Track-based event recognition in a realistic crowded environment

Jasper R. van Huis, Henri Bouma [1], Jan Baan, Gertjan J. Burghouts, Pieter T. Eendebak,
Richard J.M. den Hollander, Judith Dijk, Jeroen H.C. van Rest

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

ABSTRACT

Automatic detection of abnormal behavior in CCTV cameras is important to improve the security in crowded environments, such as shopping malls, airports and railway stations. This behavior can be characterized at different time scales, e.g., by small-scale subtle and obvious actions or by large-scale walking patterns and interactions between people. For example, pickpocketing can be recognized by the actual snatch (small scale), when he follows the victim, or when he interacts with an accomplice before and after the incident (longer time scale). This paper focusses on event recognition by detecting large-scale track-based patterns. Our event recognition method consists of several steps: pedestrian detection, object tracking, track-based feature computation and rule-based event classification. In the experiment, we focused on single track actions (walk, run, loiter, stop, turn) and track interactions (pass, meet, merge, split). The experiment includes a controlled setup, where 10 actors perform these actions. The method is also applied to all tracks that are generated in a crowded shopping mall in a selected time frame. The results show that most of the actions can be detected reliably (on average 90%) at a low false positive rate (1.1%), and that the interactions obtain lower detection rates (70% at 0.3% FP). This method may become one of the components that assists operators to find threatening behavior and enrich the selection of videos that are to be observed.

Keywords: Surveillance, CCTV, security, tracking, behavior analysis, action recognition, threat.

1. INTRODUCTION

To improve security, the number of surveillance cameras is rapidly increasing in crowded environments, such as shopping malls, airports and railway stations. However, the number of human operators remains limited and only a selection of the video streams are observed. Automatic detection of suspicious behavior in CCTV cameras can help to handle the huge amount of data. Suspicious behavior can be characterized by small-scale subtle and obvious actions, large-scale walking patterns and interactions between people. For example, pickpocketing can be recognized during several steps of the criminal incident [22]. Sometimes, it is possible to detect the actual snatch, by analyzing actions at a small scale. On the other hand, it may be easier to detect (a group of) pickpockets while they are following the victim or when they are interacting with each other, by analyzing actions at a larger time scale. There are several approaches to detect small-scale actions by analyzing local motion [7][9][10][25]. This paper focusses on event recognition by detecting large-scale track-based patterns.

Our event recognition method consists of several steps: pedestrian detection, object tracking, track-based feature computation and rule-based event classification. In the experiment, we focused on single track actions (walk, run, loiter, stop, turn) and track interactions (pass, meet, merge, split), because they are related to activities of pickpockets that may lead to a distinction between suspicious and normal behavior. The main contribution of this paper is that we show that these events can be detected reliably at a low false positive rate.

The outline of the paper is as follows. The method is described in Section 2. The experiments and results are shown in Section 3. Finally, the conclusions are presented in Section 4.

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

J.R. van Huis, H. Bouma, J. Baan, G.J. Burghouts, P.T. Eendebak, R.J.M. den Hollander, J. Dijk, J.H.C. van Rest, "Track-based event recognition in a realistic crowded environment," Proc. SPIE, Vol. 9253, (2014). http://dx.doi.org/10.1117/12.2066081

2. METHOD

2.1 System overview

The system consists of the following components: pedestrian detection, object tracking, track-based feature computation and rule-based event classification (see Figure 1). Each of these is described in the following subsections.

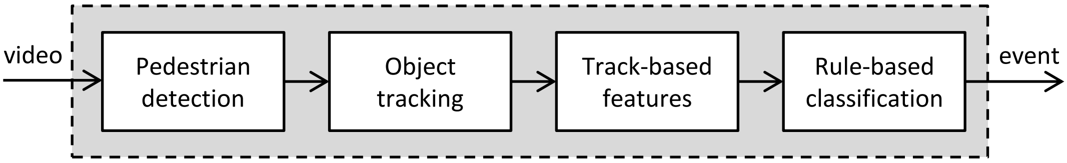

Figure 1: Overview of the method.

2.2 Real-time detection and tracking

We have a system that generates detections and tracks in each camera [4]. These tracks can be used for multi-camera tracking and re-identification [5][18][19] and further behavior analysis [3][6]. The tracks contain location information in camera coordinates (bounding boxes) as well as location information in world coordinates. In this paper, the obtained world-coordinate tracks are used to support the track-based feature computation.

