Fusion of optical flow based motion pattern analysis and silhouette classification for person tracking and detection

Johan W.H. Tangelder[*a], Ed Lebert[a], Gertjan J. Burghouts[b], Kasper van Zon[a], Marten J. den Uyl[a]
[a]VicarVision, Singel 160, 1015 AH Amsterdam, The Netherlands.
[b]TNO, Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands

ABSTRACT

This paper presents a novel approach to detect persons in video by combining optical flow based motion analysis and silhouette based recognition. A new fast optical flow computation method is described, and its application in a motion based analysis framework unifying human tracking and detection is outlined. Our optical flow algorithm represents optical flow by grid based motion vectors, which are computed very efficiently and robustly applying template matching. We model the motion patterns of the tracked human and non-human objects by the positions, velocities, motion magnitudes, and motion directions of their optical flow vectors, and build a random forest on these features. For recognition, the random forest computes a normalized score measuring the similarity of a track to a human track. Using edge detection on a motion image for each motion blob its silhouette is computed. Recognition scores are computed, which measure the similarity of the silhouettes with human silhouettes. The optical flow classifier and the silhouette classifier are used as a combined classifier. We analyze the ROC curve to set different decision thresholds on the recognition score for different scenarios. The experiments on the VIRAT test set demonstrate that for human detection the combination of the optical flow based motion method with one based on human silhouette analysis, obtains superior results, compared to the constituent methods.
Keywords: Optical flow based motion analysis, silhouette based recognition, behavior recognition, similarity measures, combined classifier, VIRAT dataset, ROC curve analysis.

1. INTRODUCTION

This paper presents a novel approach to detect persons in video by combining optical flow based motion analysis and silhouette based recognition. A new fast optical flow computation method is described, and its application in a motion based analysis framework unifying human tracking and detection. By combining optical flow and silhouettes, and by applying track based classification instead of frame based classification, recognition results are improved significantly. Optical flow is represented by grid based motion vectors, which are computed very efficiently and robustly applying template matching. A tracking framework using Kalman filtering has been implemented efficiently by predicting the future location of an object using the estimated optical flow vectors. Since motion and shape are complimentary cues for object recognition, their combination into one classifier should boost recognition results.

For optical flow based motion analysis we model the motion patterns of the tracked human and non-human objects by the positions, velocities, motion magnitudes, and motion directions of their optical flow vectors, and build a random forest on these features. We apply the random forest to measure the similarity of a track to a human track.

Using edge detection on a motion image for each motion blob its silhouette is computed. Similarity scores are computed, which measure the similarity of each silhouette with a human silhouette. Also, a normalized recognition score measuring the silhouette-based similarity of a track to a human track is obtained by accumulating the similarity measures of the silhouettes in the track.

Finally, the optical flow classifier and the silhouette classifier are used as a combined classifier. We analyze the ROC curve to set different decision thresholds on the recognition score for different scenarios. We divided the videos from the parking lot scenes from the VIRAT Ground Dataset[1] into a learn set and a test set. The experiments on the VIRAT test set demonstrate that for human detection the combination of the optical flow based motion method with one based on human silhouette analysis, obtains superior results, compared to the constituent methods.

The outline of the paper is as follows. Related work is described in section 2. We describe our approach in section 3. Experimental results are presented in section 4. Finally section 5 concludes our paper.

[*] h.tangelder@vicarvision.nl; phone +31 20 5300333; http://www.vicarvision.nl

Optics and Photonics for Counterterrorism, Crime Fighting, and Defence X; and Optical Materials and Biomaterials in Security and Defence Systems Technology XI, edited by Douglas Burgess, Roberto Zamboni, et al., Proc. of SPIE Vol. 9253, 92530L · © 2014 SPIE · CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2066460

