Fusion of Acoustic and Optical Sensor Data for Automatic Fight Detection in Urban Environments

Maria Andersson[1], Stavros Ntalampiras[2], Todor Ganchev[2], Joakim Rydell[1], Jörgen Ahlberg[1], Nikos Fakotakis[2]

[1]Division of Information Systems, FOI, Swedish Defence Research Agency, Linköping, Sweden
[2]Department of Electrical and Computer Engineering, University of Patras, Patras, Greece

maria.andersson@foi.se, sntalampiras@upatras.gr, tganchev@ieee.org, joakim.rydell@foi.se, jorgen.ahlberg@foi.se, fakotaki@upatras.gr

Abstract – We propose a two-stage method for detection of abnormal behaviours, such as aggression and fights in urban environment, which is applicable to operator support in surveillance applications. The proposed method is based on fusion of evidence from audio and optical sensors. In the first stage, a number of modality-specific detectors perform recognition of low-level events. Their outputs act as input to the second stage, which performs fusion and disambiguation of the first-stage detections. Experimental evaluation on scenes from the outdoor part of the PROMETHEUS database demonstrated the practical viability of the proposed approach. We report a fight detection rate of 81% when both audio and optical information are used. Reduced performance is observed when evidence from audio data is excluded from the fusion process. Finally, in the case when only evidence from one camera is used for detecting the fights, the recognition performance is poor.

Keywords: Abnormal behaviour detection, multiple sensor fusion, acoustic data, visual data, thermal imaging data, Hidden Markov Model

1 Introduction

1.1 Background

Today, applications such as area access authorization, home arrest, medical care, anti-terror surveillance, and forensic analysis are widespread. These applications share the need of reliable autonomous surveillance technology, enabling automated alerts, and thus support control and law enforcement mechanisms. Furthermore, a desirable functionality for such technology is multi-person behaviour analysis and short-term intention prediction, early reporting of danger and the support of alert indicators that facilitate the prevention of danger, and the management of crisis events, etc. Analysis of crowd behaviour as well as modelling of crowds is also an area of increasing interest within the safety, security and computer vision research communities. The analysis of crowds for prediction and detection of events is a complex issue among other things because people move close to each other, occlusion can hide important actions and the possibility to successfully track individuals is reduced. In addition, it is worth mentioning that for different types of crowd there are different criteria for defining normal behaviours, and a certain normal behaviour in one type of crowd should cause an alarm in another type of crowd.

In public security applications, privacy and ethical considerations are inhibiting the use of biometric processes, such as the automated voice recognition or face identification, as well as the use of automatic speech recognition technology. Respecting these restrictions, we consider only such technology for autonomous surveillance and automatic scene analysis which does not employ person identification or speech content analysis.

1.2 Scope and objectives

The present work aims at automatic detection of abnormal behaviours, such as aggression and fights, which pose danger to the average citizen, and which are of significant importance for public security applications. This effort is part of the FP7 PROMETHEUS project [1], which aims at the creation of a framework for monitoring and interpretation of human behaviours in unrestricted indoor and outdoor environments. The project targets the development of a probabilistic platform for processing and fusion of data streams from heterogeneous sets of sensors, such as microphone arrays, overview and high-definition video cameras, 3D cameras, thermal imaging sensors etc. The long-term objective is the development of technology that is applicable in real world applications, able to operate on data captured in uncontrolled environments, such as public spaces.

Specifically, in connection with this objective, in the present work we investigate a fusion method that combines evidence from audio and optical sensors for achieving robust detection of events of aggression and fights. We elaborate on real-world data captured in an outdoor public space. The proposed method for detection of abnormal behaviours is based on a two-stage processing of the input data. At the first stage a number of individual, modality-specific event detectors, referred to as low-level event detectors, perform recognition of a set

of predefined events-of-interest on the incoming data streams. The output of the detectors consists of sequences of event labels and the recognition confidence for each detected event. These sequences are fed to the second stage, which combines and disambiguates the output of the individual detectors. As the individual detectors operate on a set of sensors with complementary perception capabilities, a more accurate and reliable detection is achieved when compared to the performance obtained for each of the individual detectors alone.

The proposed technology is expected to facilitate the creation of reliable autonomous surveillance systems, which incorporate intelligent automatic alert mechanisms. These mechanisms can offer the opportunity for reducing the workload of the personnel in security control centres, and enable appropriate actions to be taken for minimizing the risk of injuries or property damage.

