Multi-Sensor Fusion for the Security Surveillance of Public Areas

Martin Litzenberger*[a], Michael Hubner [a], Bernhard Kohn [a], Kilian Wohlleben [a]
[a] AIT Austrian Institute of Technology GmbH, Giefinggasse 4, 1210 Vienna, Austria

ABSTRACT

Increasing security awareness in the public sector are leading to a more and more widespread use of surveillance applications. Although the available technologies like video processing are already well advanced, they still suffer from high false alarm rates when used under realistic conditions. We present a method for sensor fusion based on probability density maps and a rule engine. The system was tested in a public area using the combination of audio localization, audio classification and video detection using 79 simulated scenarios and 44 hours of sample data recorded over a period of several weeks. The false positive rate decreased by 60% and the event localization rate increased by 25% with the fusion approach compared to the detection performance of individual techniques.

Keywords: sensor data fusion, security surveillance

1. INTRODUCTION

With increasing security concerns in public areas, efforts to deploy monitoring and surveillance services are increasing massively. Since permanent visual inspection of many video screens in parallel is nearly impossible for the human operator, automated event detection is the only way to scale up large surveillance installations.

Video surveillance with automated video detection is still the most important sensor modality for security surveillance in public areas [1]. Audio detection is also used, especially for the detection and localization of certain strong acoustic events such as firearms or braking glass [2, 3]. However, even a small false positive rate per sensor from automatic detection leads to a significant number of false alarms as the number of sensors increases. On one hand, the permanent checking of false alarms means a considerable additional burden for security personnel and, on the other hand, leads to a loss of trust in the technology. An obvious approach to reduce the false rate is to compare data from automatic detections by different sensor modalities that monitor the same space.

A combined video detection and audio classification of critical events in 3-dimensional space has been described [4]. The fusion of different wireless positioning technologies in 2-dimensional space was reported in [5]. Often the temporal and spatial relationships of events in existing systems are only roughly analyzed, such as the assignment of a person detection from a camera image to acoustic cues detected in the vicinity of the same camera [6, 7]. Rule-based approaches with sensor installations to detect security relevant events have been described in [8].

Map-based methods can be used to determine the location of an object or security-related event. There are several approaches to achieve the necessary data association between different detections, such as feature-based and position-based maps [9]. The most common application of map-based sensor fusion today is found in the automotive sector, where a relatively small number of sensors are located in a relatively small spatial area around the vehicle [10]. Monitoring tasks with a very high number of sensors (e.g. up to 100) and for a very large geographical area (e.g. up to 1 km) based on evaluating the nearest neighbor relationships of elements in the map can become problematic. When evaluating nearest neighbor relationships, the computational effort increases dramatically with the number of sensors and the generated target hypothesis. For position-based maps, where each cell of the map represents the probability of an event or object at that location, the nearest adjacent relationship can be easily evaluated by addressing the adjacent cell of the grid. This method therefore scales much better for large sensor networks. Occupancy grid maps [11] and density maps [12] are common solutions for such location-based maps.

In this article, we introduce multi-sensor fusion using video crowd detection, audio event classification, and audio localization for detecting safety-critical events in public areas. The approach combines probability density maps with a rule engine that links data from multiple density maps. The performance of the chosen method, compared to detections of individual sensors, in terms of sensitivity and false positive rate was derived from 79 simulated scenes of safety-critical events and from 44 hours of recorded data from a public area over several weeks.

*martin.litzenberger@ait.ac.at; phone +43-50550-4111; www.ait.ac.at

2. MAP-BASED FUSION

The concept of the introduced data fusion approach is based on the aggregation of multiple density maps. The density maps spatially represent and cover the monitoring area for the different sensor modalities in use with the system, to allow a spatio-temporal comparison. In a first stage sensor detections are projected into the density maps and density cell values are updated. In the second stage, multiples of such density maps are merged in a fusion step and evaluated with a rule-based approach.