2.3 Track-based feature computation

Abnormal or suspicious behavior can be characterized by actions and interactions. For example, pickpocketing can be recognized during several steps of the criminal incident [22], e.g. while observing the environment, waiting for an opportunity, communicating to accomplice, surrounding the victim, snatching something, handing over the loot to an accomplice, and leaving the scene. We identified that the following actions and interaction would be beneficial for the analysis of behavior of a group of pickpockets: walk, run, loiter, stop, turn (actions) and pass, meet, merge, split (interactions).

The basic features that allow analysis of these actions and interactions are the following:

Primary features that are computed for each track detection instance, stored per track:

- Speed (km/hour): instant walking speeds based on 1 second history (e.g., for walk, run)
- Distance (meter): instant distance traveled in the last 1 or 2 seconds (e.g., for stop)
- Direction (degrees): instant direction based on 1 second history (e.g., for turn)
- References to other tracks that have been within 2 meters distance.

Derived features and features that are computed on-the-fly while the event is assessed.

- Direction change
- Angle between tracks
- Distance between tracks (e.g., for pass, meet, merge)
- Duration of certain speed, distance or angle between tracks

2.4 Rule-based event classification

The rules are defined based on expert knowledge of the actions and interaction and some thresholds were determined after inspection of an example in the video data. A list of rules is shown in Table 1.

Table 1: Rules for the actions and interactions.

(inter)action	rule
Walk	speed > 3 km/hour & speed < 7.5 km/hour, for a duration > 2 seconds
Run	speed > 8 km/hour for a duration of > 1 second
Loiter	speed < 3 km/hour for a duration of > 4 seconds
Stop	2-second-distance < 1 meter for more than 1 second
Turn	Angle change > 135 degrees & angle change < 50 degrees in last second before turn, while only taking directions into account where speed was > 3 km/hour
Pass	2 tracks with relations in the following order: a) tracks being > 3 meters apart for > 0.7 seconds b) then < 1.5 meter apart for > 0.4 seconds & a relative angle of >150 degrees & both having a speed of > 3 km/hour c) tracks being > 3 meters apart for > 0.7 seconds
Meet	2 tracks with relations in the following order: a) tracks being > 3 meters apart for > 1 second b) tracks being < 2 meters apart for > 3 seconds & both tracks are ‘stopped’ (2-second-distance < 1 meter for more than 1 second)
Merge	2 tracks with relations in the following order: a) tracks being > 3 meters apart for > 1 second b) tracks being < 2 meters apart for > 2 seconds & both tracks have speed > 3 km/hour for > 1 second.
Split	2 tracks with relations in the following order: a) tracks being < 2 meters apart for > 2 seconds & both tracks have speed > 3 km/hour for > 1 second & relative angle < 30 degrees b) tracks being > 3 meters apart for > 1 second

3. EXPERIMENTS AND RESULTS

3.1 Experimental setup at the shopping mall

In the shopping mall, we used the following hardware. We used a camera setup with 20 network cameras of multiple types, including four AXIS-P1346 cameras with 1920x1080 resolution at 30 frames/sec and many AXIS-211M cameras with 1280x1024 resolution at 9 frames/sec. A small region was equipped with multiple cameras that are aimed at the same location and recorded a controlled set of actions in this region, similar to IXMAS [10][25]. The other cameras are hardly overlapping and they are used to cover a large region of the shopping mall, similar to [4].

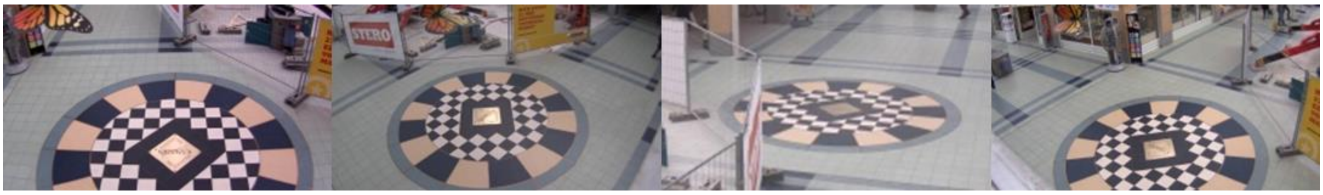

Figure 2: Four cameras are used at 30 frames/sec at one location, similar to the setup of IXMAS [25].

Our experiment was part of a larger measurement campaign that included the recording of several activities. In the small region with overlapping cameras, a controlled set of actions was performed by actors. These actions include the actions of IXMAS [25], and the actions and interactions that were defined in the previous section. During the experiment, we did not establish a perimeter, so the majority is normal behavior of passers-by. The actors performed the actions 10 times and the interactions 5 times. The events of the actors were used to compute the sensitivity of the system. Furthermore, the system generated in a period of approximately half an hour a large number of tracks and detected events. These tracks include both the tracks of actors and passers-by. Of the detected events, a small number was checked manually to analyze the number of false positives.

