SUSPICIOUS BEHAVIOR DETECTION IN PUBLIC TRANSPORT BY FUSION OF LOW-LEVEL VIDEO DESCRIPTORS

Dejan Arsić, Björn Schuller and Gerhard Rigoll

Technische Universität München, Institute for Human Machine Communication
Arcisstrasse 16, 80333 München, Germany
{arsic — schuller — rigoll}@tum.de

ABSTRACT

Recently great interest has been shown in the visual surveillance of public transportation systems. The challenge is the automated analysis of passenger's behaviors with a set of visual low-level features, which can be extracted robustly. On a set of global motion features computed in different parts of the image, here the complete image, the face and skin color regions, a classification with Support Vector Machines is performed. Test-runs on a database of aggressive, cheerful, intoxicated, nervous, neutral and tired behavior in an airplane situation show promising results.

Index Terms— Video Surveillance, Behavior Detection, Feature Fusion, Low Level Features

1. INTRODUCTION

Granting the passenger's safety in public transport is an expensive task, as additionally to closed circuit television CCTV systems a large staff is required to analyze video streams on line. Therefore it seems desirable to automatically monitor passenger compartments in trains, aircrafts and buses. The aim is to automatically detect passengers which might be a threat to others or themselves. As most of the passengers are sitting during the journey these will be observed with cameras mounted in the seat's back rests in front of the passengers. Due to this position the activity region is limited to the head and upper part of the body, as seen in figure 1. Given the fact that one camera and microphone are used for up to four seats in economy class the recorded audio signal is not processed, as voices cannot be assigned to a person yet and the level of noise in an aircraft is quite high.

In [1] we presented a frequency based fusion approach to discriminate between unruly and neutral activities. To provide the crew more detailed information between six different behaviors has to be discriminated.

Aiming at provision of practical usability, the monitoring system has to extract features and subsequently analyze them in real time in order to be able to react without any delay. Computation power is restricted by the demand to keep vehicles lightweight and energy supply is limited. This need is fulfilled by the using low-level features and a following fusion. Additionally the detectability of features has to be kept in mind, as reliable results have to be achieved. Therefore we decided to use so called global motion features, based on difference imaging, extracted from different parts of the image, here skin color regions, the face and the entire image. The classification is performed by windowing the resulting time-series of features. Each window providing a static vector is classified with Support Vector Machines (SVM). To enhance recognition results the features extracted from different image regions are processed separately and afterward fusioned in an additional step, which analyzes the SVM outputs.

The paper is structured as followed: Sec. 2 we will introduce the used feature sets, followed by their classification in sec. 3. To evaluate this approach a novel database for airplane behavior modeling has been recorded and will be presented in section 4. Finally results will be provided in sec. 5 prior to a short discussion in sec. 6.

Fig. 1. Exemplary camera position in an aircraft, with 2 persons observed at the same time

This work has partially been funded by the European Union within the FP6 IST SAFEE Project. www.safee.info. Special Thanks to Michael Schmitt.

2. LOW-LEVEL VIDEO FEATURES

Low-Level features seem to be the appropriate to solve the problem of computation power, restricted by the maximum number of processing units and the energy supply in an aircraft. Furthermore the reliability of the applied descriptors within the given scenario have to be considered. For instance a facial expression analysis as in [2] has been discarded, as the requirement of frontal faces cannot be granted and a model based image interpretation with Active Appearance Models (AAM) [3] would produce non reliable results.
Therefore we decided to use a set of global motion features [4], extracted from different parts of the image, as shown in figure 2. These are based on a simple difference image:

$$d(x, y, t) = I(x, y, t) - I(x, y, t + 1)$$

First the center of motion $m = [m_x, m_y]$ can be computed both in x and y direction:

$$m_x = \frac{\sum_{x,y} x * d(x, y, t)}{\sum_{x,y} d(x, y, t)},\ m_y = \frac{\sum_{x,y} y * d(x, y, t)}{\sum_{x,y} d(x, y, t)}$$

