Behavioral profiling in CCTV cameras by combining multiple subtle suspicious observations of different surveillance operators

Henri Bouma *, Jack Vogels, Olav Aarts, Chris Kruszynski, Remco Wijn, Gertjan Burghouts

TNO, P.O. Box 96864, 2509 JG The Hague, The Netherlands

ABSTRACT

Camera surveillance and recognition of deviant behavior is important for the prevention of criminal incidents. A single observation of subtle deviant behavior of an individual may sometimes be insufficient to merit a follow-up action. Therefore, we propose a method that can combine multiple weak observations to make a strong indication that an intervention is required. We analyze the effectiveness of combining multiple observations/tags of different operators, the effects of the tagging instruction these operators received (many tags for weak signals or few tags for strong signals), and the performance of using a semi-automatic system for combining the different observations. The results show that the method can be used to increase hits (detecting criminals) whilst reducing false alarms (bothering innocent passers-by).

Keywords: Surveillance, CCTV, behavioral profiling, security, information fusion, threat assessment.

1. INTRODUCTION

CCTV cameras are used for surveillance applications in areas, such as train stations, airports and shopping centers [3]. The complexity of the camera surveillance task and the growing importance of the prevention of incidents may lead to unnecessary bothering of innocent passers-by. When a surveillance operator recognizes subtle deviant behavior for a person it may be insufficient for follow-up actions. However, if multiple weak observations (e.g. from different operators) are combined, it may become a strong indication for a real security threat that requires intervention.

We have conducted an experiment in which we presented videos with and without incidents of theft and pickpocketing to participants. The participants were asked to detect deviant or suspicious behavior and tag these even before an incident occurred. Part of the users were instructed to respond to weak signals and create many tags, the others were instructed to respond only to the strongest signals and tag sparingly.

In this paper, we analyze the effects of combining multiple tags of different operators. That is, we analyzed the ratios of hits and false alarms as a function of number of surveillance operators and number of tags given by those operators. We also analyzed the performance of a semi-automatic system (as alert system or search engine) for combining the different observations and effects of the tagging instruction given to the operators (many tags for weak signals or few tags for strong signals).

The outline of the paper is as follows. The technical system is presented in Section 2. The experimental setup is described in Section 3 and the results are shown in Section 4. Finally, the conclusions and recommendations are presented in Section 5.

2. SYSTEM FOR TAGGING AND MATCHING

The technical system that assists the user consists of two steps: tagging and matching. The ‘tagging’ of a suspect is done when suspicious behavior is recognized in surveillance video. This tag can be generated by technology that automatically

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

Henri Bouma, Jack Vogels, Olav Aarts, Chris Kruszynski, Remco Wijn, Gertjan Burghouts, “Behavioral profiling in CCTV cameras by combining multiple subtle suspicious observations of different surveillance operators”, Proc. SPIE, Vol. 8745, (2013). http://dx.doi.org/10.1117/12.2015869

recognizes deviant behavior or actions [2][4][5][6][7] or it can be generated by humans that manually push a mouse button. In our experiments, we focused on manually generated tags.

The second step of the system is a 'matching', which tries to combine multiple tags from the first step that belong to a single individual. When multiple weak tags are combined, it can become a strong indication for threatening behavior that requires intervention. In order to combine the different tags, a person re-identification method [1] is needed to assist in matching the most similar persons. This matching method can be used in two different ways.

The first way to use the re-identification method is as an automatic alert system. In this case, it only gives an alert if a good match was found and it can be ignored for the rest of the time. If the algorithm decides there is no good match, no similar persons are shown; and if there is a good match, only one similar person is shown and the system selects the best match. Similar to a fire alarm, this setup works best if it runs at the background and only bothers an operator if action is required. However, it less suitable to handle uncertainty in the matching.