2. RELATED WORK

For an extensive review on techniques for detecting humans in surveillance videos, we refer the reader to the recent survey paper by Paul *et al*.[2]. Three main categories of object recognition methods can be distinguished. **Shape-based methods** detect humans by analyzing the outline of detected objects. Wang *et al*.[3] applied shape Fourier descriptors extracted from the silhouettes during articulated motion to detect human motion sequences. Lin and Davis[4] applied a shape-based, hierarchical part-template-matching approach for human detection including Histogram of Oriented Gradients (HOG) descriptors and scene-to-camera calibration. **Motion-based methods** are based on the idea that object motion characteristics and patterns are unique enough to distinguish humans from other moving objects. Bobick and Davis[5] developed a view-based approach for the recognition of human movements by constructing a vector image template comprising two temporal projection operators: binary motion-energy image and motion-history image. Cutler and Davis[6] presented a self-similarity-based time-frequency technology to detect and analyze periodic motion for human classification. Unfortunately, methods based on periodicity are restricted to periodic motion. **Appearance-based methods** are based on modelling appearance. Dalal and Triggs[7] introduced the HOG descriptor which describes local appearance and shape by the distribution of intensity gradients. By computing the HOG features over a number of single detection windows and classifying the obtained features using a SVM approach humans are detected. Dollar *et al*.[8] apply an image pyramid to obtain a fast pedestrian detector. Given HOG features computed at a certain scale, they reduce computing time significantly by approximating HOG features at nearby scales by reweighting these HOG features. In contrast with Dalal and Triggs[7] single object detection approach Felzenswalb and Huttenlocher[9] have taken a parts-based approach, which deals with the great variability in appearance due to body articulation. In such an approach, each part is detected separately and a human is detected if some or all of its parts are detected in a geometrically plausible configuration. In their pictorial structure approach Felzenswalb and Huttenlocher[9] apply pictorial structures to describe an object by its parts, connected with springs, and represent each part with Gaussian derivative filters of different scale and orientation. Zhu *et al*.[10] applied HOG descriptors, that vary in size, location and aspect ratio filtered by a cascade of detectors. In order to find the blocks best suited for human detection, they applied the AdaBoost algorithm to select those blocks to be included in the detection cascade.

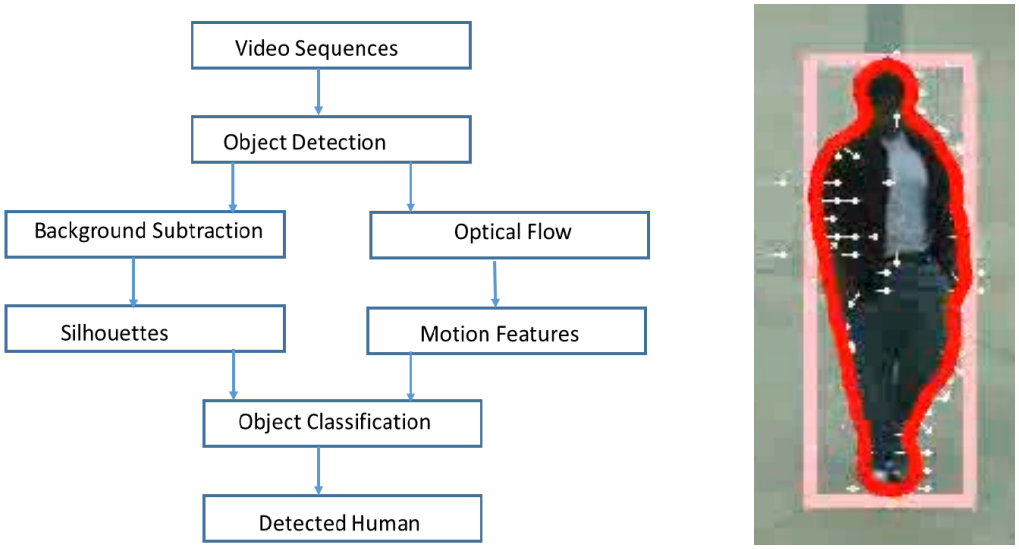
Shape-based methods do not take into account the human motion pattern as a cue for recognition. On the other hand motion-based methods ignore the human silhouette as a cue for recognition. Appearance-based methods do not take motion into account, although they may be applied in combination with motion detection to limit the search area.

Since, we assume that people are moving we focus on analyzing motion sequences, using shape-based silhouettes extracted from motion and motion-based optical flow as complimentary cues.

