Bayesian network based multi stream fusion for automated online video surveillance

Dejan Arsic, Frank Wallhoff, Björn Schuller, Gerhard Rigoll

Angaben zur Veröffentlichung / Publication details:

Arsic, Dejan, Frank Wallhoff, Björn Schuller, and Gerhard Rigoll. 2005. "Bayesian network based multi stream fusion for automated online video surveillance." In *EUROCON 2005 - The International Conference on "Computer as a Tool," 21-24 Nov. 2005, Belgrade, Serbia*, edited by Ljiljana Milić and Đorđe Paunović, 995–98. Piscataway, NJ: IEEE. https://doi.org/10.1109/EURCON.2005.1630115.

Bayesian Network Based Multi Stream Fusion for Automated Online Video Surveillance

Dejan Arsić, Frank Wallhoff, Björn Schuller, Gerhard Rigoll, *Member, IEEE*

Abstract — Video Surveillance is an omnipresent topic when it comes to enhancing security in public places and transportation systems. Fully automated behavior detection systems are desirable when it comes to cutting costs for analysing video and audio streams online. These will initiate an alarm signal autonomously if a possibly dangerous situation is detected.

The particular investigated scenario is monitoring passengers' behaviors in aircrafts. In order to work robustly in unconstrained environments many subsystems have to be developed. Though in the last years reliable approaches for required systems have been brought up, there exists a gap between reliability and computational effort. Hence a Low Level Activity representation of behaviors will be presented, which can be detected with so called weak classifiers in real time. These outputs will be interpreted by a highly sophisticated probabilistic Bayesian Network.

Keywords **— Bayesian Networks, Multi Stream Fusion, Video Surveillance, Low Level Features**

I. Introduction

Nowadays video surveillance is used for analysing events after they actually happened, e.g. to recognise 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 to support security staff in surveillance. A possible application may be automatic observation of passenger compartments in aircrafts. The goal is the detection of e.g. aggressive persons, passengers illicitly using electronic devices or just ill people only with the help of video material. At the moment audio is not considered, as only microphones attached directly to the installed cameras are used. In order to be able to allocate audio to persons, the use of microphone arrays is planned in future work.

An inevitable requirement is the implemented system's real time ability (25fps) in order to be able to react in time. This need is accomplished by partitioning a complex behaviour into several independent activities, so called Predefined Indicators (PDI), in order to detect activities on a lower semantic level with weak classifiers. By these means an additional advantage arises, as the possibility 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 is 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.

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

D. Arsič is with the Institute of Human Machine Communication, Technical University Munich, Germany (phone: +49-89-28928551; fax: +49-89-28928535; e-mail: arsic@tum.de)

F. Wallhoff is with the Institute of Human Machine Communication, Technical University Munich, Germany (e-mail: wallhoff@ei.tum.de)

B. Schuller is with the Institute of Human Machine Communication, Technical University Munich, Germany (e-mail: schuller@tum.de)

G. Rigoll is with the Institute of Human Machine Communication, Technical University Munich, Germany (e-mail: rigoll@tum.de)

II. Image Acquisition

Aircrafts provide only limited space for cameras and computation units. Additionally power supply is restricted, so that the amount of technical equipment has to be kept as low as possible. Considering these constraints and the required data both position and technical specifications of the employed cameras have to be chosen carefully.

Multiple seats are observed at the same time by positioning cameras above the passengers' heads in the bin. Fig.1 shows the resulting Field Of View (FOV) and arising problems. Passengers in the 2nd row may appear very small and also be occluded. Facial details could be recognized by installing high resolution cameras, which will create a huge amount of data and increase computation time. Full PAL (720 x 576) resolution seems to be a compromise. Artifacts arising from interlaced material are avoided by the use of progressive scan cameras. This allows undisturbed image processing.

