A HIERARCHICAL APPROACH FOR VISUAL SUSPICIOUS BEHAVIOR DETECTION IN AIRCRAFTS

D. Arsić, B. Hörnler, B. Schuller, and G. Rigoll

Institute for Human Machine Communication
Technische Universität München, Germany
(arsic - b - 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.

Index Terms— Behavior Detection, Low Level Features, Fusion, SVM, Surveillance

1. INTRODUCTION

The aim of the EU funded Project SAFEE[1] is to increase on-board security in aircrafts, by detecting potentially threatening situations automatically. The SBDS (Suspicious Behavior Detection System) work package aimed to provide a systematic solution to monitor people within enclosed spaces in the presence of heavy occlusion, analyze these observations and derive threat intentions [1]. These are then reported to the crew, which will decide which steps to take to gain control over the situation. Threats may include unruly passenger behavior (due to intoxication, etc.), potential hijack situations, and numerous other events of importance to both flight crew and ground staff. Indicators of such events have been collected to a set of so called pre-determined indicators (PDIs), such as nervousness or frequent visits to the lavatory. These PDIs have been assembled to complex scenarios, which can be interpreted as combination and temporal sequence of so called low level activities (LLA). In order to detect these three systems have been employed by various subject matter experts (SME):

- **Acoustic Event Detection**
- **Tracking of passengers**
- **Low level visual stress detection**

All observations are subsequently fed into a common scene-understanding module, which produces a reasoned output from the various elements. In order to achieve this aim aircrafts could be equipped with cameras, which are observing the isles, the cockpit region and seated passengers, and microphones spread all over the cabin. Fig. 1 illustrates possible sensor positions in the aircraft. In this work the modules for seated behavior detection have have been implemented, and outputs have been send to a scene interpretation module [1].
In order to robustly pick up visual stress indicators it is important to know how a LLA can

[1]EU Funded FP6 project, Grant Number:AIP3-CT-62003-503521, SAFEE: Security of Aircraft in the Future European Environment

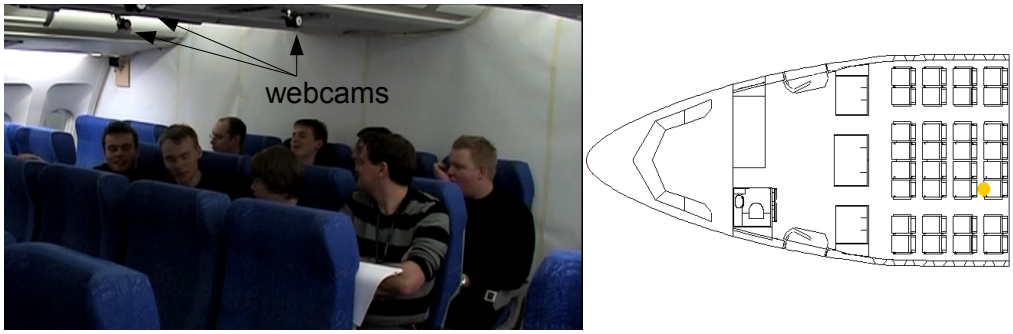

Fig. 1. Exemplary camera mounting in an A340 mock-up. The terminal on the right side indicates seat positions with unruly passengers.

be characterized. According to experts these can be further decomposed into so called Low Level Features (LLF), which can be chosen with respect to their detectability [2]. These LLFs can subsequently be combined to detect a more complex behavior.