Since the behavior is independent of the passenger's location, only changes in the direction of movement and their value is used: $\delta m_x = m_x(t) - m_x(t-1)$ and $\delta m_y = m_y(t) - m_y(t-1)$
To distinguish between motions of large or small parts of the body the mean absolute deviation $\sigma = [\sigma_x, \sigma_y]$ is computed with:

$$\sigma_x = \frac{\sum_{x,y} d(x, y, t)|x - m_x|}{\sum_{x,y} d(x, y, t)}\ \sigma_y = \frac{\sum_{x,y} d(x, y, t)|x - m_y|}{\sum_{x,y} d(x, y, t)}$$

Furthermore the changes within a series of variance are considered: $\delta\sigma_x = \sigma_x(t) - \sigma_x(t-1)$ and $\delta\sigma_y = \sigma_y(t) - \sigma_y(t-1)$. Additionally the so called intensity of motion

$$i = \frac{\sum_{x,y} d(x, y, t)}{\sum_{x,y} 1}$$

is taken into account, which describes the changes in the entire image.
Enhancement of classification results is performed by splitting the image into several parts. The complete image is used as first feature set. A second one is extracted from the face area. Face tracking is initialized with a pyramidal search in a complete frame with Neural Networks (NN) as presented by Rowley [5]. Robustness is added by training the network with upright, tiled and faces rotated in depth. To avoid the computationally expensive full search and enable real time performance the condensation algorithm [6] is applied, which predicts possible positions of detected objects in a following frame. Reinitialization rate is kept low by using skin color as additional cue. The amount of skin is measured within the predicted particles and added to the probability provided by the neural network. This way the track is not lost, if the actual gaze is in an unlearned state. Two subsequent faces cannot be simply subtracted as head movement and imprecise detection results would add additional changes. These are avoided by performing block matching in the region surrounding the face, which should be the head. The face is then shifted to the position with the lowest difference in the observed area. Finally the difference image and global motion features within the face can be computed. These will represent facial expressions without any other influence.
In order to find hands and arms in the image skin colored areas are detected with the help of the skin locus [7]. A Pixel in the RGB plane is assumed as skin colored if:

$$g > ((-0.7279) * r^2 + 0.6066 * +0.1766))$$
$$\& (g < ((-1.8423) * r^2 + 1.5294 * r + 0.0422))$$
$$\& ((r - 0.33)^2 + (g - 0.33)^2 > 0.04^2)$$

with $r = \frac{R}{R*G*B}$ and $g = \frac{G}{R+G+B}$ As the position of the face is already known the surrounding skin color regions and the face itself can be excluded. Global motion features can now be extracted from subsequent skin color images.
This way every frame is represented by a vector with only 21 entries, which should be sufficient for the classification task, as a recognition of time series will be performed.

Fig. 2. Sample from the ABC with tracked face (r), difference image and face difference (m) and a difference image of skin regions (l)

3. RECOGNITION OF BEHAVIORS

So far we have extracted 21 features from every frame within a video sequence, which now have to be classified. Unfortunately even within one class examples can show a high variance because of unknown start and end state or the length of the sample. As a further segmentation of behaviors into so called submotions [2] failed, a dynamic classification with Hidden Markov Models (HMM) [8] was dismissed. Anyhow some results will be presented in 5. Therefore we decided to switch to static classification methods.
All samples are windowed with a length of 25 frames ($1s$) without any overlap. This resulted in 4511 frames for the

total of 461 samples. A vector containing all features with 525 entries can be created. For further experiments it can be assumed that the three feature groups are statistically independent. Each group could prefer several classes during the detection task. Three vectors with 175 entries each can be now created for each frame. This feature vectors x are now classified by the use of Support Vector Machines (SVM) [9]. These are a popular approach for the discrimination in two class problems. With a couple-wise multi class discrimination strategy these vectors x can now be classified. Trials have proved, that a polynomial kernel should be preferred.
For all of the four approaches models have been trained, with the results presented in the next section. To enhance recognition in an additional step the output of the SVMs for the three small sets are combined with three different methods: A simple majority vote, the sum of the computed probabilities and a fusion of the outputs with once more SVM.
Up to now single windows of 25 frames in a video sequence have been classified. This can be easily extended to the recognition on longer unsegmented sequences. It can be assumed that an order of behaviors can be neglected and just the appearance of the classes has to be taken into account. For the classification of a video sequence the results of each window are summed up.