Each cell of a density map represents a location in the monitored area and holds a value representing the frequency of events for this location, supplied by the sensor modality associated to this map. Figure 1 illustrates the fusion concept on the example of an audio classification and a video crowd detector feeding into the fusion system. Each detection is attributed to an area by projecting the detection area into the corresponding cells of the density maps. The example shows two spatially overlapping audio detections in the left density map (indicated by two circular regions) and a rectangular area for the video crowd detection in the right map. For the sake of the illustration the detection region examples are visualized as circles and a rectangle. The actual shape of such detection areas may differ, and it is explained in the later section 3.1 how they relate to the actual detector modalities. The cells covered by the detections are updated by increasing the cell value by the confidence value delivered by the individual detector. To prevent a "runaway" of the cell values the values are clipped to 1 and all cell values are subject to temporal decay. The temporal decay introduces the "aging" of the detector information with time and is realized as an exponential decay of the cells' values. The decay time constant for this process is chosen by heuristics and depends on factors such as e.g. the update interval of the detectors. Thus, the decay time constant can be high for video detectors (video frame rate) and lower for other detector modalities.

All rule defined as the density maps participating in the fusion, their weights and a threshold completes the fusion: All density maps are periodically aggregated into one fusion map, by adding up the cell values corresponding to the same spatial location and weighting them depending on a predetermined weight for each density map. The weights are chosen to sum up to 1 to ensure that the aggregated values are again not exceeding a maximum value of 1. After aggregation the threshold is applied to all cells of the fusion map and the cells exceeding the threshold are generating an alarm. The grey-colored cells in Figure 1 illustrate such cells. The alarms' geographic location, that can be presented to a human operator on a map, is determined by the polygon enclosing all cells exceeding the threshold.

3. TEST DATA ACQUISITION AND EXPERIMENTAL VERIFICATION

The sensor system set up for testing and evaluating the fusion method consists of 2 video cameras and 6 microphones installed in a public area. Figure 2 shows the structure of the square and the layout of the sensor systems. The monitored area has a size of approximately 30m×60m. The sensor equipment is mounted on 3 lampposts on the left edge of the monitored area, identified in Figure 2 by posts 1, 2 and 3. Microphones are installed in pairs on the 3 posts, cameras are installed on posts 1 and 2. All sensors are mounted about 7m above the ground. The cameras' fields of view are indicated by the "V"-shaped polygons.

3.1 Setup, sensors, and detection algorithms for evaluation

The cameras and microphones are output to various software detectors that provide the input for our fusion framework. The algorithmic details of the software detectors used for this work are not further described in this article. The generic interface to the density maps is the "event" generated by the detector software. An event is defined by a georeferenced area of circular or polygonal shape that has the additional property type "Crowd" (C), "Audio Localization" (AL) or "Audio Classification" (AC), a probability value and a timestamp. When an event is entered into a density map, all grid cells of the map contained in the events circular or polygonal surface are updated using its probability property.

The image images from the video cameras are fed into an open-source video classification software trained to recognize crowds (C). The resulting bounding boxes in the image reference system of the cameras are converted into events with circular area using the measured field of view of the cameras for the coordinate transformation. The microphone signals flow into two different software detectors: First, 3 microphone pairs are used to estimate the direction of the dominant sound source from the signal delay between two microphones via time domain cross-correlation (AL). This information is converted into polygonal areas of triangular shape using the direction and the directional uncertainty of the source localization. The probability value for these events is derived from the audio signal quality.