2. THE AIRPLANE BEHAVIOUR CORPUS

As audiovisual databases are spares in the research community it has been decided to record a new database dedicated to the topic of on board suspicious behavior detection in aircrafts. Experts in the field of security have provided input on possible threat indicators, which are namely: aggression (a), intoxication(i) and nervousness(nr). A fully automated system should be able to pick these up during a flight and reliable separate them from neutrality (nu), cheer (c) and tiredness (t). In order to obtain data in equivalent conditions of several subjects of diverse classes we decided for acted behavior. There is a broad discussion in the community with respect to acted vs. spontaneous data, which we will not address herein. However, it is believed, that mood induction procedures favor realism in behavior. Therefore a script was used, which leads the subjects through a guided storyline: prerecorded announcements by five different speakers were automatically played back controlled by a hidden test-conductor. As a general framework a vacation flight with return flight was chosen, consisting of 13 and 10 scenes as start, serving of wrong food, turbulences, falling asleep, conversation with a neighbor, or touch-down. The general setup consisted of an airplane seat for the subject positioned in front of a blue screen. Camera and a condenser microphone AEG 1000S MK II were fixed without occlusions of the subject. 8 subjects in gender balance from 25a to 48a (mean 32a) took part in the recording. The language throughout recording is German. A total of 11.5h video was recorded and annotated independently after pre-segmentation by three experienced male labelers within a closed set as seen in tab. 1. This table also shows the final distribution of samples with total inter-labeler-agreement. This set will be referenced as ABC (Airplane Behavior Corpus) in the ongoing. The average length of the 405 clips in total is 8.4s.

aggr	*cheer*	*intox*	*nerv*	*neu*	*tired*	
96	105	33	93	79		54%

Table 1. Amount of samples for each class in the ABC

2.1. Low Level Feature Extraction

Discussions within the SAFEE consortium have shown large scale privacy issues for surveillance tasks in public transportation systems. While video has been widely accepted, as the public already has become accustomed to the presence of CCTV, the processing of acoustic features seems to be more problematic. It is commonly agreed that eavesdropping is not accepted and considered as privacy violation, while cameras are more likely to be accepted for security reasons. Therefore this work has been restricted to visual cues. The speech emotion community has recently shown great interest in the security domain and also investigated various approaches for this classification task with promising results [3].

A major issue for the aircraft application scenario is the need for real time capable feature extraction and subsequent classification. Additionally the features have to be extracted robustly from every frame. Facial feature points as defined by the MPEG7 standard [4] have been discarded, as they are not visible in all frames. Nevertheless these have been investigated in [5] by Wimmer. Therefore we decided to focus on global motion features [6] extracted from various parts of the image 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:

$$m_x = \frac{\sum_{x,y} x * d(x,y,t)}{\sum_{x,y} d(x,y,t)}, m_y = \frac{\sum_{x,y} y * d(x,y,t)}{\sum_{x,y} d(x,y,t)} \quad (1)$$

Since the behavior is independent of the passenger's location, only changes in the direction of movement and their value is used: $\delta m_{x/y} = m_{x/y}(t) - m_{x/y}(t-1)$

To distinguish between motions of large or small parts of the body the mean absolute deviation

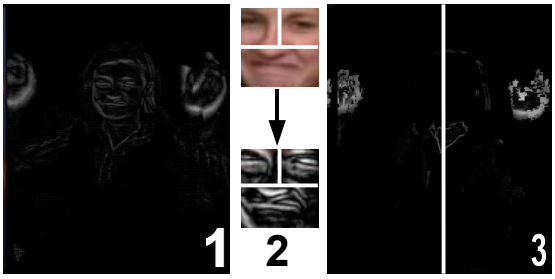

Fig. 2. Visualization of applied features: Extracted Global Motions, Face detection and facial feature extraction, Skin motion detection in the left and right half of the image.

$\sigma = [\sigma_x, \sigma_y]$ is computed with:

$$\sigma_x = \frac{\sum_{x,y} d(x,y,t)|x - m_x|}{\sum_{x,y} d(x,y,t)}$$
$$\sigma_y = \frac{\sum_{x,y} d(x,y,t)|x - m_y|}{\sum_{x,y} d(x,y,t)} \quad (2)$$

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

$$i = \frac{\sum_{x,y} d(x,y,t)}{\sum_{x,y} 1} \quad (3)$$

is taken into account, which describes the changes in the entire image.

These features are extracted, as illustrated in fig. 2, from various different image parts and categorized by following means:

Global Motion Features: In the first place the global motion features g are computed from the entire image for each frame in a video sequence.