The second way to use the re-identification method is as a semi-automatic search engine. In this case, the query is compared to all images and the results are sorted by decreasing similarity. So, the system does not select one image – as the automatic alert system – but presents a sorted list from which the user selects the appropriate result, in the same way as Google sorts millions of websites to present the best matching links on the first pages. However, it is not full automatic and it requires some interaction with the human. The disadvantage of the second way is that it is labor intensive for the human operator, but the advantage is that it leads to better results, especially in difficult situations where it is hard to separate different people, e.g. due to bad lighting or similar clothes.

3. EXPERIMENTAL SETUP

In the experiment, we showed 8 videos to participants, of which 5 videos contained a suspect and an incident – including the moment before an incident – and 3 videos did not contain suspects or (a moment before) incidents. Participants had the task to detect deviant behavior and to separate suspects from innocent people. This could be done by 'tagging' a suspicious person with a mouse click while the video was playing.

The 22 participants were randomly assigned to the instruction to give 'many' or 'few' tags. Participants that received the instruction 'many' were told that it was important to respond early when there was a suspicion, and that it was no problem when an error was made by tagging an innocent person. The other participants were instructed that every false tag should be avoided, although they should tag before an incident occurred to prevent it. The group of participants consisted of students, agents of the Flexteam of the Dutch police in Amsterdam-Amstelland and agents from the Netherlands Royal and Diplomatic Protection Service (DKDB). Because during the automatic assignment of participants all students were assigned to the instruction type 'few', we cannot compare the different groups (Table 1). The student group has therefore been ignored in the analysis of the effects of the instruction type.

Table 1: Participants that received an instruction type 'few' or 'many'.

Participant	Instruction 'many'	Instruction 'few'
Student	-	N = 5
Flex	N = 5	N = 4
DKDB	N = 2	N = 6

After all data was collected, a rectangular bounding box was manually created for each tagged person, and each person received a unique ID-number that was assigned to all related bounding boxes. In total, the participants gave 268 tags, on 36 unique persons (Figure 1). Of the 36 persons, 23 were tagged at least twice and they received 255 tags. The manually assigned ID's were used as ground-truth to verify the quality of the (semi-)automatic re-identification.

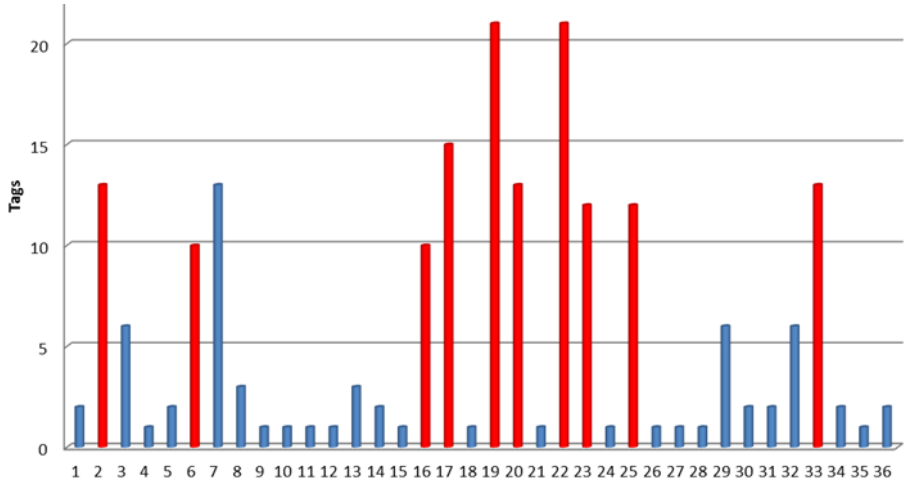

Figure 1: The number of tags each of the 36 tagged persons received, this includes both suspects (red) and innocent people (blue).

4. RESULTS

Errors are introduced by human participants, when either innocent persons are tagged (false positives, FP) or when offenders are missed (false negatives, FN). Additional errors can be introduced by the (semi-)automatic re-identification, when tags pertaining to different persons are falsely combined. In the following three subsections, the results for the three different proposed modes of analyzing and combining the human and electronic data are analyzed separately. In the first subsection (Sec. 4.1), results are shown related to the performance of human participants. In this subsection, we assume that the matching works perfectly and that all tags related to the same person are correctly grouped. In the second and third subsection, the tags of human participants are matched with the proposed method as an alert system (Sec. 4.2), or alternatively as a search engine (Sec. 4.3).