The present work bears some similarity with previous work [2], where a Hidden Markov Model (HMM) is used for detection of abnormal crowd behaviour. The HMM is based on data from visual and thermal infrared cameras. In the present work, we elaborate further on this approach and evaluate the proposed two-stage method on the PROMETHEUS database. Furthermore, in the present work we rely on an extended set of sensors, which includes overview video cameras, thermal imaging cameras and a microphone array. The introduction of the microphone array is expected to improve the situational awareness about the scene of interest, compared to having only optical sensors. For instance, in outdoor environment, occlusions or fog could prevent the proper event detection through video and thermal imaging cameras, but the microphone array will still provide evidence for the acoustic events that are characteristic for fights or aggression.

1.3 Outline

The remainder of this paper is organized as follows: In Section 2 we briefly overview related work and define the innovations of the proposed method. In Section 3, we offer an outline of the low-level event detectors for each of the individual modalities. In Section 4, we present a fusion method, which combines and disambiguates the detections provided by the low-level event detectors and performs the classification to normal/abnormal behaviour. Section 5 outlines the PROMETHEUS database and experimental setup used in the experimental validation of the proposed approach. Section 6 reports the experimental results for the low-level event detections and for their fusion. Finally, Section 7 concludes with a summary of achievements.

2 Related work

The possibilities of acoustic surveillance for hazardous situations have recently received quite a lot of attention by the signal processing community. The basic advantages are the low computational needs and its independence of illumination conditions and occlusions. A small number of related systems that operate under urban environments are mentioned in the literature. In particular, a system for gunshot and scream detection and localization in a public square is presented in [3]. The authors use forty-nine features as an input to a hybrid filter/wrapper selection method. Two parallel Gaussian Mixture Models (GMMs) were built to represent the selected feature set for discriminating screams and gunshots (respectively) from noise. Data were extracted from movie sound tracks, internet repositories, and recorded from people shouting at a microphone, while the noise samples were captured in a public square of Milan.

Detection of audio events in public transport vehicles was investigated in [4]. Four microphones recorded four different scenarios, including fight scenes, a violent robbery scene, and scenes of bag or mobile snatching. GMMs and Support Vector Machines (SVMs) based classifiers, fed by feature set composed of the first 12 Mel-Frequency Cepstral Coefficients (MFCCs) and their energy, derivatives and accelerations were used.

Related studies [5-8] deal with the modelling of crowds for different applications, where also effects of crowd psychology are considered, i.e. how people are expected to react in cases of emergency, how they move within a building or outdoors to escape in occasion of an emergency event, etc.

Certain combination of acoustic and visual sensors has been investigated for security and safety issues such as the detection of abnormal and aggressive behaviour at train and underground stations [9-10]. Detailed body-pose estimation and analysis of movements of body parts were used in [10], in contrast to the work presented here, where we do not consider such level of detail in the optical image, but rely on abstractions such as representing humans with optical flows and foreground regions.

3 Low-level event detection

The level of activity of a crowd provides information that can be used to detect abnormal crowd behaviour. Normal behaviour often corresponds to calm movements and calm speech, i.e. people standing or moving relatively slowly through the scene, without making excessive gestures. An abnormal event, however, is likely to be accompanied by more rapid movements, and shouting. We define seven crowd observations: normal activity (calm motion), intensive activities by a few, intensive activities by several, low sound, high sound, small crowd and large crowd. The observations are fed to the crowd behaviour analysis, which is based on HMM.

3.1 Detection of acoustic events

The processing of acoustic signals considered here is motivated by the monitoring system described in [11]. The block diagram of the proposed detector of acoustic events is depicted in Figure 1. Specifically, after signal pre-processing, which removes the signal offset with respect to zero mean value and smoothes any possible misalignments, the feature extraction stage takes place.