Fig. 1. Exemplary view in an Airbus cabin

External lighting or shadows often cause inconvenient illumination of objects. This is avoided by using NIR cameras and infrared illumination. Especially the object recognition task is drastically enhanced by this method.

III. Image Preprocessing

Significant importance is assigned to image preprocessing, in order to ease pattern recognition. Besides standard algorithms like histogram equalization and deinterlacing, if necessary, an algorithm for smoothing head movements has to be implemented. Usually the faces position cannot be determined exactly, so that differences within the face may appear, although the facial expression has not changed in contrast to the position, see Figure 2a. Subsequently to the face detection in two following frames block matching is performed in the area surrounding the face, illustrated as the area between the two rectangles in Fig. 2b. By sparing out the inner box, reduction of movement caused by facial expressions is avoided.

By moving the facial area of the 2nd frame in x and y

direction and computing the difference image for each possible displacement we obtain the coordinates of the face for the lowest head movement. This is indicated by the smallest difference.

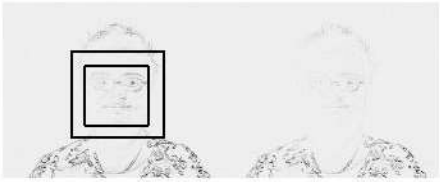

Fig. 2: Difference image before (a) and after (b) matching

IV. Low-Level-Feature Representation Of Complex Behaviors

A crucial question within the behavior detection task is the selection of suitable features. We assume, that a single feature cannot describe a complex behavior entirely. Therefore we defined so called Predefined Indicators for a detailed representation [1]. For instance a passenger planning to hijack the aircraft tends to be nervous, being anxious and observing the cabin. These PDI in turn may be characterized by observing Low Level Activities (LLA), as shown in Fig.3.

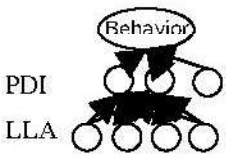

Figure. 3: Segmentation Of Behaviours In PDI And LLA

By virtue of the actual restriction of analysing only seated passenger's 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 actions frequency of occurrences. 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 favouring real time performance. The obliged initially high error rate is compensated within a multi stream fusion described later.

V. Databases

Testing and training our system for the particular application area required a large database containing both LLAs and complex behaviors.

In the first case a large LLA database has been created. It contains 10.000 images of yawning, laughing and talking faces performed by 15 persons. Furthermore 250 blinks and 10 minutes of head movement, sitting down and uprising actions have been filmed.

Figure 4: Samples taken from the LLA database

The creation of a video database containing complex behaviours is very time consuming, because a large set of scenarios has to be developed, filmed and eventually annotated. For these reasons only ten scenarios, such as hijacks, unruly passengers and fear of flying were filmed. These have been analysed both by experts in order to recognize features 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 behaviours, each of the behaviours is modelled in 25 different ways, resulting in 250 representations [1].

VI. Low-Level-Feature-Detection

Aiming towards real time capabilities of the system, it is indispensable to develop fast classifiers, at reliabilitie's cost. This will be enhanced by a subsequent intelligent fusion of the weak classifier output. Some of the implemented systems will be presented in the following paragraphs:

A. Global Motion

Global motion describes movements of persons, in our particular case sitting down, uprising, the seat state and the intensity of a passenger's movement. All these features have in common, that they can be computed fast and reliable by applying difference images:

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

In the first place it is possible to detect if a person is sitting in a seat by background subtraction of the actual grabbed frame and a stored image. If the sum of differences

$$sum_d = \sum_{x,y} d(x,y,t) \qquad (2)$$

is very small it's likely, that the seat is empty, otherwise it's probably occupied. Comparing two subsequent seat states provides information on the passenger's activities. Sitting down is characterized by a taken seat status after the seat being free, whereas a change from taken seat to empty seat indicates uprising.

