Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall

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

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

ABSTRACT

Proactive detection of incidents is required to decrease the cost of security incidents. This paper focusses on the automatic early detection of suspicious behavior of pickpockets with track-based features in a crowded shopping mall. Our method consists of several steps: pedestrian tracking, feature computation and pickpocket recognition. This is challenging because the environment is crowded, people move freely through areas which cannot be covered by a single camera, because the actual snatch is a subtle action, and because collaboration is complex social behavior. We carried out an experiment with more than 20 validated pickpocket incidents. We used a top-down approach to translate expert knowledge in features and rules, and a bottom-up approach to learn discriminating patterns with a classifier. The classifier was used to separate the pickpockets from normal passers-by who are shopping in the mall. We performed a cross validation to train and evaluate our system. In this paper, we describe our method, identify the most valuable features, and analyze the results that were obtained in the experiment. We estimate the quality of these features and the performance of automatic detection of (collaborating) pickpockets. The results show that many of the pickpockets can be detected at a low false alarm rate.

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

1. INTRODUCTION

Proactive detection of imminent incidents is required to decrease the cost of security and of security incidents [24]. The number of surveillance cameras is rapidly increasing to improve security in crowded environments, such as airports, shopping malls and railway stations. However, the number of human operators remains limited, only a selection of the video streams can be observed and the effectiveness of predictive behavior indicators is not clear. Automatic early detection of suspicious behavior in CCTV cameras can help to prevent incidents in a timely manner and handle the huge amount of data. This paper focusses on the automatic detection of suspicious behavior of collaborating pickpockets, because this illustrates the predictive value of social behavior indicators.

Threat detection – and pickpocket detection in particular – is a challenging problem for several reasons. The first challenge is that threats appear in many variations. For example, suspects may work alone or they may work in groups and they may loiter slowly or move as fast as others. The second challenge is that the environment is crowded, which may lead to occlusions. The third challenge is that people move freely through the area which cannot be covered by a single camera, which hinders robust long-term behavior analysis. The fourth challenge is that threats are a high-level semantic concept, since the threat consists of a complex social interaction between (multiple) suspects and a victim over a long period. The fifth challenge is that the behavior may be very subtle. In pickpocketing, the actual snatch is hardly visible and the suspect will try to blend in the crowd. Finally, the sixth challenge is that the number of actual threats is extremely low in comparison to the huge amount of normal passers-by. To improve the efficiency of human operators, a system for automatic pickpocket detection shall not create many false alarms.

Our novel contribution is that we present a method that detects pickpockets in a realistic and large collection of normal passers-by in a crowded shopping mall. Pickpockets demonstrate typical kinds of behavior before, during and after the actual incident. Examples of this behavior are: following an intended victim, or – in the case of cooperating pickpockets

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

H. Bouma, J. Baan, G.J. Burghouts, P.T. Eendebak, J.R. van Huis, J. Dijk, J.H.C. van Rest, "Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall," Proc. SPIE, Vol. 9253, (2014). http://dx.doi.org/10.1117/12.2066851

– interacting with each other for coordination or for handing over the loot. Our approach focusses on the walking and interaction patterns with track-based features.

Our method consists of several steps: pedestrian tracking, feature computation and pickpocket recognition. We performed an experiment with 19 actors that performed in total more than 20 validated pickpocket incidents in a crowded shopping mall. We used a top-down approach to translate expert knowledge in features and rules, and a bottom-up approach to learn discriminating patterns with a classifier. The classifier was used to separate the pickpockets from normal passers-by that are shopping in the mall. We performed a cross validation approach to train and evaluate our system. In this paper, we describe our method, identify the most valuable features, and analyze the results that were obtained in the experiment. We estimate the quality of these features, and the performance of automatic detection of (collaborating) pickpockets.

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.

2. METHOD

2.1 System overview

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

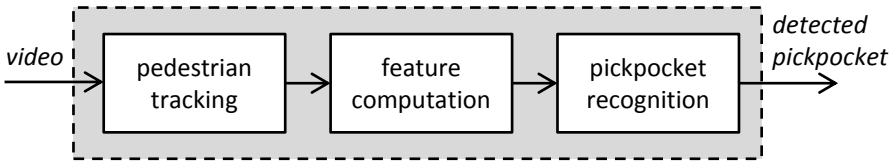


Figure 1: Overview of the method.