3. OUR APPROACH

In our approach the human detection process is modelled by two steps: object detection followed by object classification. Two common methods for object detection are background subtraction and optical flow. In this paper we apply the adaptive Gaussian mixture method implemented in OpenCV[11] for background subtraction and our own optical flow implementation, which is described in section 3.1. For object recognition we have developed a shape-based recognition method based on comparing silhouettes, found by background subtraction, and a motion-based recognition method based on extracting motion features from optical flow. Instead of applying only the silhouette classifier or the optical flow classifier, we use both as a combined classifier.

Optical flow feature analysis is combined with analysis of silhouettes derived from motion detection, as illustrated in figure 1. Both optical flow and background subtraction are applied to detect objects from a video sequence. For optical flow detection and tracking we apply a Kalman tracker, which computes for each video a number of tracks, see figure 2. Each track contains for a consecutive number of frames a detection window. By intersecting the detection windows of a track with the bounding boxes of the silhouettes detected by background subtraction, these silhouettes are associated with the track. Finally, based on analyzing the silhouettes and the optical flow the detected object is classified as human or non-human.



(a) (b)

Figure 1: Human detection pipeline: (a) The human detection system combines optical flow and silhouette analysis for human detection; (b) Silhouettes (red) analysis and motion (white) are complimentary cues for human detection.

Figure 2: Four objects tracked by the Kalman filter based on optical flow. The red detection windows contain optical flow vectors and a track identifier.

3.1 Optical flow computation

There have been two directions in the development of optical flow algorithms[12]. One has emphasized higher accuracy; the other faster implementation. We decided to apply a fast optical flow algorithm, which computes only a rather sparse optical flow. Inspired by the work of Ancona and Poggio[13] we applied 1-D kernel templates for estimating the optical flow. The optical flow in one point is calculated by the vector summation of the ***x*** and ***y*** components, which are estimated by 1-D templates. Using this approach is reducing the complexity of the problem, but still have a good estimation of the flow field. The problem is transformed to two one-dimensional searches and the search space size is reduced from quadratic to linear. Our method estimates motion vectors by using fixed grid points. This approach has been implemented on an AXIS camera with an ARTPEC-4 processor. Moreover, the optical flow algorithm has been

implemented on a GPU system. The GPU allows more processing power and with our efficient implementation we could estimate the optical flow vectors using two-dimensional search with a speed of 45 fps on a 1280x720 video[14].

3.2 Computing track scores

Given an object track, a recognition score for the object has to be derived from the recognition scores for the object detections in each detection window. Therefore the object classifier scores are fed into an accumulator network, which integrates recognition over time. We used two accumulators for each detected object, one computing a score S_{person} and one computing a score S_{other}. Finally, we take as recognition score for the track $S_{person} - S_{other}$.
These memory accumulators are applied, because missing values, noisy measurements, and fade in and fade out of partially visual objects occur.

3.3 Motion features

From the flow vectors in the detection windows W the following 7 motion features are derived:

- The detection window ratio $R = W_{width} / W_{heigth}$.
- (X, Y) with $|X| \leq 1$, $|Y| \leq 1$ denoting the location of a flow vector in W relatively to the center of W.
- The flow vector (VX, VY).
- The flow magnitude $Mag = VX^2 + VY^2$.
- The flow angle $Phi = Atan2\ (VX, VY)$.

For motion classification, on the training data for VIRAT pedestrian area scene '0102' and the parking lot scene '0401', a random forest was trained consisting of 2 trees with 32 leafs each using the motion features of the optical flow features.

During testing, for each detection window we propagate each flow vector down both trees, and use the average score as the final probability of this detection. All detections probabilities in a track are accumulated in a final human detection score, as described in section 3.2.

3.4 Silhouette analysis

By applying edge detection on the motion image computed by background subtraction, silhouettes of the moving object are extracted. For each detection window a recognition score for the silhouette is computed based on the output of a neural network recognizer on the silhouette outline and a neural network recognizer on the silhouette skeleton, which is computed by iteratively applying the morphological operations erosion and dilation on an input silhouette, see figure 3.

