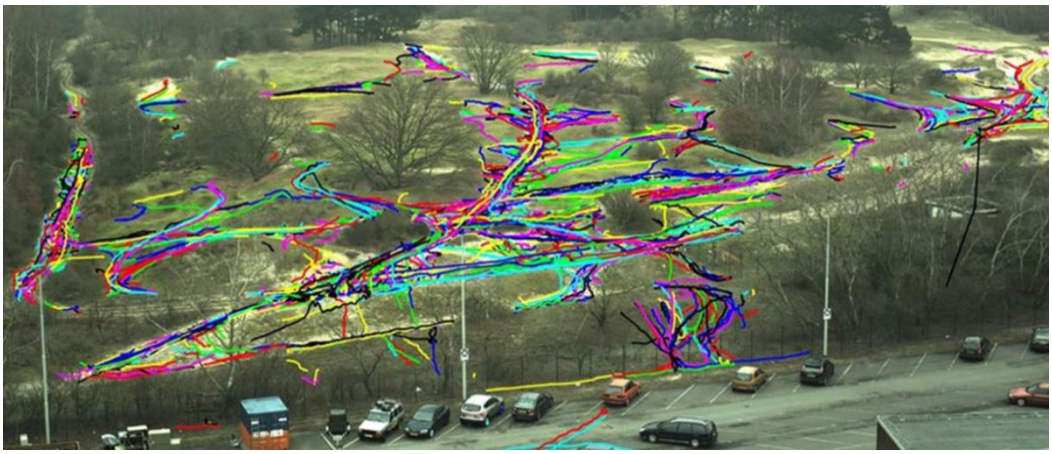

Figure 4: Tracks in the TwitCam Dunes Dataset.

Based on the tracks in the dataset, busy regions can automatically be segmented. The black curves in Figure 5 indicate the contours of these busy regions. Frequent transitions and high correlations between parts of these contours indicate the tunnels, which people use to travel from one visible region to another visible region. In the foreground (at the bottom of the image), the tunnels are derived correctly and they indicate the location of common transitions. In the background (at the top of the image), several tunnels are missing because the current dataset does not contain sufficient tracks to recognize the transitions.

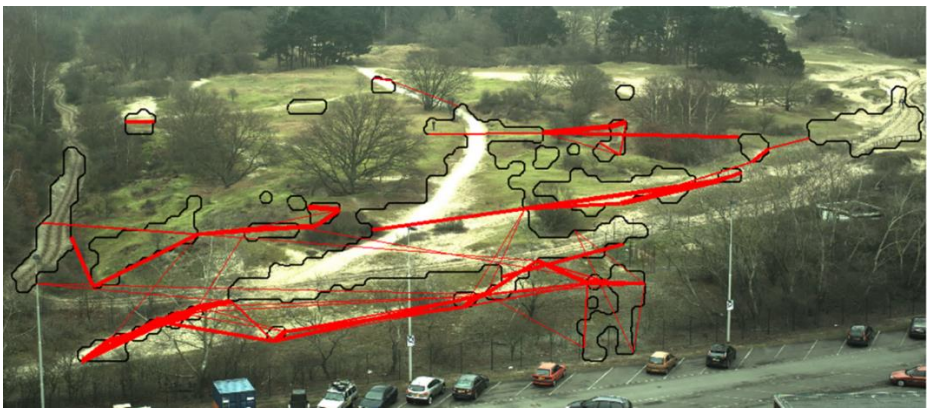

Figure 5: Segmented busy regions (black curves) and tunnels (red lines) derived from the dataset.

Figure 6 shows two examples of the scenario ‘Ambush’. The visualizations include the path of symbols found by the Ambush grammar. The left figure shows people walking on the road (left: a, b) and then they leave the road and become invisible when entering the bushes (left: c). The right figure shows people walking on the road (right: a) and they leave the road and enter the bushes for a short period (right: b) and continue leaving the road for a longer period (right: c). The system gives an alarm after a few seconds. A near-miss example is shown in Figure 7. The people do not hide in the bushes for a long time, but they walk on the path behind the bushes and appear after a few seconds at the other side of the bush. The time that the people are invisible for the camera is measured and when they reappear on one side, the detection is automatically connected to the track of the disappearing people on the other side.