2.2 Pedestrian tracking

We have a system that generates detections and tracks in each camera [5][23]. These tracks can be used for multi-camera tracking and re-identification [1][19][6][27] and further behavior analysis [4]. For each person, typically multiple tracks are generated due to presence in different cameras and track fragmentation. The tracks contain location information in world coordinates, and location information in camera coordinates (bounding boxes). In this paper, the obtained world-coordinate tracks are used to support the track-based feature computation.

2.3 Track-based feature computation

The seven main steps of a pick pocket scenario may include the following [25]: Observing the environment, waiting for an opportunity, communicating to accomplice, surrounding the victim, snatch something, handing over the loot to an accomplice, and leaving the scene. Related to this, we defined features related to the walking speed, orientation change, split and merge interactions. These track-based features were designed based on expert knowledge. The features are shown in Table 1.

Table 1: Track-based features designed with expert knowledge.

Feature label	Description
Max. distance to line BE	Maximum distance of a track to the line from begin to end point (indication of straightness of the track).
Max. orientation change	Maximum orientation change of a track
Max. distance after merge	Maximum distance between two tracks after merge
Max. distance after split	Maximum distance between two tracks after split
Max. distance before merge	Maximum distance between two tracks before merge
Max. distance before split	Maximum distance between two tracks before split
Mean speed	Average speed of a track
Min. max. dist. near crossing	Minimum of the max dist. before and after a crossing of two tracks.
Min. distance to camera	Minimum distance to the camera of a track (for removal of tracks that are too far away)
No. persons in 2m	Number of persons that is within 2 meters.
No. persons in 4m and 3s	Number of persons that is within 4 meters for at least 3 sec.
Speed [0-3] km/h	Duration that the speed is less than 3 km/h of a track
Speed [0-1] km/h	Duration that the speed is between 0 and 1 km/h.
Kinematics Histogram	Each track is described by a 2D speed and orientation-change histogram.
Kinematics GMM	Each track is described by a 2D speed and orientation-change space clustered by GMM to create a bag of words (BOW) histogram.

2.4 Pickpocket recognition

Several classifiers were used to recognize the pickpockets. All features and their combinations were classified using a Fisher linear discriminant classifier (LDC) [13], which proved to distinguish the samples best during our experiments. The only exception is the Kinematics-GMM, for which a random forest (RF) combined with a voting scheme worked better. In all cases, the support vector machine (SVM) could not be tuned to get better results.

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 [22]. The other cameras are hardly overlapping and they are used to cover a large region of the shopping mall.

In the experiment, 18 actors were used to become pickpockets and victims. The pickpockets worked several times in a group of 1, 2 or 3 persons (1, 7, 2 times respectively) and they had 60 minutes to rob multiple victims. The victims returned in 20 minutes and, in total, these victims went into the shopping mall 30 times. The victims were instructed not to cause a rumor (even if they noticed the theft) and when passers-by would intervene, they had to step out of their role and calm the passer-by. In total, more than 20 victims (of the 30) were pickpocketed. Only one victim was attended by passers-by, but this happened after the pickpocket already left the scene with the loot. During the experiment, we did not establish a perimeter, so the majority of people are shoppers during both experiments. Therefore, our dataset contains both suspicious and normal behavior.

(A) (B)

(C) (D)

Figure 2: Four examples of pickpocket incidents: (A) Pickpocket P1 steals from victim V before she bumps into P2, (B) P1 steals from V while P2 is distracting, (C) P1 steals from V and gives the loot to P2 who quickly leaves, (D) P1 steals from V while P2 is distracting.

3.2 Annotation and separation of train/test data

The total duration of the experiment was 8 hours. Automatic detection and tracking was applied to the video data [5] resulting in 258.111 tracks. Actors entered the scene multiple times as pickpocket and victim. We identified 19 pickpockets. Each of these pickpockets was annotated by combining the tracks to a ground-truth track for these persons. In total, this resulted in 1443 pickpocket tracks.

To assess the performance of the system, we performed leave-one-pickpocket-out cross validation (19 folds). Each test fold contains one pickpocket person and other tracks of passerby in the same camera where the pickpocket is present are also added to the test set. Assignment of a camera to the test set was done in segments of 5 minutes. Neighboring cameras of which the field-of-view may overlap with the cameras in the test set without an annotation of the test person are excluded from the train and test set. Also other pickpockets that are present in the test fold are ignored. The train set contains time segments of cameras in which the test-pickpocket is not visible. Note that the test folds are not mutually exclusive (since pickpockets work in groups and some are present at the same time in the same camera), but that the train set is correctly separated from the test set.