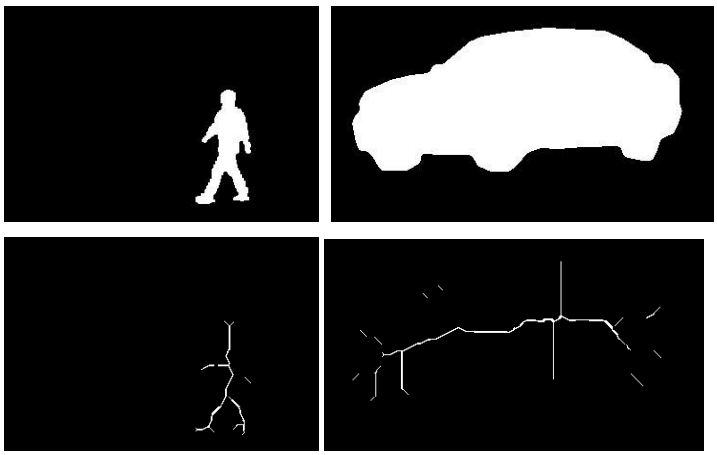


Figure 3: Skeleton of a silhouette person (left) and car (right).

3.4.1 Silhouette outline neural network classification

From a silhouette outline a silhouette outline feature vector of length 64 is derived as follows.

From the edges of the silhouette outline $(x_1, y_1, x_2, y_2, \ldots x_N, y_N)$ we compute the centroid of the silhouette (c_x, c_y), with $c_x = 1/N \sum_{i=1..N} x_i$, $c_y = 1/N \sum_{i=1..N} y_i$. We transform each edge of the silhouette outline (x, y) to polar coordinates according to

Polar angle $\phi = atan2(c_y - y, x - c_x)$

Polar distance $d = \sqrt{(c_y - y)^2 + (c_x - x)^2}$

and normalize the polar coordinates by dividing each polar distance by the maximal polar distance to obtain a scale invariant representation. We bin the angle ϕ in 64 categories and compute for each bin the mean polar coordinate of the silhouette segment intersecting the bin.

For training negative and positive silhouette outline feature vectors are clustered separately, into 12 positive and negative clusters. Finally, the 24 Euclidean distance vectors are fed into a back propagation neural network[15].

3.4.2. Silhouette skeleton neural network classification

Similar to the silhouette outline feature vectors silhouette skeleton features are computed. We bin the normalized silhouette skeleton y coordinate in 64 categories, and compute for each bin the mean normalized x coordinate of the skeleton segment intersecting the bin.

For training negative and positive silhouette outline feature vectors are clustered separately, into 12 positive and negative clusters. Finally, the 24 Euclidean distance vectors are fed into a back propagation neural network[15].

3.4.3 Neural network training

Both neural networks have been trained with 1102 examples of humans from the CASIA Gait Database[16] with 1102 examples and 460 examples of cars and parts of cars from the mpeg7shapeB[17] database and the data sets from TU Graz[18] and TU Darmstad[19]. All detections scores in a track are accumulated in a final human detection score, as described in section 3.2.

3.5 Combined motion and silhouette classification

The final human recognition score for a track is obtained by taking the mean of the accumulated motion recognition score and the accumulated silhouette recognition score.

4. EXPERIMENTAL RESULTS

The VIRAT Ground Dataset[1] is a realistic, natural, challenging video data set for video surveillance applications, and therefore our choice for the experimental verification of our method. From this dataset the VIRAT videos from the three parking lot scenes contain both a lot of human and non-human annotated examples. Moreover, a parking lot is a scene in which one would like to distinguish humans from non-humans (mostly cars). Therefore, we decided to experiment with these videos only and divided them into a learn set containing 19 videos from VIRAT scene ‘0002’, 23 videos from VIRAT scene ‘0101’, and 9 videos from VIRAT scene ‘0401’ and a test set containing 19 videos from VIRAT scene ‘0002’, 23 videos from VIRAT scene ‘0101’, and 8 videos from VIRAT scene ‘0401’. The videos from the parking lot test set in the VIRAT Ground Dataset contain 438 objects, which are annotated as human (person or bike), and 1938 objects as non-human (mainly car).

