VIDEO BASED ONLINE BEHAVIOR DETECTION USING PROBABILISTIC MULTI STREAM FUSION

Dejan Arsić,Frank Wallhoff, Björn Schuller, 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 robustly 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

Nowadays video surveillance is used for analysing events after they actually happened, e.g. to recognize people after an accident. Monitoring the resulting video stream online proves to be a cost intensive task, as supplementary to the technical equipment a large amount of human resources is required. Consequently it seems reasonable to automate video analysis and support security staff in surveillance. 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 only with the help of video material. At the moment audio is not considered, as only microphones attached directly to the used cameras are used. In order to be able to allocate audio to persons, the use of microphone arrays is planned in future work. An unevitable requirement is the implemented systems real time ability (25fps) in order to be able to react in time. This need is accomplished by partitioning a complex behavior into several independent activities, in order to apply so called low-level features (LLF) to detect activities on a lower semantic level. By these means an additional advantage arises, as the posssibility of a description with a few meaningful features is obtained. We will further introduce a low level representation of aggressive or nervous behaviours 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. Due to the simple structure of the applied low level classifiers accuracy ist still regarded to as unreliable. In order to boost detection rates the single results are fused by the use of multi-stream fusion to drastically increase robustness.

2. IMAGE ACQUISITION

Most public transportation systems offer only very limited space for the necessary video and processing units. Considering this constraint and the required data for the specific scenario both positioning and technical specification have to be chosen carefully.
High resolution cameras providing uncompressed video are the optimum for this task, but are way too expensive in procurement of equipment and data handling. Taking this into account full PAL resolution of 720×576 pixels seems to be a reliable compromise. Image material is still needed uncompressed, in order to suppress additional noise to grant undisturbed difference image computation.
Unconvenient illumination of objects, caused by external light sources or shadows, is avoided by the combination of near infrared (NIR) filters combined with infrared lamps. This enhances the reliability of person- and face detection systems and enables image processing in the night. Figure 1 shows the output of such a camera. It illustrates a possible field of view in an Airbus cabin. Multiple seats are surveilled simultaniously, which is done to keep the amount of needed on-board video devices and hence the required processing capacity very low. As consequence optimal camera placement is crucial, respecting also the freedom of movement of the crew and the passengers. Two rows with two

From: ICIP 2005, Genova, IEEE

Fig. 1. Exemplary segmented field of view in an Airbus

seats each are monitored by cameras mounted in the bins over the passengers seats. As can be seen from the image there is a tradeoff between the camera position, the resulting size of a passengers' face and occlusion of persons sitting in the second row.

3. PERSON LOCALIZATION AND TRACKING

Preprocessing prior to the analysis of a passengers behavior is determining the exact position of the person. For this purpose a view independent face detection based on Neural Networks presented by Rowley in [1] has been implemented. As in the first step only sitting persons are surveilled, we can easily extend the facial area to the whole upper body region by an empirically determined geometry, as is illustrated in figure 1. Another advantage is, that the possible facial position of each person is limited by the geometry of the chairs in the cabin, so that the field of view can be segmented according to the seats arrangement. [1] is used.

While the implemented system procures reliable hypotheses of possible face positions and also measures the gaze, high computional effort confines real time processing. Each frame is sampled by a sliding window at different scales, which results in a processing time far higher than $0.04s$ per frame. Nevertheless real time speed is indespensible, and can be accomplished by extending the current implementation with an effective algorithm and maintaining the accuracy of the original Neural Network. This so called Condensation Algorithm [2], uses the hypotheses of an initial detection for object tracking. Out of this set N particles are chosen randomly. If the number of detections is lower than the desired amount N, particles are simply doubled. In a subsequent processing step these particles are shifted and rescaled according to a prior determined dynamical model. This is adopted to the special requirements of the tracking scenario. Experience has shown, that seated passengers move their heads only slightly and therefore small dynmic drifts in scale and position are sufficient. Thereafter the particles successively undergo a second, random fusion and are tested by the neural net. Particles with a low resulting probability are dissmissed, whereas highly propable particles are kept. Speed up in compution results from reducing the windows to process from $N = 50.000$ to $N = 100$ which intrinsically is enough to track multiple persons in one frame, within the desired scenario. Prediction and testing are continuously repeated for successive frames in real time. If no particles remain, a reinitialization is required. A similar system for omni directional views has been implemented and studied in more detail in [3].

Additional reliability can be achieved by the already introduced segmentation of the image. Thereby system is enabeled 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

Before we can implement a detection system, we have to define the behaviours of interest, such as nervousness and aggressiveness. Let us act on the assumption, that these behaviours can be characterized by the observation of several low-level activities. Inevitable is the careful selection of these, which have the ability to represent in their sum a more complex behaviour. By virtue of the actual restriction to analyse only seated passengers' behaviours in airplanes or trains, the observable actions are performed with the upper part of the body and the face. Therefore single observations are chosen, respectively lip movement (yawning, speaking, laughing), eye movement (blinking, line of view) 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 state anything on the actual behaviour. Taking the time component into account, we are able to design a description using an action's frequency of occuranc. Movement in contrast is represented by the average intensity. For instance 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 favoring 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. Likewise complex classifiers had been dismissed, and weak but fast classifiers have been prefered. Global motion features for instance can easily be computed using difference images [4]. Head movement needs not to be computed