3.3 Results of pickpocket detection

The results are shown in Figure 3, Figure 4, Table 2 and Figure 5.

Figure 3 and Figure 4 show the ROC curves. Figure 3 shows an ROC curve where the true and false positive rates are expressed in the percentage of tracks. The figure shows that the methods perform much better than random. Of course, the pickpockets try to blend in and they are not constantly showing deviant behavior. Therefore, it would be impossible to perfectly separate all pickpocket tracks from passer-by tracks. Only one detection – or a few detections – for each pickpocket could be sufficient for an operator to increase the attention for a suspect. Figure 4 shows the true positive rate as the percentage of pickpocket persons that is detected. The number of detected tracks that is required to recognize a person can be varied. In this figure, we have chosen a threshold of one detections per pickpocket (left) and five detections per pickpocket (right). So, in the latter case, the 100% true positive rate can only be reached with at least 5 detections for all 19 pickpockets (= 95 detections). A threshold of 1 detection per pickpocket shows higher ROC curves (for the methods and for the random curve) than the threshold of 5 detections per pickpocket. Note that the horizontal axis focusses on the false positive rate until only 6%.

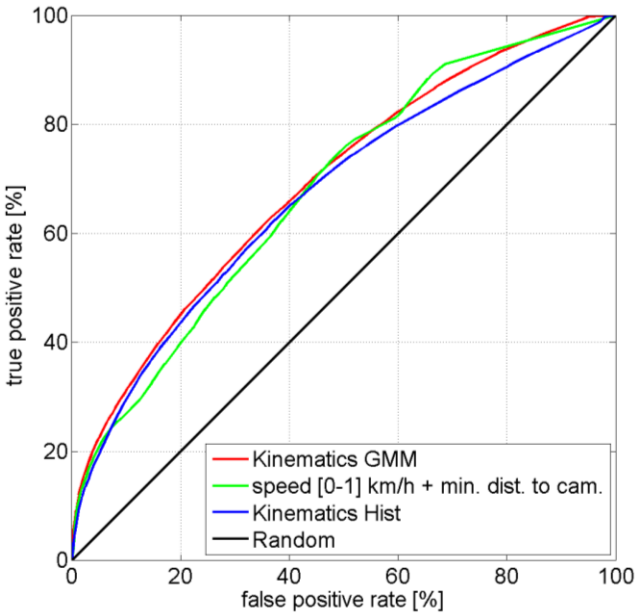


Figure 3: ROC curve with vertically the correctly detected percentage pickpocket tracks.