In our experiments it turned out that the tracks found by our Kalman filter differ significantly from the tracks annotated in the VIRAT Ground Dataset due to the following reasons. Firstly, not all annotated tracks have been detected by the Kalman tracker, because the VIRAT Ground Dataset contains track annotations of non-moving objects and very small objects, which are not detected by the Kalman tracker. Secondly, not all tracks detected by the Kalman tracker, have

been annotated in the VIRAT Ground Dataset, because they were not of interest for human behavior detection. Therefore, we could not evaluate the recognition performance straightforwardly. To alleviate this problem we decided to use the VIRAT tracks as ground truth. To each VIRAT track we add an accumulator storing the detection results. We match a detection window $\boldsymbol{W}$ with an annotated VIRAT track, if the track c ontains a rectangle $\boldsymbol{R}$, such that their Jaccard index $\boldsymbol{J} = |\boldsymbol{W} \cap \boldsymbol{R}| / |\boldsymbol{W} \cup \boldsymbol{R}| \geq 0.5$, and feed the detection scores of the silhouettes and the optical flow in the VIRAT accumulators. The analysis of the results in this section is based on using the annotated VIRAT tracks as ground truth, and the scores of the accumulators of the VIRAT tracks as test results. Figure 4 illustrates this approach. In this way we obtained 60 matched human VIRAT tracks and 144 matched non-human tracks.

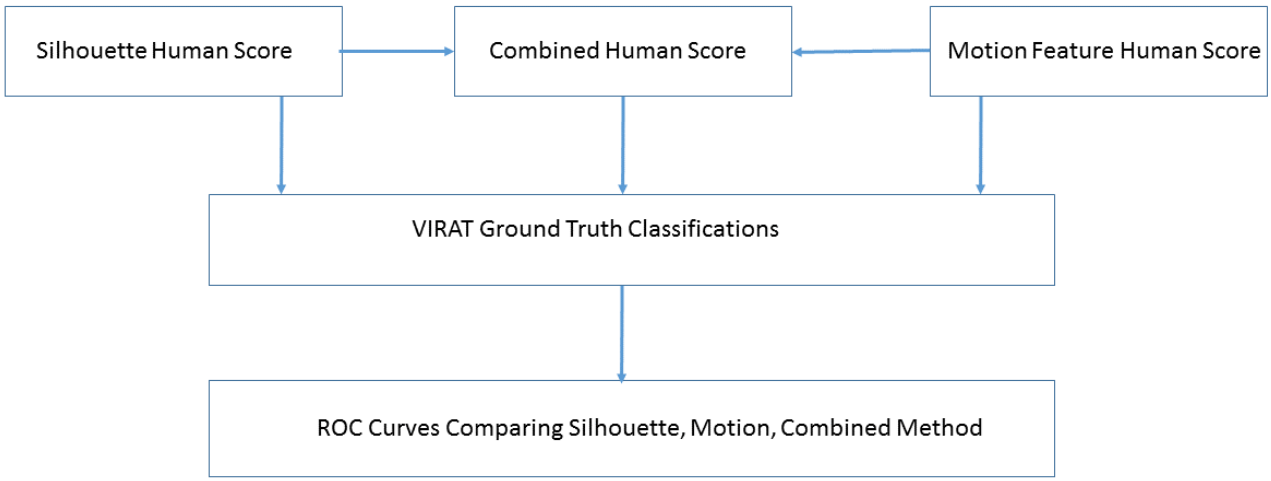


Figure 4: Evaluation pipeline.

We applied ROC curve analysis to compare the motion method, the silhouette method and the combined method. The ROC curve in figure 5 shows the recognition rate against the false alarm rate. Although, the motion method gives good results, it could not obtain very low false alarm rates, because the motion method found 4 objects with the highest recognition scores, which have ground truth non-human. Therefore, the ROC curve for the motion object starts in the point (0.088, 0.933) instead of the origin. Also very high recognition rates could not be obtained with the motion method, because the motion method classified 8 objects with the lowest recognition scores, which have ground truth human. Therefore, the ROC curve for the motion object ends in the point (0.846, 0.977) instead of the point (1.0, 1.0). We see that the distinctiveness of the random forest is rather limited. Figure 5 shows that the best overall results are obtained with the combined method.

