VISION-BASED ONLINE MULTI-STREAM BEHAVIOR DETECTION APPLYING BAYESIAN NEWORKS

Dejan Arsić,Frank Wallhoff, Björn Schuller,Marc Al Hames, and Gerhard Rigoll

Institute for Human-Machine-Communication, Faculty of Electrical Engineering,
Technische Universität Mnchen
Arcisstr. 21, D-80290 München, Germany
(arsic, wallhoff, schuller, rigoll) @tum.de

ABSTRACT

In the present treatise, we propose an approach for a highly configurable image based online person behaviour monitoring system. The particular application scenario is a crew supporting multi-stream on-board threat detection system, which is getting more desirable for the use in public transport. For such frameworks, to work robust in mostly unconstrained environments, many subsystems have to be employed. Although the research field of pattern recognition has brought up reliable approaches for several involved subtasks in the last decade, there often exists a gap between reliability and the needed computational efforts. However in order, to accomplish this highly demanding task, several straight forward technologies, here the output of several so-called weak classifiers using low-level features are fused by a sophisticated Bayesian Network.

1. INTRODUCTION

Video surveillance is an expensive task as additionally to the technical equipment large staff is needed for surveillance, either analyzing video online or actual presence. Therefore it seems desirable to automatically monitor people's behavior. A possible application may be the automatic observation of the passenger compartment of a plane. The goal is detection of e.g. aggressive persons, passengers illicitly using electronic devices or just ill passengers. In this work we present first integrated approaches and solutions in this complex research field. Aiming at provision of practical usability, the monitoring system has to compute required features in real time in order to be able to react without delay.
Fulfillment of this need is achieved by using low-level features to detect activities on a lower semantic level, meaning segmenting complex behavior into several independent activities. An additional advantage of this segmentation is comprised in the possibility of its description by a few attributes in a comfortable way. We will further present a low level representation of the bearing of aggressive or nervous persons using eye and lip movement, such as yawning or laughing. So called global motion features and the movement of the head are taken into account. All computed classifications are considered as unreliable due to the intentionally simple nature of the applied features. This circumstance is balanced by use of a multi-stream fusion to drastically increase robustness.
At the moment only video streams are analysed, as only one microphone per camera is availible. In future works microphone arrays will be installed in the surveilled areas, for also taking audio into account.

2. IMAGE ACQUISITION

Special requirements for technical equipment evolves from room constraints and limited space in most public transport systems and additional characteristics of the application area.
In order to obtain high quality images, cameras providing progressive video material with a PAL-resolution of 720×576 pixels are used. The capturing devices transmit uncompressed image-data to avoid additional noise, allowing for undisturbed difference image computation. The field of view is expanded by wide angle lenses, which results in a warped border area that can be unwarped in a preprocessing step. Furthermore Near Infrared (NIR) filters combined with infrared lamps are applied to eliminate disturbing influences caused by external lighting conditions, and to enhance reliability of the person- and face detection systems.
The amount of needed on-board video devices within the given scenario and hence the required processing capacity is kept very low by monitoring a couple of seats by a single camera. As consequence optimal camera placement is crucial, respecting also the freedom of movement of the crew and the passengers. Inside a plane one exemplary mounting position of a camera is within the overhead bins enabling simultaneous observation of three neighbored seats in two adjacent rows. Fig. 1 shows the view from the determined optimal camera position. As can be seen from the image

From: ICME 2005, Amsterdam, IEEE

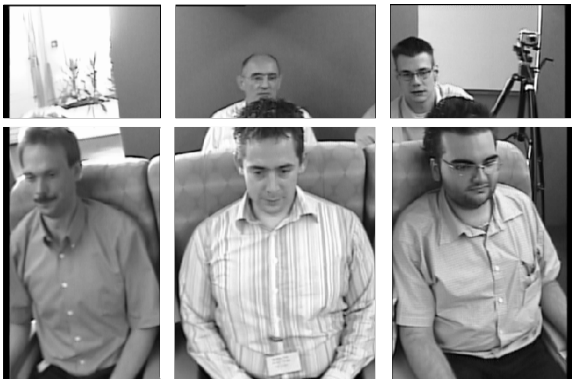

Fig. 1. Exemplary Field of view in an Airbus Cabin, and segmentation of image

there is a tradeoff between the camera position and the resulting size of the passengers' face.

3. PERSON LOCALIZATION AND TRACKING

The localization of passengers is implemented over the view angle independent detection of their faces and a subsequent expansion of the facial area to cover the upper body by an empirically determined geometry. To detect faces in still images, an extended neural network based face detection approach similar to that proposed by Rowley [1] is used.