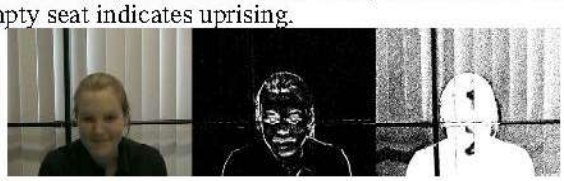

Figure 5: Original image (a), difference image (b) of subsequent frames, background differencing (c)

As Fig. 5 illustrates on the right hand side changing lighting conditions complicate a robust detection. Therefore an adaptive background model has to be created.

In addition the difference image of two subsequent frames is computed, which provides the possibility to compute the mean [2] of the movement within the image:

$$m_x = \frac{\sum_{x,y} x * d(x,y,t)}{\sum_{x,y} d(x,y,t)}, m_y = \frac{\sum_{x,y} y * d(x,y,t)}{\sum_{x,y} d(x,y,t)} \quad (3)$$

for x and y direction. More meaningful is the consequently computation of the movement's variance, which is interpreted as intensity of the movement:

$$\sigma_x(t) = \frac{\sum_{x,y} d(x,y,t) | x - m_x(t) |}{\sum_{x,y} d(x,y,t)}, \sigma_y(t) = \frac{\sum_{x,y} d(x,y,t) | x - m_y(t) |}{\sum_{x,y} d(x,y,t)} \quad (4)$$

Large variances indicate large changes within two following frames, whereas small variances are caused by little movement

B.Face Movement

As face tracking is used for person detection anyway, the head's movement can be modeled by determination of the actual position. Therefore a Multi Layer Perceptron (MLP) based face detection approach as presented by Rowley in [3] has been implemented. Wallhoff describes a similar system for "omnidirectional" views in [4]. Due to the high computation time for a full search over the image real time tracking is not possible. To circumvent this problem a system predicting the face's position and size is required. The chosen method, introduced by Isard and Blake, is called Condensation algorithm [5].

After the 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 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 = 200. Particles with a low probability are discarded. Prediction and testing are continuously repeated for succeeding frames in real time. A measurement for the intensity of the Face Movement is the average distance between 2 frames.

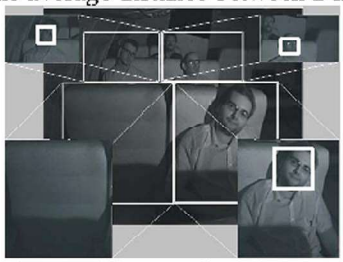

Figure 6: Face tracking results after segmentation

C.Lip Movement

Very characteristic for facial emotions and face expression are lip movements. Classification is rather challenging task as a wide range of expressions is possible and the mouth itself is a highly deformable object. Taking this into account we restricted the task to the detection of laughing, yawning and speaking. Focusing on a fast detection process computationally intense approaches as Active Shape Models and analysis of dynamic patterns have been dismissed in favor of real time processing on static images. Promising results have been provided by Support Vector Machines (SVM) [6] with difference images or edge detected images of mouths as vector input.

SVMs are a very popular approach for 2 class problems. A high dimensional vector space is spanned by positive and negative training material. Within the learning phase a non linear hyperplane is determined, which separates both classes optimally. During the classification process it is decided on which side of the plane the vector is situated.

This way three systems have been trained with images taken from our LLA database in order to distinguish the desired classes. Both difference images and edges as feature (36x18 pixel), as shown in Figure 7, were used. They achieved detection rates of up to 75% and false positive rates of 30%. These results may be accurate enough for our system as every frame is analyzed and additional features are taken into account for the correct classification of a PDI.

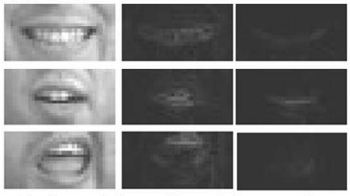

Figure 7: Laughing, Speaking and yawning

D.Eye Blinking

Blinking can be characterized by detecting the movement of the eyelid. First the eyes will close and therefore the area of the eye will become lighter, as the iris and pupil are usually darker than the skin color. In the second step the eyes will open and as result create a darker pattern.