For certain scenarios depending on the requirements the combined method is not the best choice. Our ROC curve analysis allows to set different thresholds for such scenarios. E.g. in a high security scenario at most 1 out of 100 humans may be missed, i.e. we require recognition rate ≥ 0.99. In a medium security scenario at most 1 out of 10 humans may be missed, i.e. we require recognition rate ≥ 0.9. In an easy access scenario at most 1 out of 10 detections may be false alarms, i.e. we require false alarm rate ≤ 0.1. Table 1 shows thresholds, recognition rates, and false alarm rates for these settings. Due to its low distinctiveness the motion method cannot meet the high security requirements. On the other hand it provides always a recognition rate of at least 0.933 with a low false alarm rate of 0.087. We see that for the high security case, the silhouette method is the best choice, for the medium security case the combined method is the best choice, and for the easy access use case the motion method is the best choice.

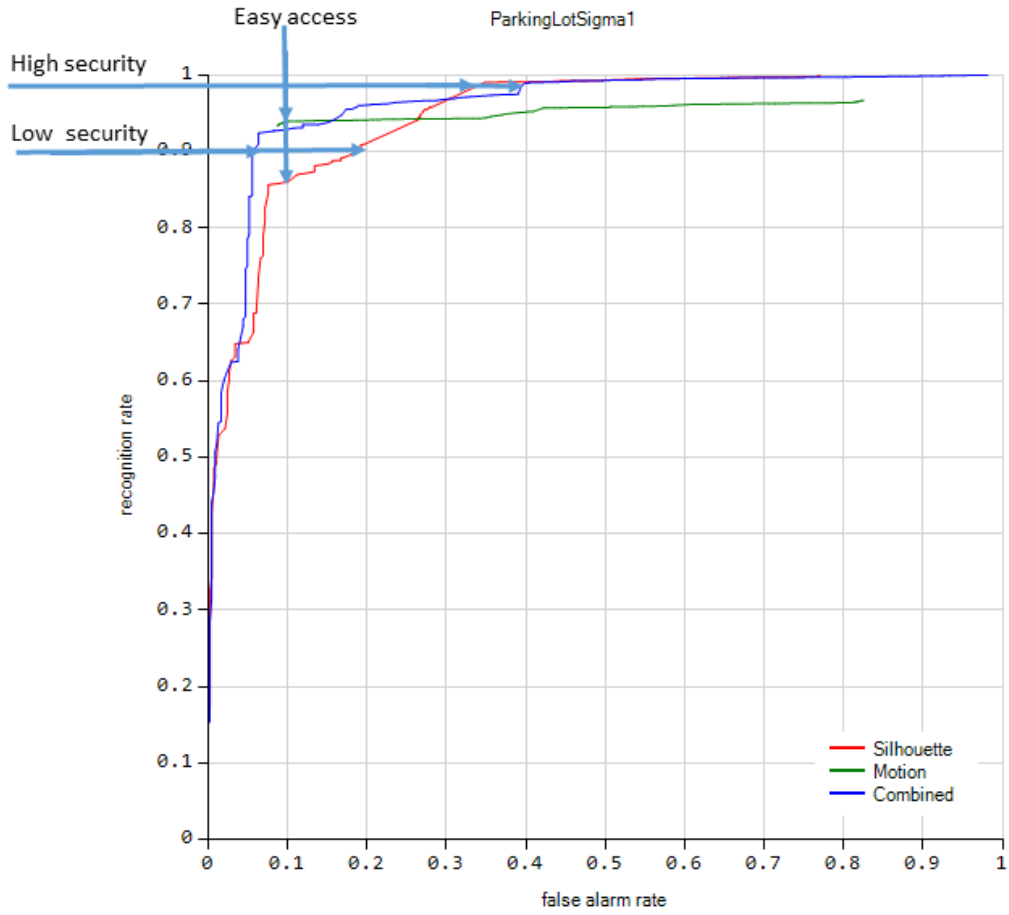


Figure 5: ROC curve analysis shows overall superior results for combining motion and silhouette analysis.

Scenario	Method	Recognition rate	False alarm rate
High security	Motion	Never recognition rate at least 0.990	
High security	**Silhouettes**	**0.990**	**0.321**
High security	Combined	0.990	0.407
Medium security	Motion	0.933	0.087
Medium security	Silhouettes	0.900	0.185
Medium security	**Combined**	**0.900**	**0.057**
Easy access	**Motion**	**0.939**	**0.100**
Easy access	Silhouettes	0.867	0.100
Easy access	Combined	0.933	0.100

Table 1: Different thresholds for different scenarios (best choice in bold). For the medium security case it is the best choice. However, for the high security case the silhouette method is the best choice, and for the easy access use case the motion method is the best choice.