4. THE AVIATION BEHAVIOR CORPUS (ABC)

Experts in the field of aircraft security defined a set of 6 activities, which could be important to detect. These are namely aggressive, cheerful, intoxicated, nervous, neutral and tired. Due to the lack of a public available database, a large database has been recorded.
In order to obtain equivalent conditions of several subjects of diverse classes we decided for acted data. It is believed, that mood induction procedures create more realistic reactions. Therefore we developed a scenario, which leads the subjects through a guided storyline. Five speakers have recorded announcements, which a hidden test-conductor will play to the actors. As a general framework a vacation flight with return flight was chosen, consisting of 13 and 10 scenes as start, serving wrong food, turbulences, conversation with a neighbor or falling asleep. Respecting a possible setup of the camera in the seat's back rest in front of each passenger, the activity radius is restricted to the head including the upper body , see figure 2. A seat for the subject was positioned in front of a blue screen. A condenser microphone and DV-camera were fixed without occlusions of the objects. 8 actors, both male and female, between 25 and 48 years in age, took parts in these recordings, which created a total of 11.5h video material. This has been presegmentated and annotated by three experienced male persons. Table 3 shows the final distribution of the 460 video clips with an average length of $8.4s$.

5. RESULTS

Due to the very limited amount of data in behavior analysis a reliable validation has to be performed. A $n - folded$ stratified cross validation (SCV) gives the opportunity to test and train disjunctive sets on the complete database. In the following we will present average result on a 5-fold SCV.
Table 1 shows the results of a 5-fold SCV on the set of 4511 segments with 25 frames length for each applied feature sets. Best results were achieved with all 21 extracted features (c), whereas global motions (g), face motions(f) and skin motions(s) perform weaker. These show higher recognition rates if a subsequent fusion is performed. A simple majority vote (mv) or the analysis of the outputs with SVM is outperformed by the addition of the provided probabilities. In total 60.9% of the segments could be correctly classified. In comparison a dynamic classification with contionous left-right HMMs has been performed both on the segments (HMM25f) and the complete sequences (HMMseq). These results state, that a static recognition should be performed. It can be seen clearly, that these recognition results remain below the performance of SVM. The fusion of the 3 subsets performed with a maximum of $49,1\%$

type:	*c*	*g*	*f*	*s*
SVM	57.5 %	55.1 %	54,9 %	51,6 %
HMM25f	47,5 %	45,3 %	41,7 %	43,1 %
HMMseq	51,3 %	49,2 %	46,8 %	47,2 %
fusion:		*mv*	*sum(p)*	*SVM*
		56.7 %	60.1 %	59,4 %

Table 1. Recognition results on single windows with the different feature sets. A comparison with HMMs on the segments and complete sequences is also provided

Table 2 illustrates the recognition results on complete video clips from the ABC. It can easily be seen, that the analysis of windows classified with the sum of probabilities reaches the best results. Interestingly the classification with the complete feature set is outperformed by far, although it had the highest accuracy on the set of segments.

type:	*c*	*g*	*f*	*s*	sum(p)
	60.8 %	59.3 %	57.3 %	56,6 %	66,52 %

Table 2. Recognition results on video data with with the different feature sets

The confusions of the final test-run with the 3 single feature sets and a following addition of the probabilities provided bay the SVMs are shown in table 3. As can be seen most confusions occur between the classes cheerful and aggressive and

between tired and aggressive. This phenomenon can easily be explained with the data set. People pretending to be tired would yawn and move their hand to the face, so the motions look quite similar. Cheerful actors were mostly overacting and moving a lot or just laughing a little bit, which would also explain the confusions. In the table also the f_1-measures are presented to show the ratio between recall and precision. Apart from intoxicated behavior, which is recognized worst, all classes are recognized almost balanced.