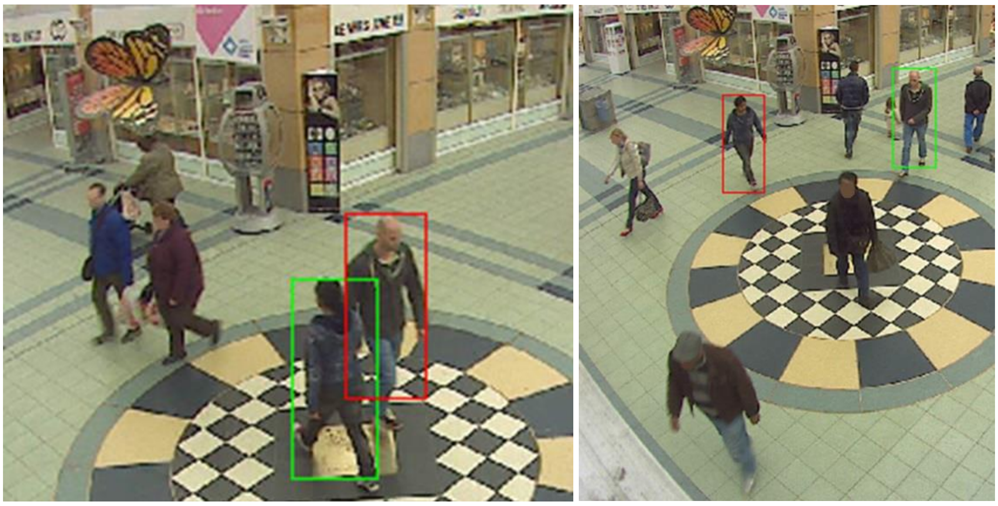

Figure 3: Examples of the action ‘pass’ (left) and ‘merge’ (right).

3.2 Results of event recognition

An example of the track localization in world-coordinates during an (inter)action is shown in Table 2.

The events of the actors (10 actions and 5 interactions) were used to compute the sensitivity of the system. The results are shown in Table 3 and Table 4. Based on the number of acts and the number of true positives (TP), we observe that the average sensitivity for actions is 90% and for interactions 70%.

The system generated in a period of approximately half an hour (between 28 and 40 minutes) more than 1000 tracks, and the number of detected events ranges between 9 and 823. These tracks include both the tracks of actors and passers-by. Of the detected events, 30 (or all detections if there are less than 30) were checked manually to compute the number of false positives (FP) and the percentage of false positives. This enables an estimation of the total number of FP in all tracks (Est. FP in tracks), the percentage of false positives per track (FP/track), and the average number of false positives per minute (FP/min). The results are also shown in Table 3 and Table 4. The average number of FP per minute appears to be 0.5 FP/min for actions and 0.2 FP/min for interactions, and the average number of FP per track appears to be only 1.1% for actions and 0.3% for interactions.

So, the results show that most of the actions can be detected reliably (on average 90%) at a low false positive rate (1.1%), and that the interactions obtain lower detection rates (70% at 0.3% FP). The small number of acts introduces high uncertainty in the scores. For example, the 90% confidence interval of the sensitivity of the actions lies between 80% and 96% sensitivity (using the available 50 samples and a binomial distribution).

Table 2: Example of location (vertical axis) versus time (horizontal axis) of six actions.