5. CONCLUSIONS

In this paper, we presented a new method to detect persons in video by combining optical flow based motion analysis and silhouette based recognition. We showed that in general the combination of optical flow and silhouette analysis obtains superior results, compared to using the individual methods. However, we found by applying ROC curve analysis in a high security scenario the silhouette method should be preferred, and in an easy access scenario the motion method. In the analysis of the results it turned out, that the distinctiveness of the random forest was rather low (false alarm rate below 0.088 and recognition rate above 0.933 could not be reached). Improving this disappointing distinctiveness of the current random forest implementation is a future research issue.

ACKNOWLEDGEMENT

The work for this paper was supported by a grant from TNO's Small Business Innovation Research Program (**SBIR 2011**) in the project "patroonherkenning voorkomt vals alarm".

REFERENCES

[1] Oh, S. *et al*., "A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video", CVPR, (2011).
[2] Paul, M., Haque S.M.E., and Chakraborty, S. "Human Detection in Surveillance Videos and its Applications – a review". EURASIP Journal on Advances in Signal Processing 176, (2013).
[3] Wang, L., Geng, X., Leckie, C., and Kotagiri, R. "Moving Shape Dynamics: a Signal Processing Perspective" Proc. CVPR, 1-8, (2008).
[4] Lin, Z., and Davis, L.S., "Shape-based Human Detection and Segmentation via Hierarchical Part-Template Matching", IEEE Transactions on Pattern Analysis and Machine Intelligence 32(4):604-618, (2010).
[5] Bobick, A.F., and Davis, J.W., "The Recognition of Human Movement using Temporal Templates", IEEE Transactions on Pattern Analysis and Machine Intelligence 23(3):257-267, (2001).
[6] Cutler, L., Davis, L.S., "Robust real-time periodic motion detection, analysis, and applications", IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8):781-796, (2000).
[7] Dalal, N., and Triggs, B., "Histograms of Oriented Gaussians for Human Detection", Proc. CVPR, (2005).
[8] Dollar, P., Belongie, S., and Perona, P., "The Fastest Pedestrian Detector in the West", Proc. BMCV, (2010).
[9] Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D., "Object Detection with Discriminatively Trained Part Based Models"", IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9):1627-1645, (2010).
[10] Zhu, Q., Avidan, S., Yeh, M.-C., Cheng, K.-T., "Fast Human Detection using a Cascade of Histograms of Oriented Gradients", Proc. CVPR, 1491-1498, (2006).
[11] Zivkovic, Z., and van der Heijden, F. "Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction". Pattern Recognition Letters 27, 773-7780, (2006).
[12] Hongche, L., Hong, T.-H., Herman, M., and Chellappa, R. "Accuracy vs efficiency trade-offs in optical flow algorithms", Computer Vision and Image Understanding: CVIU, 72(3):271–286, (1998)
[13] Ancona, L., and Poggio T. "Optical flow from 1-d correlation: Application to a simple time-to-crash detector", International Journal of Computer Vision, 14(2):1573–1405, (1995).
[14] Deen, J-.W., "GPU Implementation for an Appearance Based Real-Time Object Tracker", Master Thesis, Free University, Amsterdam (2014).
[15] Hecht-Nielsen, R., "Theory of the Backpropagation Neural Network" IJCNN, 593-605, (1989).
[16] CASIA Gait Database. http://www.sinobiometrics.com, CASIA (2005).
[17] Shape data for the MPEG-7 core experiment CE-Shape-1, Part B.
http://www.cis.temple.edu/~latecki/TestData/mpeg7shapeB.tar.gz, (2006).
[18] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer, "Generic Object Recognition with Boosting", IEEE Transactions on Pattern Analysis and Machine Intelligence 28(3) (2008).
[19] B. Leibe, A. Leonardis, and B. Schiele. "Combined object categorization and segmentation with an implicit shape model". Proceedings of the Workshop on Statistical Learning in Computer Vision (2004).