4.1 Results of human participants

In this subsection, we describe the results of the human tagging without assistance of our re-identification method. In these experiments, we assume that the matching works perfectly and that all tags related to the same person are correctly grouped. We focus the discussion on questions about the operators, such as how many tags are needed to reliably recognize a suspect, does the instruction matter, how fast do the operators respond, and what is the effect of the number of operators on the performance?

Figure 1 shows that all suspects (except person 6) receive more tags than their innocent counterparts. Table 2 allows a more detailed analysis of the data in Figure 1. For example, if the threshold of the required number of tags on a single person before being tagged as a suspicious individual, is set to 5, then 12 people are tagged as being positive (TP = 10 and FP = 2). If the threshold is lowered, the number of TP's (or hits) remains the same, but the number FP increases to 6. So, in this case, a high threshold gives better specificity. If we compare the instruction types, we see that the number of the participants with the instruction 'few' have less false positive (FP) scores but also slightly less true positive (TP) scores than the participants with the instruction 'many', as expected. The decision to choose the best instruction type depends on whether it is important not to miss a suspect in high threat cases (sensitivity) or it is important not to bother innocent people (specificity).

The reaction times show that participants with the instruction 'many' respond faster than those with instruction 'few' (10.4 versus 13.8 seconds respectively) and the students appeared to be much slower than the other groups (22 versus 10 and 12 seconds), which could be related to the instruction or to their lack of experience to judge behavior.

Table 2: Number of TP, FP, TN, FN for both instructions separately and combined (without students). The tag threshold relates to the number of tags required on a person to obtain a positive label.

Instruction		**Tag threshold**				
		1	**2**	**3**	**4**	**5**
few + many	TP *(hit)*	10	10	10	10	10
few + many	FP *(false alarm)*	12	6	4	4	2
few + many	TN *(correct reject)*	14	20	22	22	24
few + many	FN *(miss)*	0	0	0	0	0
few	TP	10	9	9	7	2
few	FP	4	2	1	1	0
few	TN	22	24	25	25	26
few	FN	0	1	1	3	8
many	TP	10	10	10	9	8
many	FP	8	3	2	2	1
many	TN	18	23	24	24	25
many	FN	0	0	0	1	2

Although many operators and a high threshold may give a better performance, it will often be impractical and financially infeasible to let many people observe one scene. Therefore, we varied the number of participants from 1 to 10, to analyze the effects of multiple observations. There is a quality difference between the observers and the performance of a random subgroup can therefore be significantly influenced by the choice of specific people from the pool of observers. To correct for this effect, a group of N observers was chosen 1000 times randomly from the total pool of observers and their scores were averaged over all 1000 iterations. The results are shown in Table 3.

Table 3: Percentage of TP/offenders and FP/innocent for various tag thresholds and number of participants assuming perfect matching, without student operators, 10 offenders and 144 innocent people.

%	**Number of participants**	**Tag threshold**									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
TP/(TP+FN)	1	71									
	2	90	51								
	3	97	78	37							
	4	99	91	64	28						
	5	100	96	82	53	22					
	6	100	99	92	72	44	17				
	7	100	100	97	85	62	36	14			
	8	100	100	99	93	77	52	31	12		
	9	100	100	100	97	88	68	45	27	10	
	10	100	100	100	99	95	81	60	38	24	9
FP/(TN+FP)	1	2.4									
	2	4.3	0.5								
	3	5.9	1.1	0.2							
	4	7.2	1.7	0.6	0.1						
	5	8.5	2.3	0.9	0.3	0.1					
	6	9.7	2.9	1.2	0.6	0.2	0.0				
	7	10.6	3.4	1.5	0.8	0.3	0.1	0.0			
	8	11.6	4.0	1.9	1.0	0.6	0.3	0.1	0.0		
	9	12.5	4.4	2.2	1.2	0.7	0.4	0.1	0.0	0.0	
	10	13.4	5.0	2.5	1.5	0.8	0.6	0.3	0.1	0.0	0.0