Face Motions: Real time face tracking based on an initialization with a Neural Network [7] combined with the condensation algorithm [8] is utilized to restrict feature extraction the facial region, where face motions can be computed. Besides the computation on the entire detected face f_c, features are extracted from three different face regions f_3 in order to model the face

more detailed. With the upper half split in two parts both the left and right eye position are approximated, where the most dominant movement is produced by eye blinks. Furthermore a feature extraction is performed in the lower half of the face, which will model movements of the mouth.

Skin Motions: In a third stage hands are detected by applying a simple skin detection algorithm [9]. As the position of the face is already known it can be assumed, that remaining skin parts are representing hands and arms. Skin motion features are computed either computed for the entire frame s or separately for the right and left arm s_2, by simply splitting the video stream in the middle.

In addition to the face motion features the face's displacement δf is determined in x and y direction based both on the face detector output and the smoothed detector output. Face motion in depth are considered by the computation of size changes. Furthermore the position change of the left and right eye brow $\delta brow$ is computed over time. Both the face displacement δf and $\delta brow$ have been added to the facial features f_3. Table 2 shows the detailed number of all 81 extracted features for each type and the entire feature set c.

g	f	f_3	s	s_2	c
9	9	36	9	18	81

Table 2. Number of features extracted in each frame for each LLF class.

2.2. Suspicious Behavior Detection

So far up to 81 features have been extracted from every frame within a video sequence, which now have to be classified. Various different classifiers have been evaluated with the extracted feature sets to find the most appropriate solution. The utilized classifiers can be basically divided into two groups: Dynamic classifiers are able to classify features of variable length, which are naturally given by video sequences of different durations. Hidden Markov Models [10] have been established as standard method for the classification of data with unknown length. As the sample contain only one single activity no further segmentation has been required and detection could be performed without dictionary. Static classifiers in contrast are only able to process feature vectors with constant length. Therefore the video sequences have to be preprocessed to guarantee a fixed vector length. The probably most convenient solution is to process every single frame in a video independently, which would guarantee a fixed length. This method incorporates two major drawbacks: Classification becomes more expensive and hence real time performance cannot be granted. Additionally dynamic feature changes are neglected, which could be important to describe a behavior. By segmenting each video with a window with a constant length of 25 frames without any overlap, dynamic changes can be still respected. Subsequently the resulting vector x with the size of $N = 25 \times features$ can be classified. From the variety of static classifiers support vector machines (SVM) [11] have been chosen. Due to the segmentation of the video sequence a subsequent fusion of the segment based classification has to be performed. This is done either by a simple majority vote or a further analysis of the produced probabilities. In order to provide a reliable evaluation, various feature configurations have been tested. Likewise it has been tested, whether the classification of combined large vectors (early fusion) or the classification of smaller feature sets and a subsequent fusion (late fusion) should be preferred [12]. As most common classifiers produce scores for each class, these can be combined to a stronger classifier. Although Jaeger [13] proposes to combine classifier outputs based on their confidence, a simple accumulation of scores is frequently sufficient, in case the classifiers can be considered as independent. The largest combined probability is then used as

detector output.
Up to now the behavior detection task has been considered as six class problem. Naturally the reduction of classes will result in higher detection results, even guessing by chance will be more reliable. Likewise unnecessary classes could be removed leaving the interesting ones behind or several classes could be summed up to one single class with more training examples. Basically there are only two important classes for the passenger surveillance task: unruly behavior, including aggression, nervousness and intoxication, and neutral behavior, including cheer, neutrality and tiredness. Nevertheless it is important to receive more detailed information on the observed behavior, than just an alert signal. Therefore a hierarchical approach, as illustrated in fig. 3, is proposed. The arising question is now how to design the classifier's stages, as the tree can be build up with a large variety of possibilities. Anyways, this work suggest to introduce two more classes, namely unruly and normal, for the above mentioned reasons. In the first stage the by far simpler two class problem of separating neutral an unruly behavior is solved. Experiments have shown a sufficient performance of SVMs for this task. Subsequently two three class problems, with respect to the output of the previous stage, have to be solved with. Therefore two separate models have been trained trained for the unruly or neutral classes.