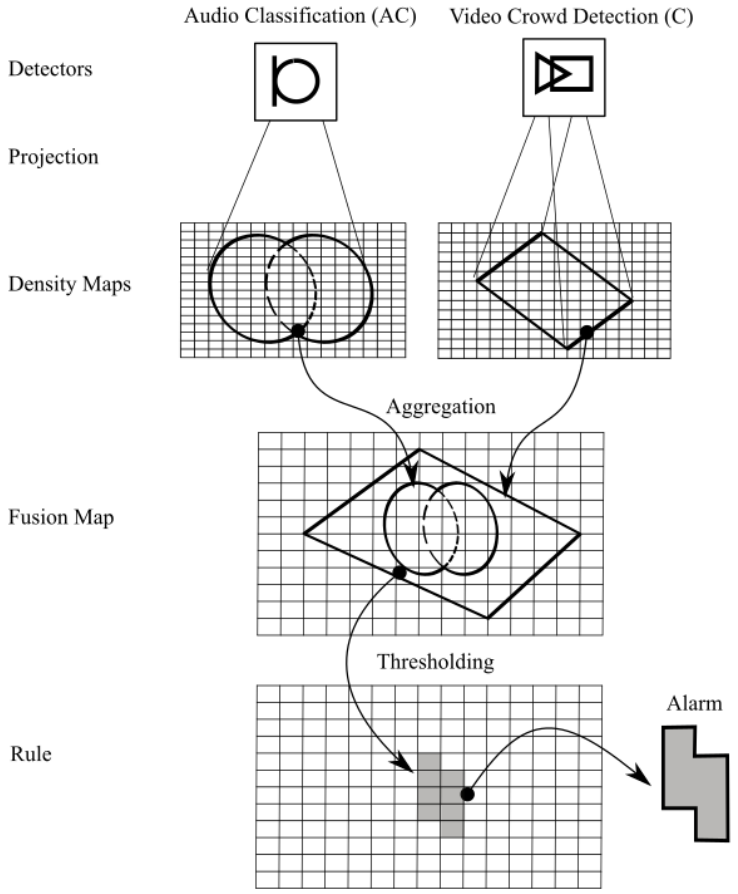

Figure 1. Schematics of the fusion concept.

Secondly, a microphone at each post feeds into a commercial audio classification software that distinguishes 4 different audio events: glass break, human aggression, weapon gunshot alarms and car alarms (AC). Any detection above a classification probability threshold generates an event with a circular range of 20m radius around its microphone location.

Each detector feeds into individual density maps with different time constant, to take into account the individual temporal dynamics of the detectors, and its weight contributes to the overall fused result.

3.2 Recorded scenario data

In order to evaluate sensitivity and false positive rates, we need test data where the underlying truth is known (ground truth). Test data was obtained using two methods: First, since not all events we want to record can be expected in reality during a test operation, they had to be simulated. Safety-critical events were simulated by actors on site day and night. The re-enacted scenes were divided into different scenarios. Table 1 gives an overview of the recorded data and the respective scenarios. Exemplary video images from the two cameras of a simulated attack scene are shown in Figure 3.

On the other hand, scenes in which none of the defined scenarios take place are required to derive the false positive rates. Data from everyday scenes taking place in the same public area was recorded over a period of several weeks across different seasons. From these data sets, 344 samples of everyday sequences with an average duration of 7 minutes were randomly selected.