The table shows that increasing the number of participants leads to a higher sensitivity (TP/offenders) and increasing the tag threshold leads to a lower FP-rate (FP/innocent). For example, one operator can detect 71% of the offenders at a FP-rate of 2.4. If a situation requires that 90% of the offenders is detected, at least two participants are needed with a threshold of 1 observation. This same setting leads to a bothering of 4.3% innocent people, which is moderately acceptable. To decrease the number of false detections, we could use four participants and a detection threshold of 2, which leads to approximately the same percentage of detected offenders (91%) and a lower percentage of false detections (1.7%).

4.2 Results of method as alert system

In this subsection, we describe the results of human participants with our method [1] as alert system, where the algorithm automatically decides whether there is a good match. If the algorithm decides that there is no good match, no similar persons are shown; and if there is a good match, only one similar person is shown and the system selects the best match. We will focus on how well the correct person can be retrieved in a database, and how this affects the percentage of detected offenders and innocent people for various tag thresholds and number of participants. The alert system allows an analysis of the worst-case scenario, since in an operational system the matching could be improved by tracking [9] or spatio-temporal information [3].

To analyze the performance of the alert system, we performed an experiment, in which we used 1000 iterations for each of the 36 persons. Each iteration, this person is selected as query and randomly 35 persons from the overall database are selected to fill a test-database. The test-database does not necessarily include the query person (to make true negatives possible), but if the same person is in the database, it is based on a different tag. After selection, the similarity score is computed between the query image and the images in the database. If the score is higher than a similarity threshold, the best match is returned (positive), and if it is lower then no similar person is returned (negative). If the returned person ID is correct, it is a TP and otherwise a FP. If nothing was returned while the query person ID is present in the database, it is a FN.

The results of the simulation are presented in a Receiver Operator Characteristic (ROC) curve in Figure 2 and the separate TP, FP, TN and FN values are shown in Figure 3. This last figure shows that best (TP+TN) results are obtained at a similarity threshold of approximately 0.45, which corresponds with a sensitivity of 0.61 and a specificity of 0.92.

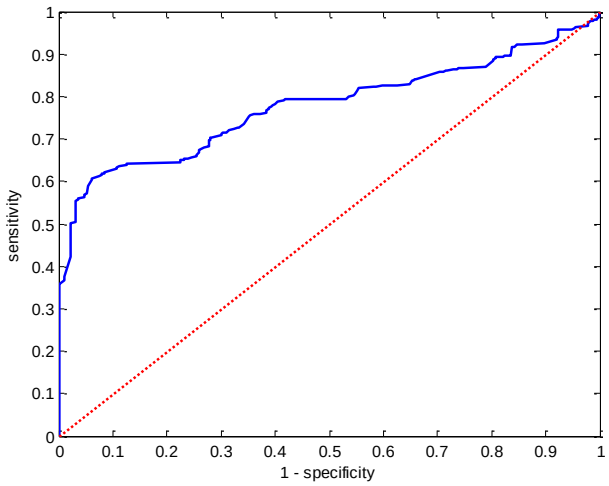

Figure 2: ROC curve of the alert system, which shows the sensitivity TP/(TP+FN) and specificity TN/(TN+FP) for various similarity threshold values.

Similar to the experiment in the previous subsection, we vary the number of participants and the tag threshold. In the previous subsection errors due to matching were ignored, but our alert system with a similarity threshold of 0.45 will result in elevated FP and FN scores. In a separate further experiment, a tagged query is only compared to other tags in the same video (so not against the full database). The results are shown in Table 4. The table shows that both the percentage of detected offenders and the detected innocent people are lower than in the ideal situation. Lowering this

threshold will lead to higher TP and FP detection rates. In the ideal case without matching errors, four participants and a tag threshold of 2 leaded to the 91% detected offenders and 1.7% false detections. With the automatic matching these numbers become slightly worse: 88% and 2.8% respectively. One operator has a sensitivity of 71% and a FP-rate of 2.4. Better results can be obtained by 3 participants and a tag threshold of 2, resulting in 75% at 1.8 FP/innocent.