Although this underlying system provides robust face detection including the possibility of estimating the gaze direction, a problem arises considering computational effort: Application of sampling with a sliding window technique at different scales leads to high calculation times. This fact makes real time face detection almost impossible. However, to circumvent the computational problems, but simultaneously keep the precision of the approach, an intelligent algorithm to predict the position and sizes of faces in image sequences is obeyed. The chosen method, introduced by Isard and Blake, is called Condensation algorithm [2]. After an initial detection process N particles, each representing the possible location and size of a face, are randomly selected, allowing for the presence of any number of faces in an image. In a next step these particles are shifted and rescaled according to prior estimated dynamical models, which may be adjusted to the specific context. In the present situation small movements are sufficient, as sitting passengers tend to show only such. After the dynamic drift all particles undergo a second, random diffusion. Thereafter only these N predictions have to be tested by the neural net, for example N=50, instead of the full search with 50.000 samples. Particles with a low probability are discarded. Prediction and testing are continuously repeated for succeeding frames in real time. If no particles remain, a re-initialization is required. A similar system for omni directional views has been implemented and studied in more details in [3].

Tracking faces of sitting passengers can be simplified by splitting up the actual image into several small images according to the seats arrangement, as can be seen in 3. A huge advantage, compared to face tracking on the whole image, is the systems ability to automatically reinitialize on small areas if the track is lost in a part of an image. Likewise other parts of the image are not affected by the time consuming reinitialization.

4. LOW LEVEL ACTION CLASSIFIERS

Prior to the presentation of the actual implementation of a detection approach, let us define the behaviours of interest, such as nervousness and aggressiveness. It is assumed, that these behaviours can be characterized by observing several low-level activities. An important step is the selection of these descriptions, able to represent in their sum a more complex behaviour. As most passengers in a train or airplane are sitting, the observable actions are performed in the upper part of the body and the face. Therefore single observations are chosen, respectively lip movement (yawning, speaking, laughing), eye movement (blinking) and global motion (head/body movement, sit down, stand up, being present/absent). Unfortunately the simple presence of an activity in a single frame does not contain any information of the actual behaviour the frequency of the activity is used for the description. Movement in contrast is represented by the average intensity. A nervous person often blinks with the eyes, tends to move with a higher frequency, stands up and sits down several times and might talk and laugh little. Respectively, a frequently yawning person can be assumed tired with a higher probability. Before a proper detection, such scenarios must be analyzed and defined. We decided for simple but fast classifiers favouring real time performance. The obliged initially high error rate is compensated within a multi stream fusion described later.

All desired low-level features have to be detected in real time. So complex classifiers had been dismissed, and weak but fast classifiers have been implemented. Global motion features for instance can easily be computed using difference images [4]. Head Movement needs not to be computed separately, as we compute the faces position in every single frame. Eye Movement, such as blinking, can also be computed with difference images, as a closed eye is lighter than an open eye. So just changes of the actual state have to be detected, applying a decision stump with learned values. The detection of Lip movement is far more complex and is performed applying Support Vector machines (SVM) [5].

Due to the lack of available real world data a database with training and test material has been created. 10.000 images of yawning, talking and laughing performed by 15 subjects are comprised. Further 250 blinks and 10 minutes of head

Activity	Seat	Rise	Sit Down	Movement	Head	Blink	Yawn	Lough	Speak
Sleep	0	0	0	2	3	0	0	0	0
Talk	15	0	1	3	3	25	0	0	10
Sit	0	0	0	5	3	15	0	0	0
Watch TV	0	0	0	3	2	20	0	10	0
Tired	0	0	0	4	2	10	25	3	5
Kid	30	3	3	15	18	30	0	0	25
Nervous	0.5	1	2	9	7	25	0	5	10
Aggressive	30	1	0	25	14	20	0	15	30

Table 1. APart of the created behavior database. Each activity is represented by its frequency

movement, uprising and sitting down actions are contained. The creation of a video database containing behaviours we intend to detect is very time consuming as an almost unlimited number of scripts has to be developed, and these have to be filmed with actors. The subsequent annotation of the resulting material is also very time consuming. Therefore a group of experts determined various frequency based representations of ten different behaviors. Table 1 shows a part of the collected data, which describes the frequency of each low-level feature based on a minute. This way 250 representations were developed by experts, which should be sufficient for first tests of the developed fusion algorithms.

5. MULTI STREAM FUSION USING BAYESIAN NETWORKS (BN)

Most expert systems describing probabilistic relationships between patterns contain a degree of uncertainty, as probabilities cannot be estimated exactly. Especially in our desired application area the collection of training material and the determination of statistical dependencies between actions and behaviours are rather inconvenient and unsatisfying.

The mathematical background of BNs [6] provides the opportunity of integration of incomplete and uncertain information in a hybrid architecture [7]. Due to the alluded benefits BNs enjoy growing popularity in knowledge modelling concerning artificial intelligence as well as in pattern recognition tasks. The major theoretical basics and capabilities of BN's in probabilistic reasoning are summarized here: Every BN consists of a set of nodes representing state variables X. The nodes are connected by directed acyclic edges expressing quantitatively the conditional probabilities of nodes and their parent nodes, see figure 5. A BN can be completely described in structure and conditional probabilities by its joint probability distribution. Let N denote the total of random variables, and the distribution can be calculated as:

$$P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P\left(X_i | \text{parents}(X_i)\right) \quad (1)$$