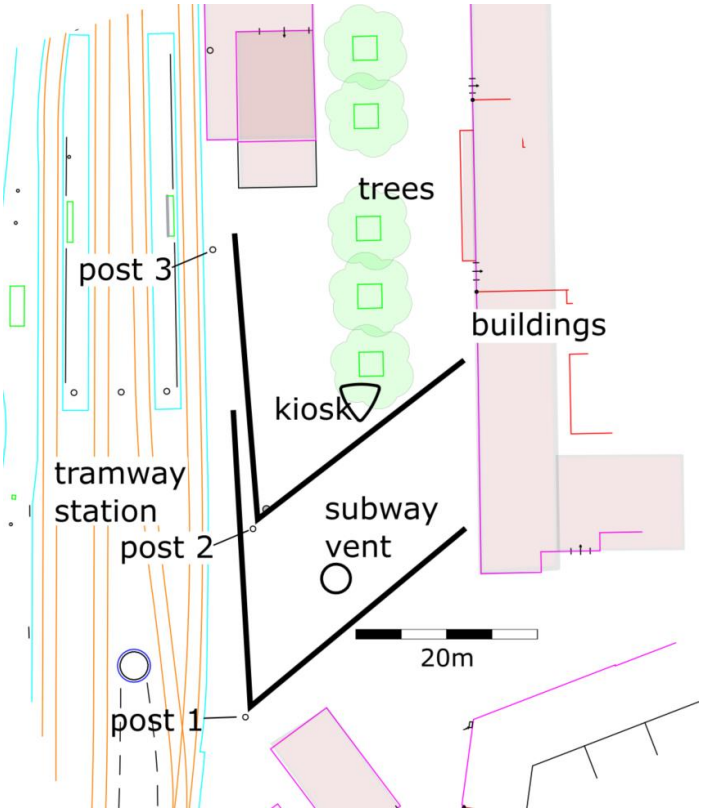

Figure 2. Schematical overview of the surveillance area and the sensor installation used for this work. The "V" shaped polygons indicate the field of view of the two surveillance cameras used.

For the statistical evaluation of the result, the ground truth must be determined. We consider an alarm to be correct (true positive) if the place and time match the ground truth. The ground truth for the simulated scenes was created with the help of a graphical annotation tool, in space as well as in time.

For the everyday scenes, it was not practical to review all the data due to its sheer volume. Therefore, no ground truth was created for the everyday scenes. In the course of the systematic evaluation, however, all alarms generated by the fusion algorithm were manually checked using the recorded videos. If scenes were found that warranted an alarm, they were considered true positives (e.g. drunk people screaming), if not, they were treated as false positives for the evaluation.

Table 1. Overview recorded test scenario data.

Test	Scenarios	Number of scenes	Total duration
Simulated	Attacks	20	11 min.
	Riots, hooliganism	22	33 min.
	Robbery	16	10 min.
	Damage to property	21	19 min.
	Total	79	73 min.
Everyday scenes	Random samples	344	44 hours

Figure 3. Two video frames of an aggression scenario recorded on the public test area with the two cameras located on posts 1 (lower frame) and 2 (upper frame). The two elements visible in the pictures, the kiosk (upper frame, right) and the cylindrical subway ventilation (lower frame, right) are indicated in the schematics of the public test installation in Figure 2 for orientation.

3.3 Experimental Verification

Binary classification methods are used to evaluate the results for each scene individually. First, the duration of each scene is divided into evaluation epochs. The length of the evaluation epoch is a fixed duration in seconds. Each scene has a series of alarms triggered by the fusion. For evaluation, in each scene actual events were annotated in space and time. For the spatial annotation a rectangular bounding-box around the actual event location ("ground truth window") was defined. Figure 4a shows an example of such a spatial annotation bounding-box for a crowd of people in front of the kiosk.

Each epoch is classified in this way into true positive, false positive, true negative and false negative epochs and a confusion matrix is produced from this. For the evaluation we have chosen the following derivatives of the confusion matrix:

- For sensitivity, a scene is considered a true positive (TP) detection if there is at least one valid alarm for all real cases (or positives P) in the data.
- The false positive rate, expressed in false alarms per 12 hours, FPR_{12h} is determined by counting the incorrect positions (FP) across the data set in relation to the duration of day and night. Day and night periods are weighted with 12 hours each. It is important to note that up to this step, only a temporal classification of the alarms is given. It is also necessary to evaluate the position of the alarms.

(a) (b)

Figure 4. (a) Video frame of a surveillance scene, showing a crowd of people, and the related annotation bounding-box manually entered in the map of the surveillance area. (b) Concept of deriving localization error from manual annotations. The red crosses indicate the fusion alarm locations, the red rectangle indicates the annotation bounding-box for the ground-truth of the related event.

To complete this, the spatial distance of an alarm location to the ground truth location was calculated. Since the time instant of the alarm must be met, only the true positives from the temporal verification are used to calculate this distance.

Figure 4b shows and example for the distance calculation used in the verification. For alarms whose center is within the ground truth window, the distance is defined as zero (see example d=0 in Figure 4b). If the center lies outside, the shortest distance to the ground truth window border is calculated (see examples d1 and d2 in Figure 4b). The localization rate of the alarms can be calculated, which indicates how many alarms are located inside the ground truth window, in relation to the number of all alarms.