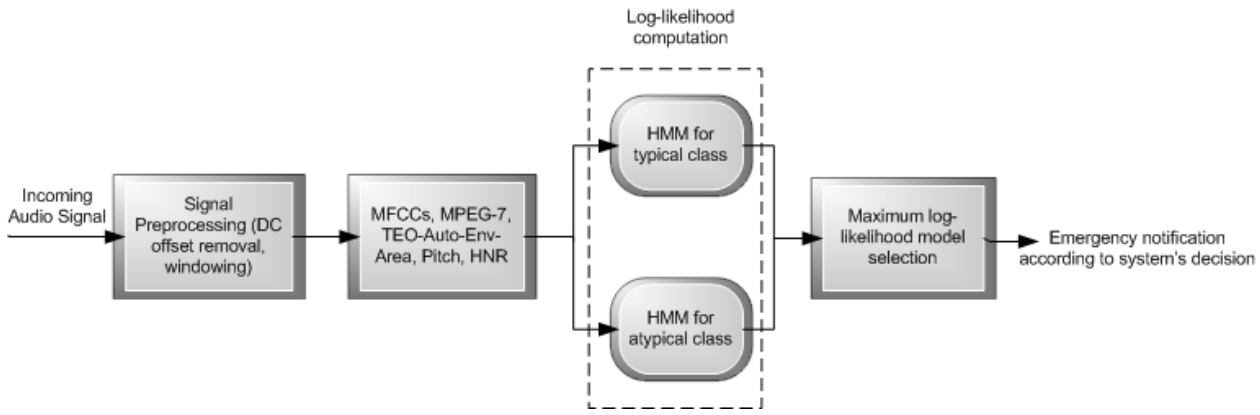

Figure 1. Block diagram of the audio event detection and classification component.

The computed audio features are fed to the pre-defined statistical models and a degree of resemblance between each model and the feature values is computed using the Viterbi algorithm [12]. This degree of resemblance is in the form of log-likelihood, while the final decision is made by determining the maximum log-likelihood.

Due to the specific nature of the audio events of interest – an abnormal vocalic reaction such as these accompanying the events of panic, fights etc – we use few additional audio descriptors, which contribute to capturing better the characteristics of abnormal vocalic reactions. Thus, to this end we rely on the following feature set:

1. *Mel-Frequency Cepstral Coefficients* (*MFCCs*): comprise a Mel-scaled projection of the log spectrum, followed by a parameter decorrelation step. The MFCC time-derivatives are appended to the final vector, as well.
2. *MPEG-7 Audio Protocol Low Level Descriptors* [13] are currently the state-of-the-art audio parameterization in generalized sound recognition. In the present work we use: a) Audio waveform min and max and b) Audio spectrum flatness, and both of them are complementary to MFCCs.
3. *Teager Energy Operator (TEO) and Intonational Features*: Due to the relatively large range of abnormal vocalic events that we have to deal with, we utilize acoustic parameters that capture the variations that the airflow pattern exhibits when it comes to abnormal circumstances. Thus we utilized TEO-Autocorrelation-Envelope Area [14], pitch, pitch derivative, and harmonicity to noise ratio. The last three were computed using PRAAT [15] software which is optimized for the processing of speech signals.

The feature extraction algorithms elaborate on audio signals with 16 kHz sampling rate and 16 bit quantization. As regards the pattern recognition methods, we experimented with generative classifiers. Three types were considered: GMMs, ergodic/fully connected HMMs and left-right HMMs. The recognition of acoustic events is based on the assumption that every source emits sounds which follow consistent patterns, their so called audio signatures. We estimate and subsequently identify the patterns using statistical techniques. HMMs have the ability to capture not only static aspects of the feature sequence but also their evolution in time. They break up the sequence into a predefined number of states and try to learn the relations between them. Ergodic HMMs allow for every possible transfer between the states while in the case of left right HMMs there are no directed loops in the automation. They result in an n by n matrix where each element represents the probability of the transition between different states. Thus, the element (i, j) is the probability of moving to state j at time t+1 given state i at time t. It should be noted that each state is modelled by a GMM based on components with diagonal covariance matrices.

3.2 Detection of events from optical streams

We use optical streams to estimate the crowd size and the degree of activity from people in the crowd.

3.2.1 Crowd activity estimation

The level of crowd activity is measured by computing the optical flow in the scene (in this case in the visual image). If a person is walking quickly, running, or moving his or her arms rapidly, the magnitude of the optical flow will be larger compared to when a person is moving slowly or standing still. An average value of the magnitude of the optical flow is estimated. The optical flow from detected persons was used in [2] to detect activities, in contrast to the present work where there is no need of first detecting persons; the optical flow is instead calculated for the whole scene. The optical flow will therefore be a measure also of the crowd dynamics as a whole.

3.2.2 Crowds size estimation

The crowd size does not alone indicate normal or abnormal behaviours. The crowd size is more a representation of the crowd state that together with other features (from other sensors) can indicate normal or abnormal behaviour.

To estimate the crowd size, we use prior knowledge of the approximate number of pixels per person associated to the distance between camera and crowd. Background subtraction is done to obtain the foreground pixels, which are assumed to represent all persons in the scene. The

number of people is obtained by dividing the total amount of foreground pixels by the number of pixels per person.

What is considered to be a large crowd will differ from case to case. For example, in a small city area a large crowd may be 20-25 persons. At large sport events, large crowds are probably hundreds or thousands of persons.

We use a thermal infrared camera to measure the crowd size. This camera is advantageous since it is not affected by cast shadows. On the other hand, high environmental temperatures may cause poor contrast in the thermal camera leading to an underestimation of the crowd size.