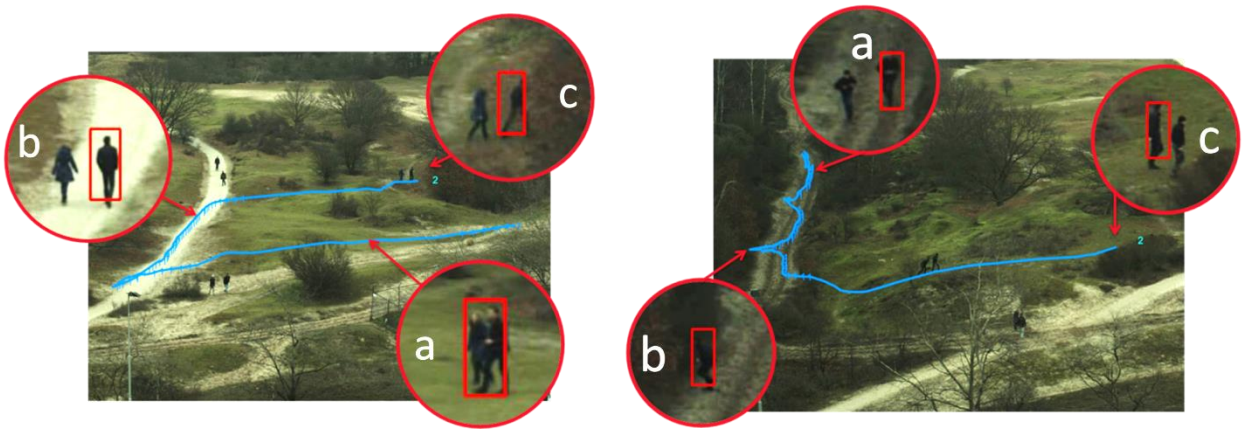

Figure 6: Two example images of scenario Ambush.

Figure 7: A near-positive example of the scenario Ambush. Two people walk on the path (top left). The rectangles indicate the automatic detections and the curve indicates the track of these detected persons. They disappear behind bushes (top right). The current undetected location is indicated manually for this publication with a red ellipse. After approximately 20 seconds, they appear again on the other side of the bushes (bottom left). When the persons are clearly visible, they are automatically detected and connected to the lost track (yellow line).

An example of the Follow scenario is shown in Figure 8. This figure shows a person (P1) who walks through the scene and is followed by another (P2). During this time, the follower can for example choose to hide from the followee (e.g. behind a bush), direct his attention elsewhere (lean on a fence) or adjust his tempo to keep an optimum distance to the followee. The video frame at the beginning of the track of person 1 was selected, although the sentence was recognized at a later moment.

Figure 8: Example image of scenario Follow (detected duration is approximately 60 seconds).

Figure 9 illustrates three different events that have been recorded for the IED-dig scenario. The scenario consists of three parts: two persons walking on a road, loitering for some time close to the road, and digging. Each event involves all three parts at three different times, at two different locations in the scene (80 meters apart), with different actors and different durations of each of the parts.

Figure 9: Example images of IED-dig in three scenarios. The visualizations include the path of symbols found by the IED-dig grammar by means of colors (from blue to green, best viewed in color). The figures show a person digging.

Table 1 indicates the number of true positives (TP), false positives (FP) and false negatives (FN) for sentences generated by the grammar. The sentences produced by the grammar are evaluated by the following procedure. First, we remove sentences that are highly similar, i.e., which have evolved to the same sequence (but with slight variations due to multiple detections in a small spatiotemporal volume). Second, we assess a sentence as correct if it intersects with the ground truth, where we allow some slack due to a small misalignment in some predefined spatiotemporal volume of maximal 3 meter and 3 seconds.

With the IED-dig grammar, all three scenarios are correctly detected, at the cost of one false positive. The duration parameters of the grammar are the variables that determine the performance most. If these duration variables are increased by a factor of two, all scenarios are detected but at a larger cost of 13 false positives. More generalization is achieved as more scenarios will be detected, yet more alarms will have to evaluated by the operator of the system. Whether this is acceptable, depends on the task, priority and the working process of the operator.

The Ambush sentence leads to 46 detection, of which 19 correspond to people that leave the road and hide in the bushes and 27 correspond to false positives. The dataset contained 4 other ambush scenarios that were missed by the system.

Table 1: Grammar sentences: detection results.