These changes can be easily detected by applying a threshold on the sum of differences on the slightly modified difference image of the eyes:

$$d = \sum_{x,y} \frac{I_2(x,y) - I_1(x,y) + 255}{2} \quad (5)$$

Small values indicate a closing eye, whereas larger values indicate an opening one. Fig. 8 visualizes this approaches' functionality.

Experiments have shown detection rates of 95,7% on our database, although the position of the eye has only been estimated empirically within the face.

Figure 8: Blink detection using difference images

VII.Multi-Stream-Fusion

As already mentioned, the actual classification of behaviours is performed by the fusion of the single classifier outputs. A rule based approach for this task may be the simplest solution, but will not take all possible relationships between features 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.

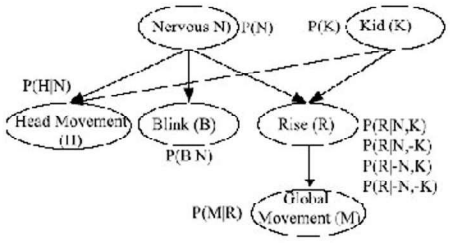

Figure 9: Exemplary BN for the LLA Nervousness

Bayesian Networks [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 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 9. 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:

$$P(X_1, \ldots, X_N) = \prod_{N=1} P(X_i|\text{parents}(X_i)) \quad (5)$$

Figure 9 illustrates an example for a possible implementation of a BN structure for a multi-stream fusion system, whose topology is derived from expert knowledge. The root node in the BN resembles the classification of the momentary behaviour of a passenger, here "Nervous Passenger". 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(N) will be assigned to the behaviour "Nervous Passenger" if a high probability for the activities "Blinking" and "Rising" is computed. At the same time the probability of the event "Whining" will also rise. In order to describe the behaviour more precise additionally the activity "Blinking" is used. As the nervous person is rising and sitting down quite often the frequency of "global motion" is going to rise additionally. 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 "Speak" may be used. A nervous person will most likely not talk or laugh a lot, but be more quiet. This description can be expanded for every activity the behaviour depends on.

In order to detect and classify a multitude of behaviours each behaviour is represented by one or more such networks. In a second step all these networks are meshed, in order to distinguish neutral and unruly behaviour. This creates a in fact fully connected network. Independencies between behaviours and activities have not been taken into account manually, as the BN is able to compute them during training.

VIII. Conclusion And Outlook

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 (DBN) [9] or Time Delayed Neural Networks (TDNN) is considered, which internally store previous inputs and computed results for the actual output.

References

[1] D. Arsić, F. Wallhoff, B. Schuller, G. Rigoll: "Vision-Based Online Multi-Stream Behavior Detection Applying Bayesian Networks", in *Proceedings 6th International Conference on Multimedia and Expo ICME 2005*, Amsterdam, The Netherlands, 06.-08.07.2005.

[2] M. Zobl, F. Wallhoff, G. Rigoll "Action Recognition in meeting scenarios using global motion features", in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*(PETS-ICSV), Graz Austria, March 2003, pp 32-36.

[3] H. Rowley, S. Baluja "Neural Network Based Face Detection", in IEEE *Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp 23-38, 1998.

[4] F.Wallhoff, M.Zobl, and G. Rigoll "Face Tracking in meeting scenarios using omnidirectional views" in *Proceedings International Conference on Pattern Recognition (ICPR) 2004, vol. 4, pp 933-936, Aug. 2004*

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

[6] N. Cristianni "Support Vector and Kernel Machines" in *ICML Tutorials* 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* (ICASSP9, May 2004, vol.1, pp 577-580

[9] Zoubin Ghahramani "Learning Dynamic Bayesian Networks", in *Lecture Notes in Computer Science*, vol. 1387, pp. 168-197, 1998