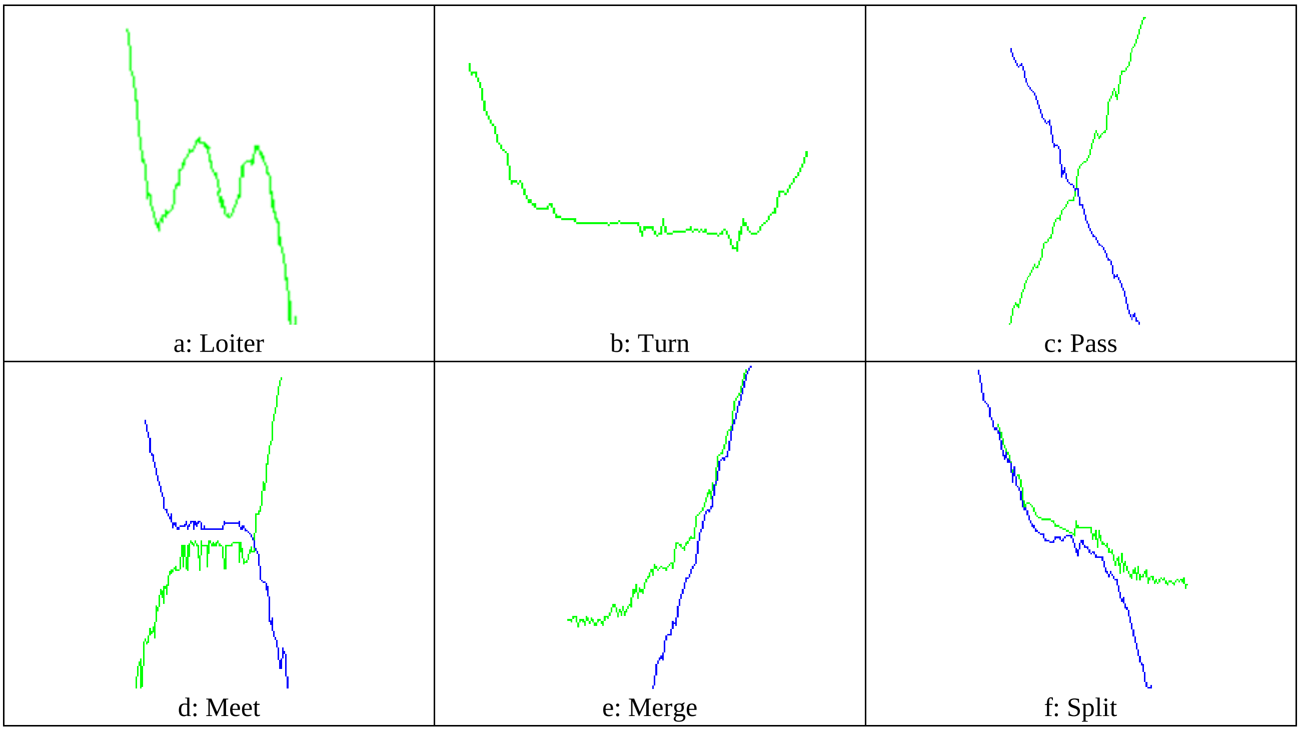

Table 3: Performance of the system on actions.

Event	**Time (min)**	**Tracks**	**Acts**	**TP (in 10)**	**Sensitivity (in 10 acts)**	**Detected event (in tracks)**	**FP (in max 30)**	**FP / max30 (%)**	**Est. FP (in tracks)**	**FP / track (%)**	**FP / min**
Walk	38:23	1831	10	10	100%	823	0	0%	0.0	0.0%	0.0
Run	38:51	1773	10	9	90%	10	0	0%	0.0	0.0%	0.0
Loiter	39:40	1810	10	9	90%	165	1	3%	5.5	0.3%	0.1
Stop	39:15	1785	10	10	100%	269	6	20%	53.8	3.0%	1.4
Turn	39:08	1795	10	7	70%	78	16	53%	41.6	2.3%	1.1

Table 4: Performance of the system on interactions.

Event	**Time (min)**	**Tracks**	**Acts**	**TP (in 5)**	**Sensitivity (in 5 acts)**	**Detected event (in tracks)**	**FP (in max 30)**	**FP / max30 (%)**	**Est. FP (in tracks)**	**FP / track (%)**	**FP / min**
Pass	29:56	1429	5	4	80%	93	0	0%	0.0	0.0%	0.0
Meet	30:59	1468	5	3	60%	9	4	44%	4.0	0,3%	0.1
Merge	28:22	1402	5	4	80%	15	7	47%	7.0	0,5%	0.2
Split	28:44	1433	5	3	60%	27	7	26%	7.0	0.5%	0.2

4. CONCLUSIONS

In this paper, we presented a method for event recognition by detecting large scale track-based patterns. The method consists of several steps: pedestrian detection, track generation, track-based feature computation and rule-based event classification. The experiments included several actions (walk, run, loiter, stop, turn) and interactions (pass, meet, merge, split). The experiments in a crowded shopping mall showed that the actions can be detected reliably (on average 90%) at a low false positive rate (1.1%), and that the interactions have lower detection rates (70% at 0.3% FP). Future work may use this approach as a component for a system that assists operators to find threatening behavior and enrich the selection of videos that are observed. Future work may also combine small-scale motion analysis with large-scale track-based patterns.

ACKNOWLEDGEMENT

The work for this paper was supported by the Netherlands top sector 'High Tech Systems and Materials', roadmap Security, in the project 'Passive sensors'. The authors acknowledge the "Centrum voor Innovatie en Veiligheid" (CIV) and the "Diensten Centrum Beveiliging" (DCB) in Utrecht for providing the fieldlab facilities and support. The company AXIS is acknowledged for generously providing several cameras of the type 'P1346' for the duration of the multi-camera action-recognition experiment.