4 High-level fusion for fight detection

The high-level fusion is performed by using a HMM [12]. In brief, the HMM consists of two stochastic processes. The underlying (hidden) process can be observed through another stochastic process that produces sequences of observations O_S. The states S represent some unobservable conditions of the system. In each state there is a certain probability of producing any observable system outputs O together with a probability indicating the likely next states. The HMM for a normal crowd is described by the following parameters:

$$\lambda_N = (A, B, \pi, S, O_S), \quad (1)$$

where A is the probability distribution of state transitions, B is the probability distribution of observations in each state and π is the probability distribution of the initial state. A, B and π can be obtained by training λ_N on relevant training data.

For behaviour recognition the likelihood L_N for an observation sequence $O_S = (O_1, O_2,\ldots,O_T)$ is calculated according to Eq. (2).

$$L_N = (P(O_S \mid \lambda_N) \quad (2)$$

The result of L_N is compared to a threshold T_N that represents expected normal crowd behaviour. If $L_N < T_N$ the crowd behaviour is likely to be abnormal. If $L_N > T_N$ the crowd behaviour is likely to be normal.

4.1 HMM for fight detection

λ_N is aimed at detecting fights, which are represented by quick movements and atypical sound events. We selected a HMM with two states (S_1 and S_2), which refer to calm motions (standing and walking) for S_1, and slightly increased activities (predominantly walking) for S_2, still belonging to normal behaviour. Both states include certain segments of unusual observations that can come from incorrect sensor detections and the fact that unusual observations may occasionally occur also in normal behaviour. We selected seven observation symbols (O_1 – O_7), which are:

- O_1: Normal activities, i.e. walking, standing,
- O_2: Increased activities, i.e. walking and more intensive movements by a few,
- O_3: Strongly intensive activities by several,
- O_4: Low sounds,
- O_5: High sounds,
- O_6: Small crowd,
- O_7: Large crowd.

Since we do not have enough recorded training data on normal behaviour in crowds we have derived training data based on the knowledge of what is associated with normal behaviour. The Expectation-Maximization (EM) algorithm was used for the training process.

5 Experimental setup

5.1 PROMETHEUS database

The multimodal multisensory PROMETHEUS database [16] was created in support of RTD activities aiming at the creation of a framework for monitoring and interpretation of human behaviours in unrestricted indoor and outdoor environments. The database consists of approximately four hours of recordings, representative for two application scenarios: smart-home and public security (airport and ATM – bankomat, surveillance). Three indoor sessions, in the smart-home and the airport scenarios, were recorded in the Greek language with average duration of a session of approximately 20 minutes. The outdoor sessions were almost entirely in the English language as spoken by non-native speakers. Two sessions with duration of approximately 30 and 50 minutes represent the ATM scenario, and an additional 76 minutes represent the public area security scenario. Except for these pre-designed scenes, where the actors improvised guided by task cards, the database contains recordings from the intersession breaks, and thus, portion of the recordings was not pre-designed.

Each recording session is comprised of multiple action scenes concatenated to a single sequence, where each action scene is implemented a number of times by different actors. In the present work, we report results on three fight scenes (in the following denoted as scene 1, scene 2 and scene 3). These scenes have a cumulative length of approximately four minutes and represent abnormal multiple-person interaction episodes, which include abnormal behaviours, as well as abnormal behaviours such as aggression, fight, a person brought down, and people helping the sufferer. The sound, video, and thermal sequences corresponding to these episodes were annotated with respect to sound event type as well as human location and action.

5.2 System setup

In the present work, we are dealing with keysound spotting. More specifically, acoustic signals which are characterized by a long duration need to be processed for the purpose of abnormal sound event detection. Thus, the instantaneous value of each feature is computed over a larger frame size than the one commonly used in content-

based recognition (namely 30ms). After several experiments and based on the highest recognition rate criterion, it was decided that all sound samples should be cut into frames of 200ms with 75% overlap. Mean value removal and variance scaling are applied on the time domain signal. The FFT size is 512.

We used Torch implementation [17] of GMM and HMM, during the experimental phase. The maximum number of *k-means* iterations for initialization was 50 while both the *EM* and *Baum-Welch* algorithms had an upper limit of 25 iterations with a threshold of 0.001 between subsequent iterations [12]. Based on the performance that each technique demonstrated, we utilized left-right HMMs with 4 states and 64 Gaussian functions for each state. We randomly chose 50% of the data from the general security scenario for training the models. The testing set is consisted of three selected scenes.