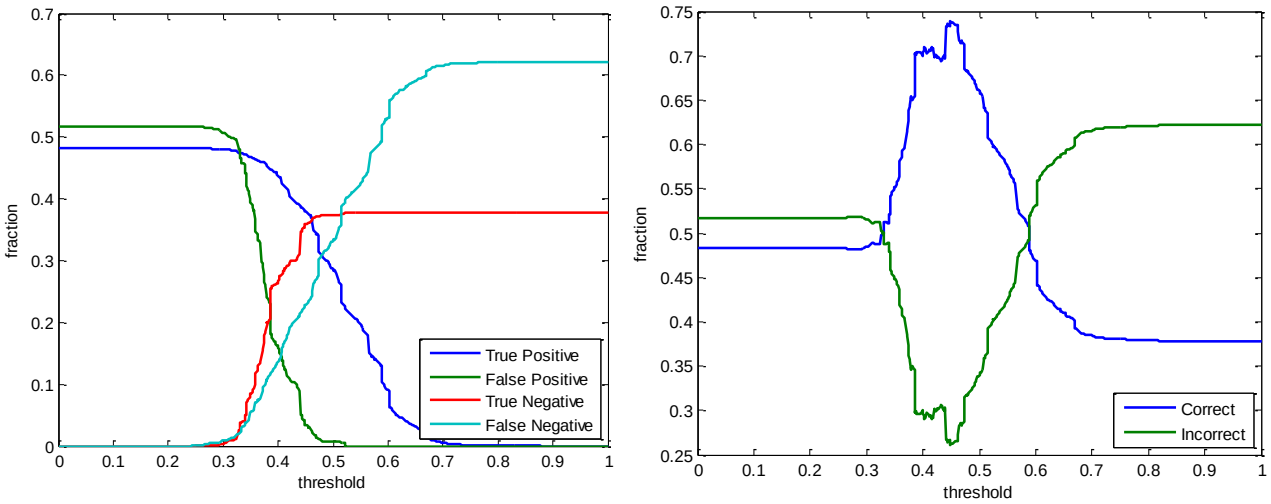

Figure 3: TP, FP, TN, FN, and the correct (TP+TN) and incorrect (FP+FN) for the alert system and different similarity thresholds.

Table 4: Percentage of TP/offenders and FP/innocent for various tag thresholds and number of participants when using the method as an alert system, without student operators, with 10 offenders and 144 innocent people.

%	Number of participants	Tag threshold									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
TP/(TP+FN)	1	71									
	2	91	53								
	3	97	75	39							
	4	99	88	66	32						
	5	100	93	79	57	25					
	6	100	97	87	72	49	20				
	7	100	99	91	81	64	41	16			
	8	100	99	94	86	76	59	36	13		
	9	100	100	97	89	81	70	51	31	10	
	10	100	100	98	92	85	78	65	46	28	9
FP/(TN+FP)	1	2.4									
	2	4.3	0.8								
	3	5.8	1.8	0.4							
	4	7.3	2.8	1.0	0.2						
	5	8.6	3.7	1.7	0.6	0.1					
	6	9.7	4.4	2.4	1.1	0.4	0.1				
	7	10.6	4.9	3.0	1.6	0.7	0.2	0.0			
	8	11.6	5.6	3.4	2.1	1.1	0.4	0.1	0.0		
	9	12.4	6.2	3.9	2.6	1.5	0.7	0.3	0.1	0.0	
	10	13.4	6.9	4.3	3.1	2.0	1.0	0.4	0.2	0.1	0.0

4.3 Results of method as search engine

There are several ways to improve the results of the alert-system, e.g. by using a tracking system [9] or spatio-temporal information [3] or by using a human-computer hybrid. In this subsection, we describe the results of the human-computer hybrid, where the query is compared to all images and the results are sorted by decreasing similarity. So, the system does not select one image – as the alert system – but presents a sorted list in which the human user selects the good result. We focus on the sorting performance of the search engine on the whole database, on all tags within each video and on pairs of tags in the videos.