REFERENCES

[1] An, L., Kafai, M., Yang, S., Bhanu, B., "Reference-based person re-identification," IEEE AVSS, (2013).

[2] Bialkowski, A., Denman, S., Sridharan, S., Fookes, C., Lucey, P., "A database for person re-identification in multi-camera surveillance networks," IEEE DICTA, (2012).

[3] Bouma, H., Baan, J., Borsboom, S., Zon, K., Luo, X., Loke, B., Stoeller, B., Kuilenburg, H., Dijk, J., "WPSS: Watching people security services," Proc. SPIE 8901, (2013).

[4] 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).

[5] Bouma, H., Borsboom, S., Hollander, R., Landsmeer, S., Worring, M., "Re-identification of persons in multi-camera surveillance under varying viewpoints and illumination," Proc. SPIE 8359, (2012).

[6] Bouma, H., Vogels, J., Aarts, A., Kruszynski, C., Wijn, R., Burghouts, G., "Behavioral profiling in CCTV cameras by combining multiple subtle suspicious observations of different surveillance operators," Proc. SPIE 8745, (2013).

[7] Bouma, H., Burghouts, G., Penning, L., et al., "Recognition and localization of relevant human behavior in videos," Proc. SPIE 8711, (2013).

[8] Bouma, H., Baan, J., Burghouts, G., e.a., "Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall," Proc. SPIE 9253, (2014).

[9] Burghouts, G.J., Schutte, K., Bouma, H., Hollander, R., "Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos," Machine Vision and Applications MVA, (2013).

[10] Burghouts, G., Eendebak, P., Bouma, H., Hove, J.M., "Improved action recognition by combining multiple 2D views in the bag-of-words model," IEEE AVSS, 250-255 (2013).

[11] Burghouts, G., Schutte, K., Hove, R., e.a., "Instantaneous threat detection based on a semantic representation of activities, zones and trajectories," Signal Image and Video Processing SIVP, (2014).

[12] Dijk, J., Rieter-Barrell, Y., Rest, J. van, Bouma, H., "Intelligent sensor networks for surveillance," Journal of Police Studies: Technology-Led Policing 3(20), 109-125 (2011).

[13] Fagette, A., Courty, N., Racoceanu, D., Dufour, J.Y., "Unsupervised dense crowd detection by multiscale texture analysis," Pattern Recognition Letters, (2013).

[14] Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M., "Person re-identification by symmetry-driven accumulation of local features," IEEE CVPR, 2360-2367 (2010).

[15] Ferryman, J., Ellis, A., "Performance evaluation of crowd image analysis using the PETS2009 dataset," Pattern Recognition Letters, (2014).

[16] Ferryman, J., Hogg, D., Sochman, e.a., "Robust abandoned object detection integrating wide area visual surveillance and social context," Pattern Recognition Letters 34(7), 789-798 (2013).

[17] Gray, D., Brennan, S., Tao, H., "Evaluating appearance models for recognition, reacquisition, and tracking," IEEE Int. Workshop Performance Evaluation of Tracking and Surveillance PETS, (2007).

[18] Gray, D., Tao, H., "Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features," Proc. European Conference on Computer Vision ECCV, (2008).

[19] Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B., "Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences," IEEE Distributed Smart Cameras, (2008).

[20] Hu, N., Bouma, H., Worring, M., "Tracking individuals in surveillance video of a high-density crowd," Proc. SPIE 8399, (2012).

[21] Marck, J.W., Bouma, H., Baan, J., Oliveira Filho, J. de, Brink, M. van den, "Finding suspects in multiple cameras for improved railway protection," Proc. SPIE, (2014).

[22] Rest, J. van, Grootjen, F., Grootjen, M., Wijn, R., Aarts, O., Roelofs, M., Burghouts, G., Bouma, H., Alic, L., Kraaij, W., "Requirements for multimedia metadata schemes in surveillance applications for security," Multimedia Tools and Applications MTAP, (2013).

[23] Satta, R., "Dissimilarity-based people re-identification and search for intelligent video surveillance," PhD thesis Univ. Cagliari Italy, (2013).

[24] Wang, H., Schmid, C., "LEAR-INRIA submission for the THUMOS workshop," THUMOS Challenge: ICCV Workshop on Action Recognition with Large Number of Classes, (2013).

[25] Weinland, D., Boyer, E., Ronfard, R., "Action recognition from arbitrary views using 3D exemplars," ICCV, (2007).