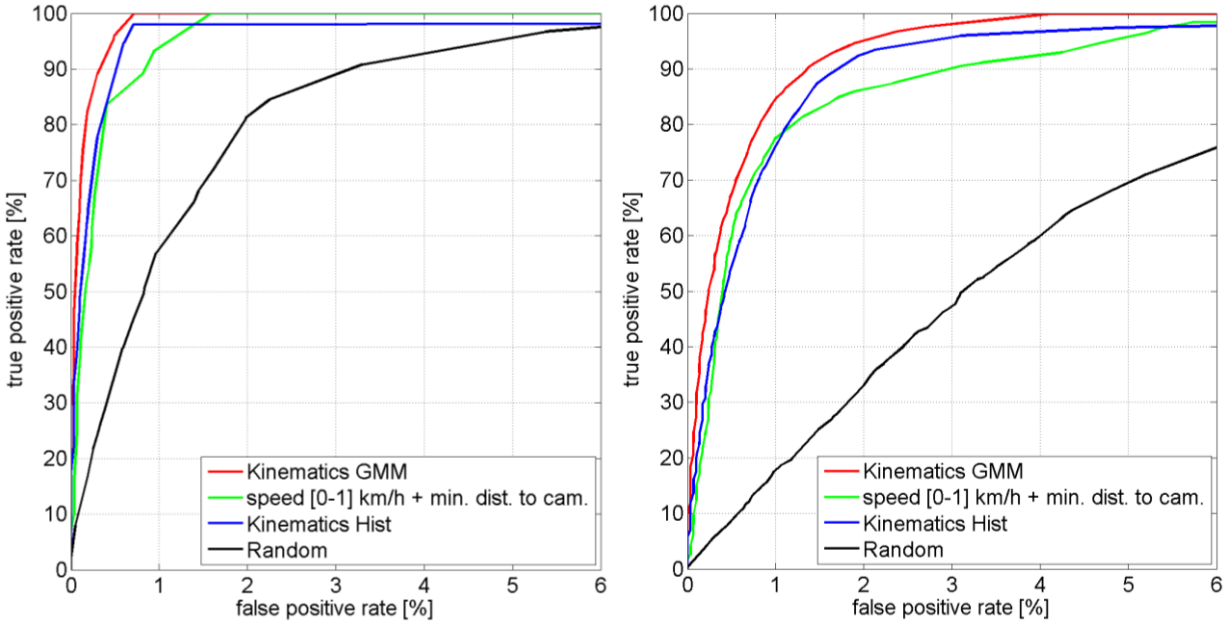


Figure 4: ROC curves with vertically the correctly detected percentage pickpocket persons for at least 1 (left) and 5 (right) detected tracks per pickpocket.

Table 2 shows the performance numbers that are extracted from Figure 3. The value of 'Sens @ 0.5%' refers to the true positive rate at 0.5% false positives and the value of 'AUC_1%' refers to the area under the curve in the range from 0 to 1% false positives. The AUC_1% is a percentage that ranges from 0 to 100%. We have chosen the area measure because its integration results in a more reliable estimate than the sensitivity and we have chosen the range until 1% because we wish to focus on a working point in this range. The table shows that Kinematics-GMM performs better than other features ($AUC_{1\%}$ = 60%). It performs significantly better than the second best feature combination in the table (at p = 0.0001) and much better than random performance. For 1 detection per pickpocket $AUC_{1\%}$ is 88.7% ± 2.4 and the sensitivity at 0.5% false positives is 95.6 ± 3.5 (numbers not included in the table). We focused on five detections per pickpocket, because we assumed that the first detection may not be suspicious enough for the operator to recognize the pickpocket.

Figure 5 shows the probability density functions of the Kinematics-GMM for pickpockets (red) and passer-by (blue). The figure shows a clear shift in confidences, which allows a separation between the two classes.

Table 2: Performance of the system on persons and on tracks (average ± standard deviation after 12 iterations). The Kinematics GMM performs best with an AUC_1% of 60% and a sensitivity of 66 at 0.5% false alarms.

Method		**Sensitivity of pickpocket (5 det. per pickpocket)**		**Sensitivity of tracks**	
Features	**Classifier**	**AUC_1%**	**Sens @ 0.5%**	**AUC_1%**	**Sens @ 0.5%**
Kinematics GMM	Random Forest	**60 ± 2**	**66 ± 3**	**6.5 ± 0.4**	**7.2 ± 0.6**
Speed [0-1] + Min dist. to cam.	Fisher LDC	51 ± 1	60 ± 3	5.0 ± 0.2	5.3 ± 0.3
Speed [0-3] + Min dist. to cam.	Fisher LDC	50 ± 1	57 ± 3	5.0 ± 0.2	5.3 ± 0.5
Speed [0-3] + No. pers. in 4m & 3s	Fisher LDC	49 ± 2	56 ± 4	4.0 ± 0.2	3.9 ± 0.4
Speed [0-3] + Max orient. change	Fisher LDC	48 ± 1	55 ± 3	4.3 ± 0.2	4.3 ± 0.3
Speed [0-3]	Fisher LDC	48 ± 1	55 ± 4	4.3 ± 0.2	4.5 ± 0.3
Kinematics Histogram	Fisher LDC	47 ± 3	50 ± 4	4.1 ± 0.3	4.5 ± 0.4
Speed [0-1]	Fisher LDC	46 ± 1	51 ± 3	4.0 ± 0.2	3.8 ± 0.4
Random	N / A	9 ± 0.3	9 ± 0.3	0.5 ± 0.2	0.5 ± 0.2

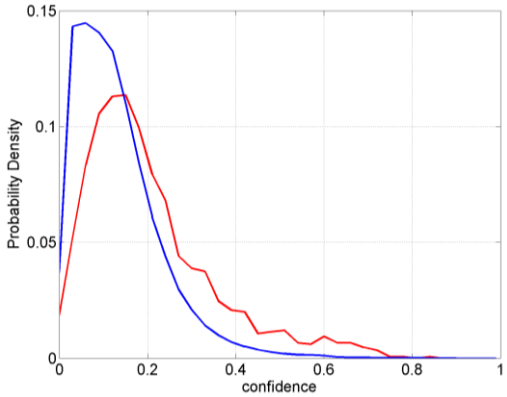


Figure 5: Probability density function of the method Kinematics-GMM, both normalized to an area of 1.0. The blue curve shows confidence histogram of the passer-by tracks and the red shows the histogram of the pickpockets.