truth	*a*	*c*	*i*	*ner*	*neu*	*t*	[#]	$f_1[\%]$
aggr	84	5	4	3	0	0	96	67.4
cheer	19	66	1	16	0	3	105	67.6
intox	3	9	12	9	0	0	33	53,3
nerv	18	0	0	71	4	0	93	67,6
neu	8	10	0	13	48	0	79	71,4
tired	21	0	0	5	3	25	54	63,3,3

Table 3. Confusions of behaviors and f_1-measures by use of SVM in a 5-fold SCV with 3 separate feature sets on the ABC

However the six classes can be reduced to the initial problem, the detection of unruly behavior, here aggressive, intoxicated and nervous, in an aircrafts. From a total of 222 unruly samples 204 $(91,8\%)$ were correctly classified, neglecting which behaviour it might be. Neutral activities in contrast have more often been mixed up with unruly ones, with 66 missclassifications in 238 samples.

6. CONCLUION AND OUTLOOK

We have shown promising results for the behavior detection task with low-level video features in section 5. Unruly behaviors can be detected with high reliability, but false alarm rate has to be decreased yet. In our future work we aim to increase the size of the presented database. Furthermore exhaustive perception tests have to be performed with individuals on the corpus, in order to be able to compare our results to human observers. Including further visual features, such as optical flow, the position of facial features and low level activities like eye-blinking. Therefore a model has to be created, which can deal with the limits of these features, for instance if only one eye is visible. To notice changes in a video stream as fast as possible a segmentation has to be introduced, as short times of unruly behavior would be suppressed by long lasting periods of neutral activities.
It seems commonly agreed, that a fusion of several input cues is advantageous [10]. Hence the additional use of low-level audio descriptors, as presented by Schuller et all [11] with recognition rates up to 86.9% on 7 classes, will be investigated. The challenging task will now be the synchronisation and fusion auf audiovisual features. This step should provide a more robust system.

7. REFERENCES

[1] D. Arsić, F. Wallhoff, B. Schuller, and G. Rigoll, "Video based online behavior detection using probabilistic multi-stream fusion," in *Proceedings IEEE International Conference on Image Processing (ICIP) 2005, Genoa*, Sept. 2005.

[2] D. Arsić, J. Schenk, B. Schuller, and G. Rigoll, "Submotions for hidden markov model based dynamic facial action recognition," in *Proceedings IEEE International Conference on Image Processing (ICIP) 2006, Atlanta*, Oct. 2006.

[3] T. Cootes and C. Taylor, "Statistical models of appearance for computer vision," *Technical report, University of Manchester, Wolfson Image Analysis Unit, Imaging Science and Biomedical Engineering, Manchester M13 9PT, United Kingdom.*, Sept. 1999.

[4] M. Zobl, F. Wallhoff, and G. Rigoll, "Action recognition in meeting scenarios using global motion features," in *Proceedings Fourth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-ICVS), Graz Austria*, Mar. 2003, pp. 32–36.

[5] H. Rowley, S. Baluja, and Takeo Kanade, "Neural network-based face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, Jan. 1998.

[6] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29(1), pp. 5–28, 1998.

[7] B. Martinkauppi M. Soriano, S. Huovinen and M. Laaksonen, "Skin detection in video under changing illumination conditions," in *Proc. 15th International Conference on Pattern Recognition, Barcelona, Spain*, 2000, pp. 839–842.

[8] Lawrence Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989, vol. 77, pp. 257–286.

[9] B. Schoelkopf, "Support vector learning," *Neural Information Processing Systems*, 2001.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[11] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion recognition in the nise applying large acoustic feature sets," in *Proceedings Speech Prosody 2006, 02.-05.05 2006, Dresden, Germany. ISCA*, Mar. 2006.