	TP	FP	FN
IED-dig	3	1	0
Ambush	19	27	4

Table 2 indicates the number of detections for the symbols on which the IED-dig sentences are based. Clearly, there are many more detections of symbols, up to 7747 for the first symbol (walking on road) to 573 for the final symbol (digging). Obviously, many people that traverse the scene are walking on the road. Therefore, the first symbol is not very discriminative. Digging does not happen often in the scene (only by our actors). The high number of detection is due to our very basic implementation of the digging symbol, mainly driven by vertical variation of the person box while standing still. This happens often, which explains the high number of detections. Interestingly, when searching for sentences of symbols 1 to 3 in a sequence with specified respective durations, the grammar is able to reduce the number of digging detections significantly to only 4 detections. This indicates the merit of the proposed method: the system operator has to validate few alarms while there is a large probability that the alarm relates to the target scenario.

Table 2: Number of detections for symbols and grammar sentences for IED-dig.

IED-dig	# Detections
Symbol 1: walk on road	7747
Symbol 2: loiter	1142
Symbol 3: dig	573
Grammar sentence	4

Table 3 indicates the number of detections for the Ambush scenario. Here we also observe that a concatenation of subsequent filtering steps leads to a reduction of the number of false alarms. Initially, the end of each track could indicate the disappearance and hiding of a person. This could lead to almost 1800 detections. Application of additional filtering conditions, such as mitigation of track fragmentation, the duration of a sentence, removal of duplicates (caused by the multi-hypothesis symbol transitions), removal of sentences in exit zones, removal on roads or at the entrance of a tunnel and a focus on suspicious areas without tracks help to reduce the number of alarms to only 46.

Table 3: Number of detections for scenario Ambush.

Ambush	# Detections
Tracks	1795
First order sentence (end of track)	1074
Remove short sentences	597
Remove duplicates	199
Remove sentence in exit zones	104
Remove roads / tunnels	62
Select only near untracked area	46
Grammar sentence	46

Figure 9 and Figure 10 show the detections for IED dig, respectively the three correct detections (true positives) and the single error (false positive). The visualizations include the path of symbols found by the IED-dig grammar by means of colors (from blue to green) and lines (connected symbols). The grammar is able to connect the stages to tracks even when some occurrences of appropriate symbols have been missed (indicated by 'NaN' in the figure). It is clear that the symbols are not perfect: there are confusions between symbols and occasionally they are missed. The grammar finds the IED-dig path through these imperfect symbols. The false positive exhibits the pattern of two people walking, then holding positions, and moving their body a little while loitering. The latter part is mistaken for digging, because it looks similar to digging according to our simple implementation of this final symbol of the IED-dig grammar.

Figure 10: Example images of a false positive of the IED-dig scenario. The system incorrectly recognizes a person digging.

3.3 Live experimental setup and demonstration

The system was applied to the recorded material of the "TwitCam Dunes dataset" and also on live video streams to demonstrate the real-time processing capabilities. The live experimental setup was aimed at a part of the original dunes scene from the recorded dataset. The live demonstration uses video with a rate of 4 frames per second and a resolution of 1920x1080 pixels. When one of the sentences/scenarios is detected, a tweet is generated that includes a short text message and an image. Examples for the Follow scenario are shown in Figure 11.

Figure 11: Example images of the Follow scenario on live data. The person with the green bounding box is followed by the person with the red bounding box (best viewed in color). The follower and followee were correctly identified in different weather conditions and in different locations (e.g. on or off path).

The performance of the Follow grammar was evaluated by manual inspection of the generated Twitter data over a period of one week. A total of 36 tweets was generated indicating a 'follow' situation. Of these 36 detections, 22 were reported by unique persons and 14 messages were caused by the same persons being signaled twice. This happened especially if a group of people were either followed or following. The tracks then switched between the persons in the group, each time causing a new tweet. In addition, this situation could occur if track fragmentation was present. A track of a person then was lost, but picked up later again, causing a double Follow tweet. These false messages could have been avoided by suppressing multiple messages in a selected time period. After correction for the duplicates, 19 of the 22 detected follow actions (86%) were reported correctly (Table 4).

Table 4: Grammar evaluation for the Follow scenario in the live situation.

	TP	FP	Total
Follow (all detections)	24	12	36
Follow (unique detections)	19	3	22

The real-time processing of activity in the scene also allows rapid inspection of periodic patterns. Figure 12 shows histograms of these patterns. The left figure shows the hours over a day and the right figure shows the days of the week. It is clear that near lunch time (13.00h) is the busiest time of the day and Sunday is the busiest day of the week.