Our re-identification algorithm [1] is used to order similar images by decreasing similarity and the image with the highest score is placed at rank 1, the second at rank 2, etc. In the ideal case, the image of the same person is always placed at rank 1, but unfortunately this is not always the case. A cumulative matching characteristic (CMC) curve is commonly used to analyze the performance, and it shows which fraction of correct matches has at most a certain rank (e.g. a value of 95% at rank 10 means that 95% is correctly ranked in the first 10 positions).

To analyze the performance of our method, all 23 persons are used that occur at least twice in the dataset (255 tags). Every image is compared with the other 254 images and the best correct match is selected. The results in Figure 4 show that 95% of the correct matches is located at rank 1.

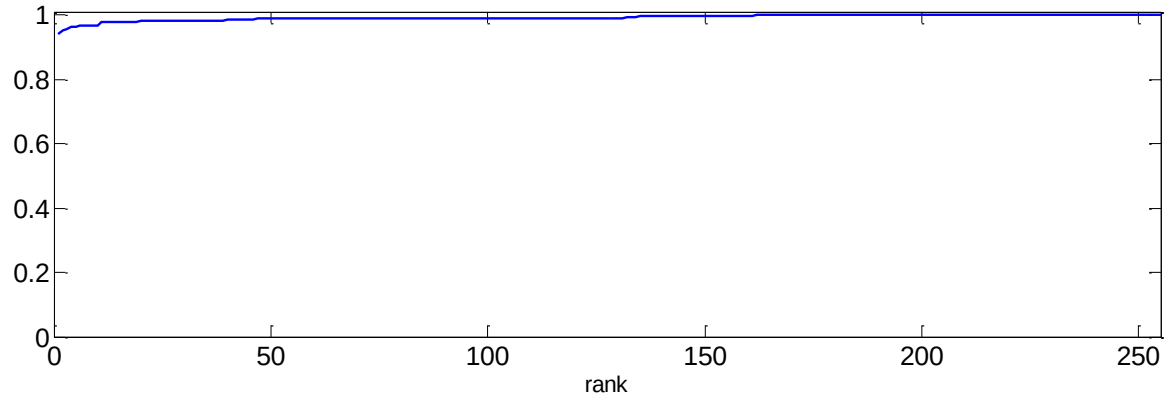

Figure 4: CMC curve of best matches over 23 persons and 255 tags in all videos.

The videos are very different in lighting, color and background. Because each person is only present in one video, the environment helps the re-identification algorithm to select the correct person. Therefore, we also analyze the performance of the algorithm to distinguish people in the same video. The results are shown in Figure 5. For more than 90% of the queries a correct person is selected at rank 1 in 7 out of 8 videos, only video 3 performs worse. This is probably caused by the larger number of people in this video (6 tagged persons, instead of 2 or 3 in other videos).