Activity	Seat	Rise	Sit Down	Movement	Head	Blink	Yawn	Lough	Speak
Talk	15	0	1	3	3	25	0	0	10
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. Part of the created behavior database. Each activity is represented by its frequency

separately, as we compute the faces position in every single frame. Eye movement, such as blinking, can also be calculated with difference images, as a closed eye is brighter than an open eye. So just changes of the actual state have to be detected, applying a decision stump with learned values. Combined with NIR illumination and the use of the red eye effect, also the pupil can be tracked, and thereafter the line of sight can be estimated [5]. The detection of Lip movement is far more complex and is performed applying Support Vector Machines (SVM) [6].

5. DATABASE

Due to the lack of a real world database for training and testing purposes a large amount of data had to be collected. In the first place a large low level feature database, containing 10,000 single images of yawning, talking and laughing performed by 15 subjects, has been built up. Furthermore 250 blinks and 10 minutes of head movement, sitting down and uprising actions have been filmed.

The creation of a video database containing complex behaviors is very time consuming, because a large set of scenarios has to be developed, filmed and eventually annotated. For this reason only ten scenarios, such as hijacks, unruly passengers and fear of flying were filmed. These both have been analysed by experts in order to recognize such features that passengers tend to show in dangerous situations. With the help of their results frequency based representations of ten different behaviours have been determined. As every person showed differences in the same behaviors, each of the behaviors is modeled in 25 different ways, resulting in 250 representations, which are illustrated in table 1.

6. MULTI STREAM FUSION USING BAYESIAN NETWORKS

The actual classification of behaviours is performed by fusion of single classifier outputs. A rule based approach for this task may be the most simple solution, but will not take all possible relationships between activities into account. Moreover all determined probabilistic relationships between patterns contain a degree of uncertainty, as these 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.

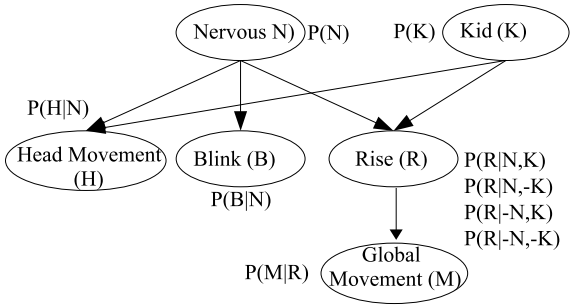

Fig. 2. Representation of "Talking To Neighbour"

Bayesian Networks (BN) [7] provide the opportunity of integration of incomplete and uncertain information in a hybrid architecture [8]. Due to the alluded benefits BNs enjoy growing popularity in knowledge modeling 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 2. A BN can be completely described in structure and conditional probabilities by its joint probability distribution. Let I denote the total of random variables, and the distribution can be calculated as:

$$P(X_1, \ldots, X_I) = \prod_{i=1}^{I} P\left(X_i | \text{parents}(X_i)\right) \qquad (1)$$

The used BN in this work is enhanced by the capability to handle soft evidences. Figure 2 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 nodes in the BN resemble the classification result of the momentary behaviour of a passenger, here "Nervous". 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 semanti-

cally lowest actions, whose states are the output of the above described low-level classifiers. High probability $P(N)$ will be obtained for the behaviour "Nervous" if high probabilities for the activities "Head Movement" and "Rise" are fed into the BN as evidences. At the same time the probability of the event "Kid" will also rise. In order to describe these more precisely the activity "Blink" is used. As the nervous person is rising and sitting down quite often the frequency of "Global Movement" is going to rise additionally. This description can be expanded for every activity a behavior depends on. In order to describe these more precisely additionally the activities "Laugh" and "Speak" may be used. A nervous person will most likely not talk or smile a lot, but be more quiet. This representation can be expanded for every activity a behaviour depends on.
In order to detect and classify a multitude of behaviors each is represented by one or more such BNs. In a second step all these networks are meshed, in order to distinguish neutral and unruly behaviour[9]. Independencies between behaviours and activities have not been taken into account manually, as the BN is able to compute them during training.

7. 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 average error rate of 11% for the classification of all 10 scenarios. Training the network to distinguish between neutral and unruly activities resulted in 8% error rate on the testset. 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 accurately. 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.
A boost in classification performance is expected by involvement of the time component, as the actually obeyed frequency representation contains only limited information regarding this aspect. In future research we consider dynamic modeling and classification by Dynamic Bayesian Networks or Time Delayed Neural Networks. These methods provide the ability to take previous activities and results into account.

8. 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.

9. 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] A. Kapoor and R. Picard, "Real-time, fully automatic upper facial feature tracking," in *Proceedings from 5th International Conference on Automatic Face and Gesture Recognition*, Cambridge, May 2002.

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

[7] 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.

[8] 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.

[9] D. Arsic, F. Wallhoff, B. Schuller, and G. Rigoll, "Vision-based online multi-stream behavior detection applying bayesian networks," in *Proceedings 6th International Conference on Multimedia and Expo ICME 2005, Amsterdam*, July 2005.