2.3. System Evaluation

In this seaction the results of the suspicious behavior classification task will be presented and compared to other work. Due to the very limited amount of data, a *n-fold* stratified cross validation (SCV) strategie has been used for a reliable evaluation. This way disjunctive sets have been trained and tested with the entire database.
Tab. 3 shows the results of a 5-fold SCV on segments with 25 frames length for all applied feature sets. The 2class (S2), 6class (S6) and

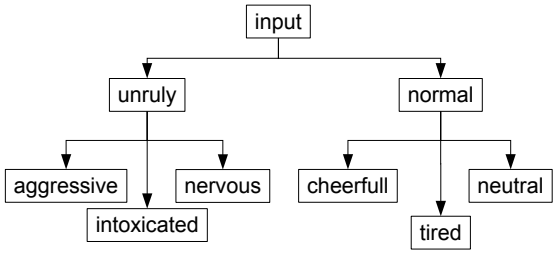

Fig. 3. Hierarchical classification of unruly and normal behavior patterns with SVMs. The introduction of two additional classes enhances the overall recognition rate of the six class problem

HMM (H6) segmentation approaches were used to classify a total of 4511 frames, while the classification of unruly and normal behavior included (Su3, Sn3) only about half the data. Due to the lower dimensionality the classification of smaller feature sets outperforms the classification of the larger ones. This observation is extremely noticeable for the entire feature set, as the amount of features drastically exceeds the number of examples. Summing up the scores of the three smaller sets, namely global motion g, skin motion s and face motion f, creates the best recognition rate for all approaches. This can be easily explained, as each activity is probably best characterized by one of the feature classes. The trained classifiers hence are able to separate different samples in a better way than other ones. As noted previously the reduction of classes and the smaller database size will result in a higher recognition rate, which also can be seen in tab. 3. Further it has to be noted, that unruly behaviors are by far better discriminated than normal ones, which indicates the small inter class variance of the normal behaviors.
As entire video sequences have to be classified the segments have to be combined in the last step. Therefore once more the classifier outputs have to be combined, which is done by a simple addition of the scores over the duration of the video. As it is unlikely, that the weakest clas-

method	g	f	f_3	s	s_2	g+f+s	c	sum
S6	55.1	54.9	52.1	51.6	51.8	57.5	44.7	60.1
S2	73.1	72.0	70.9	76.2	75.5	75.2	69.2	80.5
Su3	72.9	77.9	74.3	75.1	73.3	79.5	68.3	81.2
Sn3	64.3	71.4	70.6	67.9	63.8	73.9	61.4	74.1
Hf6	45.3	41.7	41.3	43.1	40.9	47.5	29.7	49.3

Table 3. Recognition rates of windows with 25 frames.

sifiers will outperform the stronger ones after a temporal integration only the results with the most promising feature sets are shown in tab.4. As can be seen the temporal integration of the single classifiers is able to remove some of the errors of the segment based recognition process, where especially the classifiers of the tree based approach reach high classification rates. Once more it can be observed, that the normal class is harder to discriminate than the aggressive one. Yet the tree based classification (St6) shows the best performance for this classification task. Obviously the recognition rates do not significantly differ for the HMM based classification of segments (Hf6) or entire sequences (Hs6), and fail compared to the static classification with SVMs. A more detailed analysis of the achieved

	g	f	s	g+f+s	sum
S6	59.3	57.3	56.6	60.8	66.5
S2	81.8	79.9	84.5	85.9	87.9
Su3	81.2	85.2	83.3	87.0	90.7
Sn3	66.7	83.8	70.4	79.6	75.2
St6	69.3	72.5	71.7	73.8	**74.9**%
Hf6	50.1	45.3	46.7	49.5	52.6
Hs6	49.2	46.8	47.2	51.3	52.3

Table 4. Recognition rates for entire sequences both for 6 class problem and the hierarchical approach, which also shows the best performance.

results is provided by the confusion matrix in tab. 5. This matrix illustrates the confusions between the classes. As can be seen nervous behavior can be recognized best, whereas intoxication is recognized worst. $f1$ measures for the other classes are distributed almost equally.