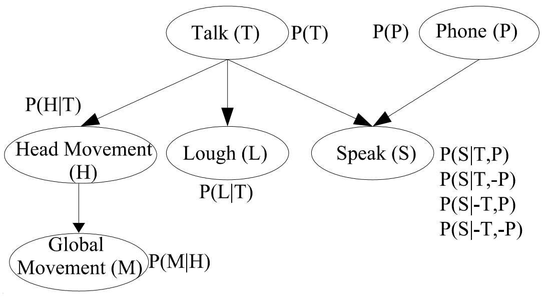

Fig. 2. Representation of "Talking To Neighbour"

Figure 5 illustrates an example for a possible implementation of a BN structure for a multi-stream fusion system, whose toppology is derived from expert knowledge. The root node in the BN resembles the classification of the momentary behaviour of a passenger, here "Talking To Neighbour". This is achieved by the correct mapping of the nodes representing facial actions and movement and to the associated behaviour. These nodes themselves are characterized by the probabilities of the semantically lowest actions, whose states are the output of the above described low-level classifiers. A high probability P(T) will be assigned to the behaviour "Talking to Neighbour" if a high probability for the activity "speaking" is computed. At the same time the probability of the behaviour "Using Phone" will also rise. In order to describe the behaviour more precise additionally the activities "Laugh" and "Head Movement" are used. During a conversation a person will be looking in the same direction all the time. A consequence to little head movement is probably only little Global Motion. This description can be expanded for every activity a behaviour depends on. In order to describe the behaviour more precisely additionally the activities "Laugh" and "Head Movement" are used. During a conversation a person will be looking in the same direction all the time. A consequence to little head movement is probably only little Global Motion. This description can be expanded for every activity a behaviour depends on. In order to detect and classify a multitude of behaviours a fully meshed network as shown in figure 5 has been imple-

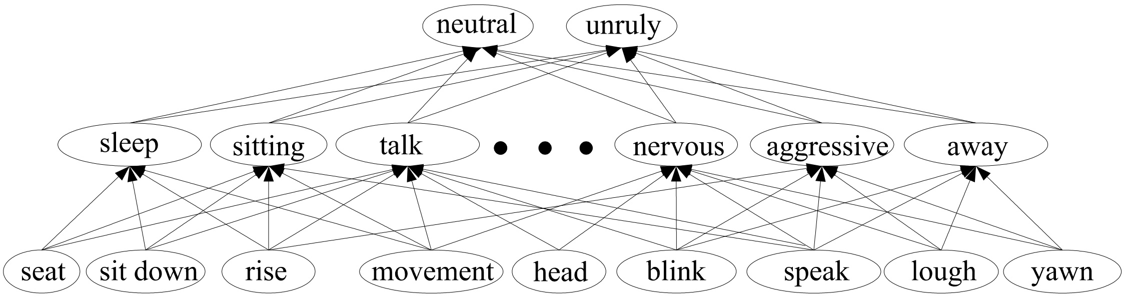

Fig. 3. Examplary Bayesian Network

mented. The Network classifies the passenger's momentary behaviour as threat (unruly) or no threat (neutral). Independencies between behaviours and activities have not been taken into account manually, as the BN is able to compute them during training.

6. RESULTS AND CONCLUSION

In this paper we introduced an approach towards fully automated behaviour detection in public transportation vehicles. It is assumed that behaviours can be segmented into low-level activities, which can be detected in real-time. To prevent high error rates, the output of several weak classifiers is fused in a second entity, a prior trained Bayesian Network. The implemented approach has been trained with 200 randomly chosen samples taken out of an artificial behaviour database. Reclassification of the training material resulted in an average error rate of 2.1%. Testing the network with 50 training disjunctive samples resulted in an error rate of 11.3%. While these seem promising results, the error rate is not acceptable for a real life application. Performance may be enhanced by creating a larger representative behaviour database, so that behaviours are described more accurate. A basic problem remains, that in some cases different behaviours can be described by the same observations, for example a person talking to her neighbour or being on the phone using a hands free set. In such cases it seems reasonable to introduce more low-level features in order to differentiate between similar behaviours.

In the future, more complete real-world data has to be collected in order to grant a more complete definition table of possible behaviours. A boost in classification performance is expected by a stronger involvement of the time component, as the actually obeyed frequency representation contains only limited information regarding this aspect. The use of Dynamic Bayesian Networks DBNs or Time Delayed Neural Networks TDNNs is considered, which internally store previous inputs and computed results for the actual output.

7. ACKNOWLEDGEMENT

This work has been partially funded by the European Union within the SAFEE project(Security of Aircraft in the Future European Environment) of the 6th FP.

8. REFERENCES

[1] 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.

[2] 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] F. Wallhoff, M. Zobl, G. Rigoll, and I. Potucek, "Face tracking in meeting room scenarios using omnidirectional views," *Proceedings Intern. Conference on Pattern Recognition (ICPR)*, vol. 4, pp. 933–936, Aug. 2004.

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

[6] E. Charniak, "Bayesian networks without tears: making bayesian networks more accessible to the probabilistically unsophisticated," *AI Magazine*, vol. 12, no. 4, pp. 50–63, 1991.

[7] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine belief network architecture," in *Proceedings IEEE Intern. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2004, vol. 1, pp. 577–580.