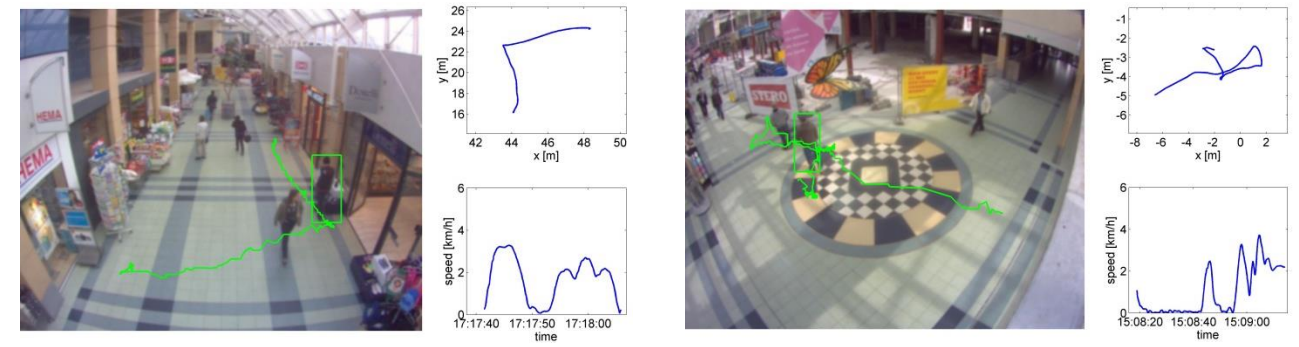
3.4 Analysis and discussion of results

The first point of discussion is related to operational use. We have shown that the system gives a much better performance than a random selection of tracks. However, the question remains whether it is this sufficient for operational use. The video-content analysis (VCA) system will only be used if it improves the performance of an operator, e.g., by making him more effective or more proactive. In order to answer that question, we make an estimate of the number of true detections and false alarms. We use the result in Table 2 indicating a sensitivity (recall) of 66% at 0.5% false positives. This operating point contains 1.29K false alarms (0.5% of 258K tracks) and 104 true positive detections (1443 pickpocket tracks with a true positive rate of 7.2%, which is more than 19 pickpockets with 5 detections per pickpocket and a sensitivity of 66% because some pickpockets received more than 5 detections). On average, this leads to 2.9 alarms per minute (1.39K positives in 8 hours) and 1 out of 13 detections actually is a pickpocket (104 true positives out of 1.39K positives results in a precision of 7.5%). For a system that requires active interaction with the operator for each alert (e.g. an alarm that must be switched off), this performance would be insufficient. In that case, the number of alarms should be much lower. However, when a random selection of cameras is replaced by our system for camera selection with only a bounding-box overlay that indicates the suspect (a weak alert that may be ignored), it may improve the performance of the operator, since the number of pickpockets on the spot view is enriched. The current system can be used if the catching of pickpockets has high priority – or as an improvement for idle operator time – because it only enriches the camera selection for the operator, but it still assumes a human operator that makes decisions. Better estimates of increased performance would require additional experiments with a human operator in the loop.

The main features of our system relate to low speed and large orientation change. This is also confirmed by the examples shown in Figure 6. Typical examples of false positives of the current setup are a loitering group of people and old people that are walking very slowly (with a walker). The best feature is the Kinematic GMM, which does not reduce all information to one or two values, but preserves the complete distribution of values over the complete duration of the track.

4. CONCLUSIONS

In this paper, we presented a method for the challenging problem of automatic detection of pickpockets. Experiments in a crowded shopping mall showed that the system can generate five detections on each pickpocket with a sensitivity of 66% at only 0.5% false alarms. For a single detection on each pickpocket, the sensitivity is even higher (95.6% @ 0.5FP). The main features of the system are based on speed and orientation change and we observed that it is important to preserve the complete distribution of features values from a track. In our experiments, this would lead to less than three false alarms per minute, which could be acceptable to improve camera selection.