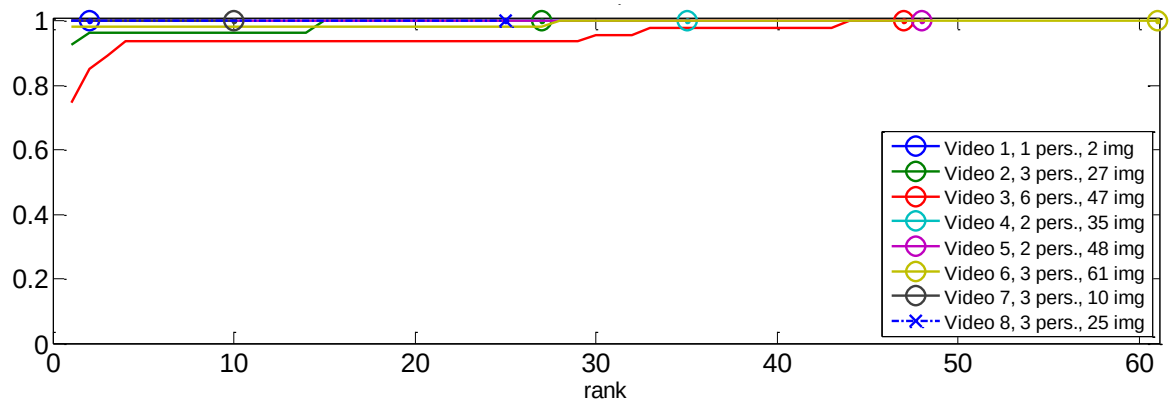

Figure 5: CMC curve of the best matches of the same person per video.

Selecting the best score of persons that occur more than two times may lead to an overestimation of the average system performance. To obtain more accurate estimates we select a database with 23 unique persons and 23 unique queries (Figure 6). At rank 1, the sensitivity is 76%, which would probably be insufficient for an automatic alert system. There are two ways how the human-computer hybrid can lead to improved results. One approach is to present only the first 5 results. This would hardly take time of the user and increase the sensitivity to approximately 90%. Another approach lets the user search the complete database until the correct match is found. If the database is randomly sorted, this would be very time consuming. However, the average CMC is over 90%, which allows a human (assisted by the search engine that sorts the database) to find a hit in the database more than 5 times faster than without the search engine.

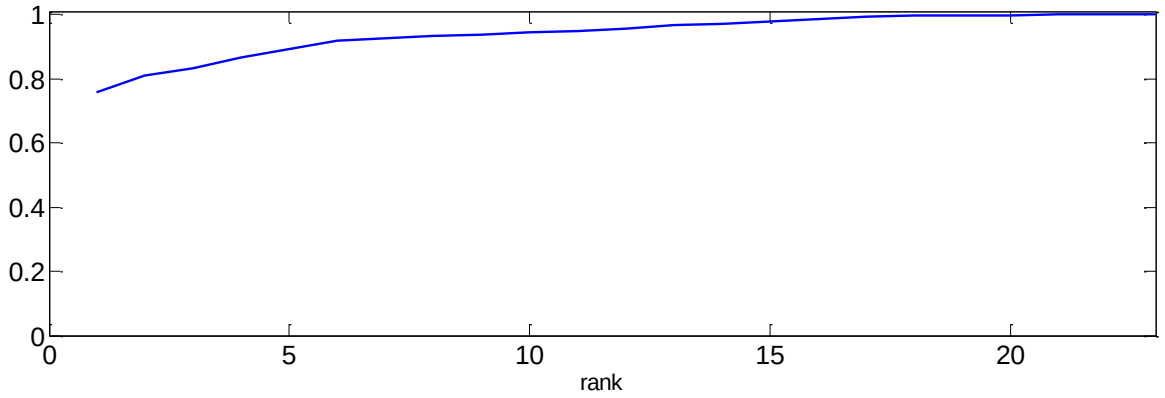

Figure 6: CMC curve of 23 query-target pairs.

5. CONCLUSIONS AND RECOMMENDATIONS

The overall conclusion of this preliminary study is that combining multiple weak observations may lead to a strong indication for threatening behavior that needs intervention. In this paper, we analyzed the effects of combining multiple tags of different operators. That is, we analyzed optimal numbers of tags that give good indications of persons with malicious behavior and the accompanying chances of targeting innocent passers-by, and we analyzed the number of operators needed to reach a predefined performance in our setup. To match multiple tags we used a re-identification system for combining the different observations as an alert system or as a search engine. Finally, we analyzed the effects of the tagging instruction for these operators (many tags for weak signals or few tags for strong signals).