Data from the visual and thermal infrared cameras are put into an analysis framework. In the framework the camera images are calibrated and synchronized. The calibration is done by using the Matlab calibration tool. When HMM calculations are performed that also includes acoustic input data, the acoustic data are introduced in the framework in the form of sound decisions.

5.3 Performance measure

The accuracy of the proposed event detection method is reported in terms of correct detection rate, D_T, in percentages:

$$D_T = 100 \times \frac{T_P + F_N}{T_N + F_N + T_P + F_P} [\%] \quad (3)$$

where T_P and T_N stand for true positives and true negatives, and F_N and F_P are false negatives and false positives, computed on per frame basis. In the present work, we weight equally the misclassification of the normal and the abnormal events.

6 Experimental results

We used four sensors for observing the three fights. Figure 2 shows the views from the three optical cameras. The upper visual camera to the left is referred to as camera 1 in the following discussion and the upper visual camera to the right is referred to as camera 2. The thermal infrared camera below has the same view as camera 1. In Figure 2, the same fight can be seen from the different views, representing the same time. The microphone array is placed close to the desk in the scene.

6.1 Results from the low-level detection

6.1.1 Acoustic information

The method that was chosen for the audio processing is essentially detection by classification. We experimented with several window sizes which varied from 0.03 seconds to 1 second with 0.01 seconds skip step. The best performance was achieved while using a window of 0.5 seconds. Subsequent decisions which have the same label are merged into one with the same start time and cumulative increase of duration. Such a smoothing scheme removes single-frame detections. An example of the sound recognizer outputs is illustrated in Figure 3.

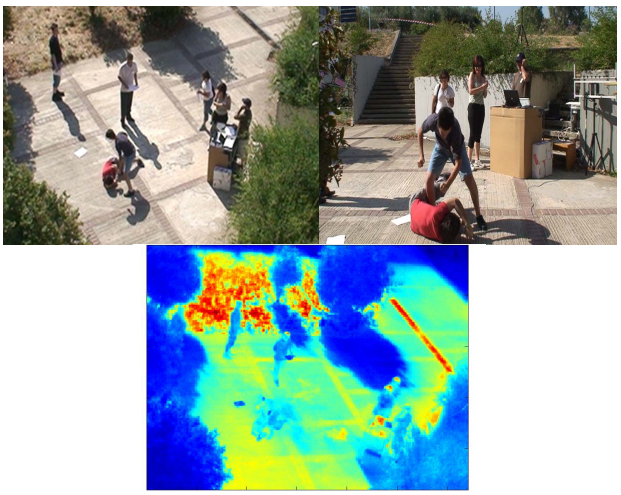

Figure 2. Views from the optical cameras. The upper visual image to the left represents the view from camera 1 and the upper visual image to the right represents the view from camera 2. The lower image represents the thermal infrared camera.

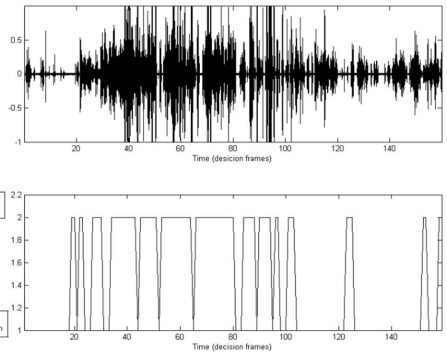

Figure 3. The sound waveform (top) and the decisions made by the sound recognizer (bottom) on the sequence denoted as scene 1, which includes an episode with a fight (see Figure 2). In the bottom plot decision value "2" corresponds to an abnormal situation, and decision value "1" corresponds to normal situations.

When the output curve increases, an abnormal situation is detected by the system. The waveform belongs to a fight scene, where the fight starts approximately at t=1.8 sec. The percentages are calculated using per window analysis and the resulted correct detection rate is 87% for the three fight scenes.

6.1.2 Optical information

Figure 4 presents optical flows from camera 1 and camera 2 for scene 2. Three levels of optical flow are defined for each camera to distinguish between different degrees of movements, according to the observation symbols O_1, O_2

and O_3 (see section 4.1). The levels are different in the two cameras. This is because they have different possibilities to detect the movements. The distance between camera 2 and the crowd is shorter than the distance between camera 1 and the crowd. Camera 2 is therefore able to detect more movements and consequently obtain higher optical flow values. The first levels (*optical flow*=30 in camera 1 and *optical flow*=400 in camera 2) represent the threshold for increased activities, compared to normal activities. The second levels represent the threshold for strongly intensive activities, compared to increased activities.

The thresholds have been derived by observing the optical flows under known conditions, i.e. when the types of events in the scene are known. 5 optical flow values are calculated per second. Each fight scene has a duration of 80 seconds which in Figure 4 corresponds to *Time* = 400. The fight starts at *Time* = 200.