A: Pickpocket, GMM=0.71

B: Pickpocket, GMM=0.79

C: Passer-by, GMM=0.6

D: Passer-by, GMM=0.6

Figure 6: Four examples of tracks with a high confidence score. Note that both the pickpockets (A+B) and the passerby (C+D) contain fragments with low speed and/or large orientation change.

Future work may include the improvement of threat recognition by combining track-based analysis with other features (e.g. action recognition for local motion [9][10] or re-identification information for long-term analysis [5][7]) or by extending it to other forms of complex threatening behavior [2][26][27]. Recently, we started a project in collaboration with the Royal Marechaussee and Qubit Visual Intelligence at Amsterdam's Schiphol Airport, where we intend to detect activities such as people falling on the ground and theft of bags amid large crowds. Other work may include the development and harmonization of multimedia metadata schemes which are suited to represent the output of systems like this [25].

ACKNOWLEDGEMENT

The work for this paper was conducted in the project 'Passive sensors' in the Netherlands top sector 'High Tech Systems and Materials'. The creation of the pickpocket dataset was a joint effort with the TNO safety and security research programs Deviant Behavior, funded by the Government of the Netherlands. 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 providing cameras of the type 'P1346' [22].

REFERENCES

[1] An, L., Kafai, M., Yang, S., Bhanu, B., "Reference-based person re-identification," IEEE AVSS, (2013).
[2] Andersson, M., Patino, L., Burghouts, G., et al., "Activity recognition and localization on a truck parking lot," IEEE AVSS, (2013).

[3] Bialkowski, A., Denman, S., Sridharan, S., Fookes, C., Lucey, P., "A database for person re-identification in multi-camera surveillance networks," IEEE DICTA, (2012).
[4] 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).
[5] 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).
[6] 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).
[7] Bouma, H., Vogels, J., Aarts, A., et al., "Behavioral profiling in CCTV cameras by combining multiple subtle suspicious observations of different surveillance operators," Proc. SPIE 8745, (2013).
[8] Bouma, H., Burghouts, G., Penning, L., et al., "Recognition and localization of relevant human behavior in videos," Proc. SPIE 8711, (2013).
[9] 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 SIVP, (2014).
[10] Burghouts, G., Schutte, K., Bouma, H., et al., "Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos," Machine Vision and Applications, (2013).
[11] Burghouts, G., Eendebak, P., Bouma, H., Hove, J. ten, "Improved action recognition by combining multiple 2D views in the bag-of-words model," IEEE AVSS, 250-255 (2013).
[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] Duin, R.P.W., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D. and Tax, D.M.J., "PRTools4, A Matlab Toolbox for Pattern Recognition," http://prtools.org, Delft University of Technology (2004).
[14] Fagette, A., Courty, N., Racoceanu, D., Dufour, J.Y., "Unsupervised dense crowd detection by multiscale texture analysis," Pattern Recognition Letters, (2013).
[15] 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).
[16] Ferryman, J., Ellis, A., "Performance evaluation of crowd image analysis using the PETS2009 dataset," Pattern Recognition Letters, (2014).
[17] 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).
[18] 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).
[19] Gray, D., Tao, H., "Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features," Proc. European Conference on Computer Vision ECCV, (2008).
[20] 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).
[21] Hu, N., Bouma, H., Worring, M., "Tracking individuals in surveillance video of a high-density crowd," Proc. SPIE 8399, (2012).
[22] Huis, J.R., van, Bouma, H., Baan, J., Burghouts, G., e.a., "Track-based event recognition in a realistic crowded environment," Proc. SPIE 9253, (2014).
[23] 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).
[24] Rest, J. van, Nunen, A. van, Roelofs, M., "Afwijkend gedrag," TNO Report, (2014).
[25] Rest, J. van, Grootjen, F., Grootjen, M., et al., "Requirements for multimedia metadata schemes in surveillance applications for security," Multimedia Tools and Applications MTAP, (2013).
[26] TACTICS Consortium, "D3.1 Conceptual Solution Description," www.fp7-tactics.eu, (2013).
[27] Satta, R., "Dissimilarity-based people re-identification and search for intelligent video surveillance," PhD thesis Univ. Cagliari Italy, (2013).
[28] Waern, A., Andersson, M., Petersson, H., "Multi-sensory surveillance for moving vehicles," SPIE Newsroom, (2014).
[29] Wang, H., Schmid, C., "LEAR-INRIA submission for the THUMOS workshop," THUMOS Challenge: ICCV Workshop on Action Recognition with Large Number of Classes, (2013).