The current results indicate that our approach to combining multiple tags leads to increased effectiveness both on the side of increasing hits and decreasing false alarms. Most notably, this analysis can make comprehensive how shifting a criterion toward preventing incidents may automatically lead to increased false alarms and how this latter effect can be reduced by including more operators. The decision to choose the best instruction type depends on whether it is important not to miss a suspect in high threat cases or it is important not to bother innocent people. Furthermore, with the current state-of-the-art, it is better to use the re-identification method as an interactive search engine than as automatic alert system.

The limited size of the dataset in this initial study prohibits drawing firm conclusions. The number of persons in the database is low, and each suspect is only present in one video while each video has a different environment. Therefore, the performance of the re-identification algorithm is probably over-estimated. Despite these factors this study clearly demonstrates the power of combining different weak tags to a single strong one. This method is especially useful when the observations of several operators (who are watching different cameras) is supported by an intelligent data information network. To analyze the performance of the automatically combined tags, a future experiment will include several operators that watch different cameras in which the same suspect can be observed. Reliable statistical analysis was not conclusive because the students were not equally distributed over the instructions, the numbers were too low and the variances were not equal in the subgroups. For good analysis of the instruction type, assignment to this type to must be more balanced and other variations (e.g. operator experience) should be limited.

ACKNOWLEDGEMENT

The authors thank the students, and the agents of the Flexteam and DKDB for their participation in the user experiments. The work for this paper was supported by the Ministry of Security and Justice in the project for 'Early recognition of deviant behavior' (Topic 1) and by the Dutch top sector 'High Tech Systems and Materials', roadmap Security, in the project 'Passive sensors'.

REFERENCES

[1] Bouma, H., Borsboom, S., Hollander, R. den, Landsmeer, S., Worring, M., "Re-identification of persons in multi-camera surveillance under varying viewpoints and illumination," Proc. SPIE 8359, (2012).
[2] Bouma, H., Hanckmann, P., Marck, J.W., Penning, L., Hollander, R., Hove, J.M. ten, Broek, S.P. van den, Schutte, K., Burghouts, G., "Automatic human action recognition in a scene from visual inputs," Proc. SPIE 8388, (2012).
[3] Bouma, H., Baan, J., Landsmeer, S., Kruszynski, C., Antwerpen, G. van, Dijk, J., "Real-time tracking and fast retrieval of persons in multiple surveillance cameras of a shopping mall," Proc. SPIE 8756, (2013).
[4] Bouma, H., Burghouts, G., Penning, L. de, Hanckmann, P., Hove, J.M., Korzec, S., Kruithof, M., Landsmeer, S., Leeuwen, C. van, Broek, S. van den, Halma, A., Hollander, R. den, Schutte, K., "Recognition and localization of relevant human behavior in videos," Proc. SPIE 8711, (2013).
[5] Burghouts, G.J., Penning, L., Hove, J.M., Landsmeer, S., Broek, S.P., Hollander, R. den, Hanckmann, P., Kruithof, M., Leeuwen, C., Korzec, S., Bouma, H., Schutte, K., "A search engine for retrieval and inspection of events with 48 human actions in realistic videos," Int. Conf. Pattern Recognition Applications and Methods ICPRAM, (2013).
[6] Burghouts, G., Schutte, K., "Spatio-temporal layout of human actions for improved bag-of-words action detection," Pattern Recognition Letters, (2013).
[7] Burghouts, G., Hollander, R. den, Schutte, K., Marck, J., Landsmeer, S., Breejen, E. den, "Increasing the security at vital infrastructures: automated detection of deviant behaviors," Proc. SPIE 8019, (2011).
[8] 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).
[9] Hu, N., Bouma, H., Worring, M., "Tracking individuals in surveillance video of a high-density crowd," Proc. SPIE 8399, (2012).
[10] Wijn, R., Van den Berg, H., Lousberg, M., "On operator effectiveness: The role of expertise and familiarity of environment on the detection of deviant behavior," Personal and Ubiquitous Computing, (2011).