Figure 13 shows the number of tracks for each day in January 2016. This allows comparison with days with more than 6 sun hours (8, 17, 18 and 28 Jan), days with more than 10 mm rain (4, 7, 12, 14, 27 Jan), and days with hardly sun hours (4 – 7, 11, 14, 21, 24, 27 Jan). [2] We observe that the most busy day was a sunny Sunday (17 Jan) and the quietest day was a Thursday with hardly sunshine (21 Jan). The busiest Friday was also a day with many sun hours (8 Jan) and the quietest Sunday was a day without sunshine (24 Jan).

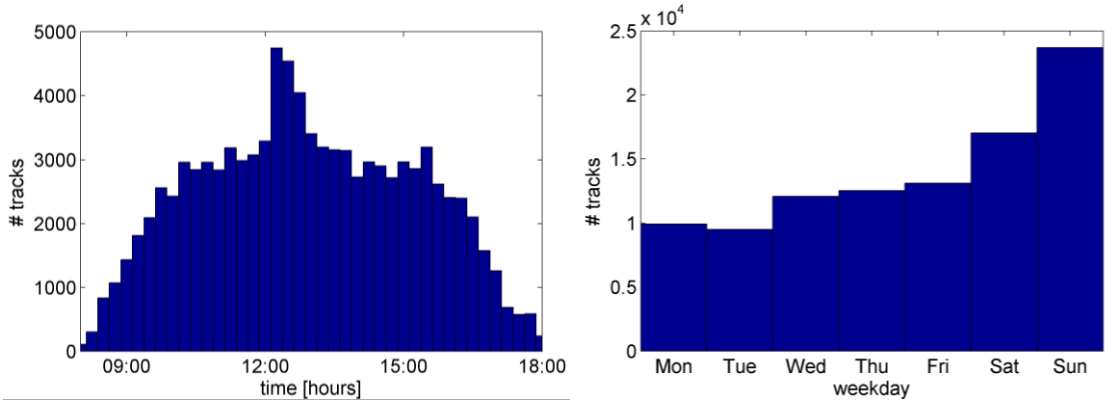

Figure 12: Periodic analysis of the live information over 3 winter months. The left figure shows that the dunes are more busy around lunch time (13.00h) than on other times. The right figure shows that Sunday is the busiest day of the week.

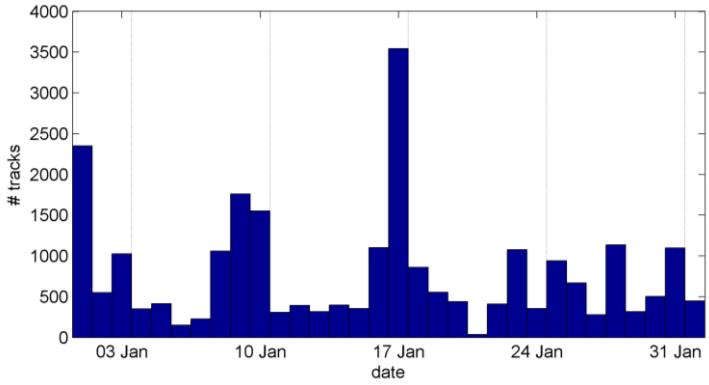

Figure 13: Number of tracks per day in January 2016.

[2] http://nieuws.weeronline.nl/28-1-2016-januari-2016-warm-nat-en-aan-de-zonnige-kant/

4. CONCLUSIONS

In this paper, we presented a novel real-time system for image understanding based on generic instantaneous low-level processing components (symbols) and flexible user-definable and user-understandable combinations of these components (sentences) at a higher level for the recognition of specific relevant events in the security domain. We showed that the detection of an event of interest can be enhanced by combining it with multiple short-term preparatory actions.

ACKNOWLEDGEMENT

The work for this paper was conducted in the project "TwitCam: Image Understanding", which was funded by the MIST research program of the Dutch Ministry of Defense.