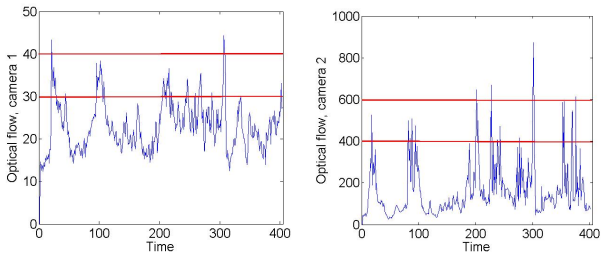

Figure 4. Optical flows for scene 2 in camera 1 (left) and camera 2 (right). See also Figure 3 for the different views of the cameras.

Figure 5 presents the crowd size estimation for scene 2, based on thermal infrared data. As above, 5 crowd size calculations are done each second. The changes in crowd size depend mostly on that people enter and/or leave the scene, or move behind each other. The changes also depend on data uncertainties, because of high environmental temperatures. In high temperatures, the contrast in thermal infrared data is occasionally poor, leading to incorrect variations in crowd size.

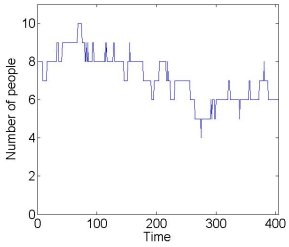

Figure 5. The variation of crowd size for scene 2. The variation is due to people entering or leaving the area, but also if they move within the area which can result in occlusion.

The fight starts at *Time* = 200. The correct number of people varies between 6 and 9. There is a reduction in crowd size at *Time* = 270. At this time the fight is ongoing and occlusion effects are present, which cause a reduction in the estimated crowd size. However, the measure can give an approximate crowd size, which is enough to see that the crowd is small.

6.2 Results for the high-level fusion

The high-level fusion, based on HMM, combines the evidence from the low-level audio and optical detectors in order to provide a more robust detection of atypical situations. Firstly, the observations from different low-level detectors are synchronized so that they correspond to the same time frame. In the following we shall consider that the system makes decision every 0.5 seconds, which is the frame size that we will consider as unit for time in the figures that follow. Each of the fight scenes has a duration of 80 seconds, which in the diagrams, presented in Figures 6 and 7, corresponds to *Time* = 160.

Figure 6 shows the log-likelihood of normal behaviour $P(O_S|\lambda_N)$ for four different scenes. The upper-left panel shows $P(O_S|\lambda_N)$ for the normal crowd behaviour. This panel is contrasted to the three fights shown in the other panels in Figure 6.

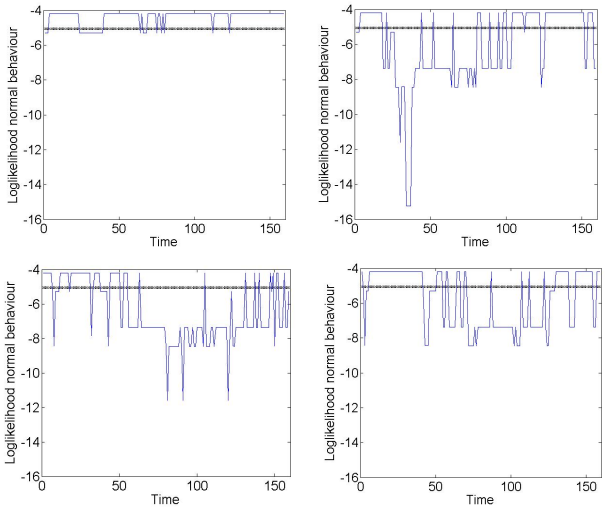

Figure 6. Fusion of audio and optical evidence. In all diagrams the plots with thin solid line show $P(O_S|\lambda_N)$ for: (i) upper-left diagram - normal crowd, (ii) upper-right diagram - fight in scene 1, (iii) lower-left diagram - fight in scene 2, (iv) lower-right diagram - fight in scene 3. The dashed line at level $P(O_S|\lambda_N)$ = -5 in all diagrams represents the threshold T_N, above which normal behaviour is reported .

The thick dashed line at level $P(O_S|\lambda_N)$ = -5 represents the threshold T_N above which normal behaviour is reported, i.e. when $P(O_S|\lambda_N) \geq T_N$. Instead, when $P(O_S|\lambda_N) < T_N$ the crowd behaviour is assumed to be abnormal. T_N has been derived by calculating $P(O_S|\lambda_N)$ based on a known normal observation sequence.