4. FINDINGS

Sensitivity, false positive rate, and localization rate are calculated separately for day and night scenes for all test data (simulated and everyday scenes). The length of the epoch was chosen for all evaluations with 5 seconds. Table 2 shows the results of the fusion using three density maps (AC, AL, C). The false positive rate and localization rate decrease and increase, respectively, with including more detector modalities to the fusion system. In particular, the false positive rate was reduced significantly by more than 60%, while sensitivity was only slightly reduced (3%). However, the localization rate increased by up to 25%, which means that by adding more detectors with different characteristics, a significant improvement was achieved compared to using individual detectors alone.

5. CONCLUSION

In this article density maps were combined with a rule-based fusion approach. Events with a georeferenced region of circular or polygonal shape with additional properties type, probability and timestamp of detectors were used as input for density maps.

The fusion results of video and audio detection data collected on a public area equipped with two video cameras and six microphones were evaluated. Over 44 hours of test data were recorded, containing 79 simulated security-related scenarios. The fusion of three density maps (audio classification, audio localization and mass detection) was successfully used to reduce the false positive rate by 60% and the localization rate by 25% while maintaining sensitivity. Empirically, it can be concluded that it is crucial that detectors with different properties (locality and specificity) must be used to improve fusion quality.

Table 2. Evaluation of the fusion with up to three detectors: Audio classification (AC), audio localization (AL), crowd detection (C) evaluated on the basis of a data set with simulated scenes and everyday scenes. Epoch length was 5 seconds.

Used Detector Modalities	Sensitivity	FPR_{12h}	Localization
Day			
AC	81,02%	65	34,76%
AC + AL	81,02%	32	51,21%
AC + AL + C	78,02%	27	63,75%
Night			
AC	78,35%	88	30,95%
AC + AL	78,35%	41	51,05%
AC + AL + C	75,45%	33	56,83%

REFERENCES

[1] V. Tsakanikas and T. Dagiuklas, "Video surveillance systems-current status and future trends," *Computers & Electrical Engineering*, vol. 70, Nov. 2017.

[2] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference On*, IEEE, 2007, pp. 21–26.

[3] J. Stachurski, L. Netsch, and R. Cole, "Sound source localization for video surveillance camera," in *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference On*. IEEE, 2013, pp. 93–98.

[4] J. Kooij, M. Liem, J. Krijnders, T. Andringa, and D. Gavrila, "Multi-modal human aggression detection," *Computer Vision and Image Understanding*, vol. 144, pp. 106–120, Mar. 2016.

[5] J. Bohn and H. Vogt, "Robust probabilistic positioning based on high-level sensor-fusion and map knowledge," ETH Zurich, Report, 2003, accepted: 2017-1108T13:23:39Z.

[6] M. Andersson, S. Ntalampiras, T. Ganchev, J. Rydell, J. Ahlberg, and N. Fakotakis, "Fusion of acoustic and optical sensor data for automatic fight detection in urban environments," in *2010 13th International Conference on Information Fusion*, Jul. 2010, pp. 1–8.

[7] I. Lefter, G. J. Burghouts, and L. J. M. Rothkrantz, "Automatic Audio-Visual Fusion for Aggression Detection Using Meta-information," in *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, Sep. 2012, pp. 19–24.

[8] J.-R. Coffi, C. Marsala, and N. Museux, "Adaptive complex event processing for harmful situation detection," *Evolving Systems*, vol. 3, no. 3, pp. 167–177, Sep. 2012.

[9] C. Lundquist, L. Hammarstrand, and F. Gustafsson, "Road Intensity Based Mapping Using Radar Measurements with a Probability Hypothesis Density Filter," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1397–1408, Apr. 2011.

[10] M. E. Bouzouraa and U. Hofmann, "Fusion of occupancy grid mapping and model based object tracking for driver assistance systems using laser and radar sensors," in *Intelligent Vehicles Symposium (IV), 2010 IEEE*. IEEE, 2010, pp. 294–300.

[11] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, Jun. 1989.

[12] O. Erdinc, P. Willett, and Y. Bar-Shalom, "The BinOccupancy Filter and Its Connection to the PHD Filters," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4232–4246, Nov. 2009.