REFERENCES

[1] Bouma, H., Baan, J., Landsmeer, S., Kruszynski, C., Antwerpen, G., Dijk, J., "Real-time tracking and fast retrieval of persons in multiple surveillance cameras of a shopping mall," Proc. SPIE 8756, (2013).
[2] Bouma, H., Hanckmann, P., Marck, J., Penning, L., et al., "Automatic human action recognition in a scene from visual inputs," Proc. SPIE 8388, (2012);
[3] Bouma, H., Baan, J., Burghouts, G., Eendebak, P., Huis, J. van, Dijk, J., Rest, J. van, "Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall," Proc. SPIE 9253, (2014).
[4] Burghouts, G., Schutte, K., Hove, R. ten, et al., "Instantaneous threat detection based on a semantic representation of activities, zones and trajectories," Signal Image and Video Processing 8(1), 191-200 (2014).
[5] Burghouts, G., 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," Machine Vision and Applications 25(1), 85-98 (2014).
[6] Burghouts, G., Eendebak, P., Bouma, H., Hove, J.M. ten, "Improved action recognition by combining multiple 2D views in the Bag-of-Words model," IEEE Int. Conf. Advanced Video and Signal-Based Surveillance AVSS, 250-255 (2013).
[7] Cheng, D. S., Setti, F., Zeni, N., Ferrario, R., & Cristani, M., "Semantically-driven automatic creation of training sets for object recognition," Computer Vision and Image Understanding 131, 56–71 (2015).
[8] Coppi, D., Campos, T. de, Yan, F., "On detection of novel categories and subcategories of images using incongruence," Proc. Int. Conf. Multimedia Retrieval, 337 (2014).
[9] Dollar, P., Appel, R., Belongie, S., Perona, P., "Fast feature pyramids for object detection," IEEE Trans. Pattern Analysis and Machine Intelligence 36(8), 1532-1545 (2014).
[10] Elhoseiny, M., Saleh, B., Elgammal, A., "Write a classifier: Zero-shot learning using purely textual descriptions," IEEE Int. Conf. Computer Vision, 2584–2591 (2013).
[11] Habibian, A., Mensink, T., Snoek, C., "VideoStory: A New multimedia embedding for few-example recognition and translation of events," ACM Int. Conf. Multimedia, 17-26 (2014).
[12] Hanckmann, P., Schutte, K., Burghouts, G., "Automated textual descriptions for a wide range of video events with 48 human actions," ECCV LNCS 7583, 372-380 (2012).
[13] Hollander, R., Bouma, H., Baan, J., Eendebak, P., Rest J. van, "Automatic inference of geometric camera parameters and inter-camera topology in uncalibrated disjoint surveillance cameras," Proc. SPIE 9652, (2015).
[14] Karpathy, A., & Fei-Fei, L. "Deep visual-semantic alignments for generating image descriptions, IEEE CVPR, 3128-3137 (2015).
[15] Kunze, L., Burbridge, C., Alberti, M., Tippur, A., Folkesson, J., Jensfelt, P., Hawes, N., "Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding," IEEE/RSJ Int. Conf. Intelligent Robots and Systems, 2910–2915 (2014).
[16] Lampert, C. H., Nickisch, H., Harmeling, S., "Attribute-based classification for zero-shot visual object categorization," IEEE Trans. Pattern Analysis and Machine Intelligence 36(3), 453–65 (2014).
[17] Li, L., Socher, R., Fei-Fei, L., "Towards total scene understanding: classification, annotation and segmentation in an automatic framework," IEEE CVPR, 2036-2043 (2009).

[18] Norouzi, M., Mikolov, T., Bengio, S., et al., "Zero-shot learning by convex combination of semantic embeddings," Int. Conf. Learning Representations, (2014).
[19] Rohrbach, M., Stark, M., & Schiele, B., "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," IEEE CVPR, 1641–1648 (2011).
[20] Sanroma, G., Patino, L., Burghouts, G., Schutte, K., Ferryman, J., "A unified approach to the recognition of complex actions from sequences of zone-crossings," Image and Vision Computing 32(5), 363-378 (2014).
[21] Satsangi, Y., Whiteson, S., Oliehoek, F., "Exploiting submodular value functions for faster dynamic sensor selection," Proc. AAAI Conf. Artificial Intelligence, 3356-3363 (2015).
[22] Schutte, K., Bouma, H., Schavemaker, S., e.a., "Interactive detection of incrementally learned concepts in images with ranking and semantic query interpretation," IEEE Content-Based Multimedia Indexing, (2015).
[23] Socher, R., Ganjoo, M., "Zero-shot learning through cross-modal transfer," Adv. NIPS, 935-943 (2013).
[24] Turakhia, N., Parikh, D., "Attribute dominance: What pops out?" IEEE ICCV, 1225–1232 (2013).