The upper-right panel shows $P(O_S|\lambda_N)$ for scene 1, that includes a fight between two people. As can be seen $P(O_S|\lambda_N) < T_N$ for the major part of the period, and well corresponds to the actual development of actions in the scene. The actual fight starts at *Time* = 75 and ends at *Time* = 100. But the fight is preceded by a quarrel that starts at *Time* = 25, when the two persons shout in a loud voice.

The fight is observed by camera 1, but the evidence from the optical flow is not so strong in the beginning of

the fight. The fight can not be observed by camera 2, since the fight takes place outside the perceptive area. The microphone array detects high volume sounds and thus contributes to early detection of the fight action, corresponding to a reduction of $P(O_S|\lambda_N)$ to a level that should give an alarm to the security operator. The fight, including the quarrel, is well detected by the HMM method.

At *Time* = 40 there is a motorcycle crossing the area, causing an increase in optical flow, sound and crowd size and leading to a strong reduction in $P(O_S|\lambda_N)$. This event also corresponds to atypical event – a motorcycle passing through the scene.

Next, the lower-left panel shows $P(O_S|\lambda_N)$ for scene 2, where the fight starts at approximately *Time* = 80 and ends at *Time* = 120. Also in this case the fight is preceded by a quarrel, starting at *Time* = 50, with high volume sound. In this case both cameras observe the fight, which will give effect for $P(O_S|\lambda_N)$ when high volume sound and high optical flow values coincide in time (i.e. when $P(O_S|\lambda_N) \cong -12$). Figures 4 and 5 show the optical flows and crowd size for scene 2. The fight including the quarrel is well detected.

Finally, the lower-right panel presents results for scene 3, where the fight start at approximately *Time* = 120 and ends at *Time* = 130, with a quarrel starting at *Time* = 50. This fight has shorter duration, when compared to the previous action sequences. The fight is observed by camera 1 where intense actions are registered for a short time period. Likewise scene 1, the fight cannot be observed by camera 2 and therefore, the sound gives important indications on the fight. Due to the information obtained from the sound event detector, this abnormal event is detected with a good accuracy.

A video demonstration of the proposed system, in the PROMETHEUS dataset, is available at [18].

6.1.3 Summary of results

The experimental results offered practical validation of the proposed two-stage fusion method (and the assumption that the use of high-detail action analysis is not necessary for proper detection of fights). The information from different types of complimentary sensors that do not necessarily produce detailed or accurate information by themselves, at all time steps, turned out to be sufficient for accurate fight detection. The good result is achieved by fusing the different low-level events detected by the different modalities, and therefore the network of sensors as a whole, enables the accurate analysis of the scenes.

The correct fight detection rate D_T for the proposed fusion approach is closely related to the type of sensor network and the chosen HMM parameters. The crowd should not produce high volume sounds and intensive actions for longer time periods. Observations indicating abnormal events are O_3, O_5 and O_7. If O_S includes these observations, $P(O_S|\lambda_N)$ will be lower than T_N. $P(O_S|\lambda_N)$ is reduced even more if these observations will coincide in time. Based on Eq. (3) we obtain D_T = 81%, which represents an average value for the three fights. In this value we include also the preceding quarrels.

For this type of sensor network we have so far investigated three fights. However, the HMM based method has also been investigated for other fights in other scenarios and test data [2]. In those cases we used only visual and thermal infrared cameras and we had also a somewhat different set of observations. The method shows promising results also for those cases.

As a comparison with other types of sensor networks on the PROMETHEUS data set we have made some further analyses with a reduced sensor network. If no indication on high volume sound can be obtained for the fight in scene 3, the indication of a fight is vague, as this is presented in Figure 7. In scene 3 the audio information is important since the information from the optical flow includes large uncertainties (and the fight can not be observed by camera 2).

The evidence from audio is advantageous since the microphone array does not have the same limitations such as the limited perceptive view, restricted to a portion of the monitoring area, as it is the case for the optical cameras. Moreover, the thermal infrared camera offers important evidence about the crowd size in poor light conditions, when the performance of the visual cameras is strongly reduced.

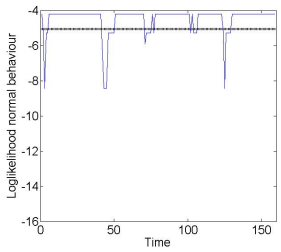

Figure 7. $P(O_S|\lambda_N)$ for the fight in scene 3 without audio information. Information is obtained only from the thermal infrared camera and the two visual cameras.

If there is only one camera observing the fights the information on the fights is poor and uncertainty is significant (see e.g. one of the diagrams in Figure 4). If the fights take place far away from the camera it will be difficult to distinguish sufficient changes in the optical flows to get reliable indications on the fight actions.