In the past other promising approaches have been evaluated with the ABC corpus, and shall now be compared to the presented results. Deformable models have been fitted to the passengers' faces in [14] for feature extraction. After feature reduction with a sequential floating forward search, the remaining 157 features were used for a time-series-analysis, which resulted in 61.1% recognition rate. This shows the advantage of the approach presented in this work, which did not rely on complex facial features. Further experiments with acoustic behavior recognition in [15], showed a classification approach with a large set of low level audio descriptors and functionals, resulting in 73.3% recognition rate. Evidently both acoustic and visual behavior detection methods seem to operate at the same level. More reliable results can be achieved by combining audiovisual features [15], which has been conducted in [5] with a recognition rate of 81.1%, though relying on facial features.

3. CONCLUSION

In this work we have presented a method for visual suspicious behavior detection in aircrafts. The hierarchical classification of low level features with a tree of SVMs achieved the best known results by now. The fusion with further acoustic features could provide even more reliable recognition results.

truth	a	c	i	nr	nu	t	f_1[%]
a	85	0	6	2	3	0	77.3
c	11	78	3	0	13	0	81.3
i	5	5	17	2	4	0	58,6
nr	8	0	3	73	9	0	83.9
nv	7	2	0	4	64	2	71,1
t	8	2	2	0	8	34	75,5

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

4. REFERENCES

[1] N. L Carter and J. M. Ferryman, "The safee on-board threat detection system," in *International Conference on Computer Vision Systems*, May 2008, pp. 79–88. 1

[2] 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, Italy*, Sept. 2005, pp. 606–609. 2

[3] B. Schuller, M. Wimmer, D. Arsić, T. Moosmayr, and G. Rigoll, "Detection of security related affect and behaviour in passenger transport," in *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia.* 2008, pp. 265–268, ISCA, 22.-26.09.2008, ISSN 1990-9772. 3

[4] J. Ostermann, "Animation of synthetic faces in mpeg-4," *Computer Animation*, pp. 49–51, 1998. 3

[5] M. Wimmer, B. Schuller, D. Arsić, B. Radig, and G. Rigoll, "Low-level fusion of audio and video feature for multi-modal emotion recognition," in *Proc. 3rd Int. Conf. on Computer Vision Theory and Applications VISAPP, Funchal, Madeira, Portugal*, A. Ranchordas and H. Araujo, Eds., 2008, vol. 2, pp. 145–151. 3, 6

[6] 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. 3

[7] 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. 3

[8] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29(1), pp. 5–28, 1998. 3

[9] 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. 4

[10] 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. 4

[11] B. Schoelkopf, "Support vector learning," *Neural Information Processing Systems*, 2001. 4

[12] D. Arsić, B. Schuller, and G. Rigoll, "Suspicious behavior detection in public transport by fusion of low-level video descriptors," in *Proceedings 8th International Conference on Multimedia and Expo ICME 2007, Beijing, China*, June 2007, pp. 20018–20021. 4

[13] S. Jaeger, "From informational confidence to informational intelligence," in *In proceedings 10th International Workshop on Frontiers in Handwriting Recognition, IWFHR*, October 2006, pp. 173–178. 4

[14] B. Schuller, M. Wimmer, D. Arsić, G. Rigoll, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proceedings ICASSP 2007, IEEE, Honolulu, Hawaii, USA, 15.-20.04.2007*, Apr. 2007. 6

[15] M. Schröder, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, and B. Schuller, "Towards responsive sensitive artificial listeners," in *Proc. 4th Intern. Workshop on Human-Computer Conversation, Bellagio, Italy*, 2008, 06-07.10.2008. 6