7 Conclusions

Investigating the problem of automatic fight detection in urban environments, we evaluated a two-stage method, which relies on a set of complementary sensors and a number of independent event detectors in the first stage, and on fusion and disambiguation of their decisions in the second stage. The experimental results obtained on normal and abnormal scenes from the PROMETHEUS database, whose content is purposely designed to represent scenes characteristic for outdoor surveillance applications, demonstrated the practical usefulness of the proposed method. It was shown that reliable detection of aggression

and fights can be achieved without detailed tracking of the body parts, which reduces the requirements to the positioning and calibration of the sensors. The last facilitates the utilization of the proposed method in real-world applications.

Acknowledgements

The research reported was partially supported by the PROMETHEUS project (FP7-ICT-214901) "Prediction and interpretation of human behaviour based on probabilistic models and heterogeneous sensors", co-funded by the European Commission under the Seventh' Framework Programme.

References

[1] PROMETHEUS project web-site: www.prometheus-FP7.eu

[2] M. Andersson, J. Rydell and J. Ahlberg, *Estimation of crowd behaviour using sensor networks and sensor fusion*, 12th International Conference on Information Fusion, Seattle, WA, USA, pp. 396-403, July 6-9, 2009.

[3] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci and A. Sarti, *Scream and gunshot detection and localization for audio-surveillance systems*, in AVSS, London, England, pp.21-26, September 5-7, 2007.

[4] J.-L. Rouas, J. Louradour and S. Ambellouis, *Audio Events Detection in Public Transport Vehicle*, in *IEEE Intelligent Transportation Systems Conference*, Toronto, Canada, pp. 733-738, September 17-20, 2006.

[5] H. Singh, R. Arter, L. Dodd, P. Langston, E. Lester and J. Drury, *Modelling subgroup behaviour in crow dynamic DEM simulation*, Applied Mathematical Modelling, Vol. 33, pp. 4408-4423, 2009.

[6] O. Oguz, A. Akaydin, T. Yilmaz and U. Güdükbay, *Emergency crowd simulation for outdoor environments*, Computer & Graphics (2010) doi: 10.1016/j.cag.2009.12.004 (article in press)

[7] A. Smith, C. James, R. Jones, P. Langston, E. Lester and J. Drury, *Modelling contra-flow in crowd dynamics DEM simulation*, Safety Science, Vol. 47, pp. 395-404, 2009.

[8] S. C. Moore, M. Flajslik, P. L. Rosin, D. Marshall, *A particle model of crowd behaviour: exploring the relationship between alcohol, crowd dynamics and violence*, Aggression and Violent Behavior, Vol. 13, pp. 413-422, 2008.

[9] C. Carincotte, X. Desurmont, B. Ravera, Bremond, F, J. Orwell, S. Velastin, J. Odobez, J. Corbucci, B. J. Palo and J. Cernocky, *Toward generic intelligent knowledge extraction from video and audio: the EU-funded CARETAKER project*, IEE Conference on Imaging Detection and Prevention (ICDP), London, UK, pp. 470-475, 13-14 June, 2006.

[10] W. Zajdel, J. D. Krijnders, T. Andringa and D. M. Gavrila, *CASSANDRA: audio-video sensor fusion for aggression detection*, IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), London, UK, 2007.

[11] S. Ntalampiras, I. Potamitis and N. Fakotakis, *On acoustic surveillance of hazardous situations*, in International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, pp. 165-168, April 19-24, 2009.

[12] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, In Proc. of the *IEEE*, Vol. 77, No. 2, , pp. 257-286, 1989.

[13] H.-G. Kim, N. Moreau and T. Sikora, *MPEG-7 Audio and Beyond: audio content indexing and retrieval*. Wiley Publishers, October 2005.

[14] G. Zhoun, J. H. L. Hansen and J.F. Kaiser, *Nonlinear feature based classification of speech under stress*, IEEE Transactions on Speech and Audio Processing, vol. 9, no. 2, pp. 201-216, March 2001.

[15] PRAAT software available at http://www.praat.org.

[16] S. Ntalampiras, D. Arsić, A. Störmer, T. Ganchev, I. Potamitis, and N. Fakotakis, *Prometheus Database: A Multimodal Corpus for Research on Modeling and Interpreting Human Behavior*, In Proc. DSP-2009, Santorini, Greece, 2009.

[17] Torch machine learning library. Available at: http://www.torch.ch

[18] Video demonstration of a fight. Available at: http://www.wcl.ece.upatras.gr/dalas/doku.php?id=demos.