A Bayesian Approach - Data fusion for robust detection of vandalism and trespassing related events in the context of railway security

1st Michael Hubner
Digital Safety & Security
Austrian Institute of Technology GmbH
Vienna, Austira
michael.hubner@ait.ac.at

2nd Kilian Wohlleben
Center for Digital Safety & Security
Austrian Institute of Technology GmbH
Vienna, Austira
kilian.wohlleben@ait.ac.at

3rd Martin Litzenberger
Center for Digital Safety & Security
Austrian Institute of Technology GmbH
Vienna, Austira
martin.litzenberger@ait.ac.at

4th Stephan Veigl
Center for Digital Safety & Security
Austrian Institute of Technology GmbH
Vienna, Austira
stephan.veigl@ait.ac.at

5th Andreas Opitz
Center for Digital Safety & Security
Austrian Institute of Technology GmbH
Vienna, Austira
andreas.opitz@ait.ac.at

6th Stefan Grebien
Intelligent Acoustic Solutions
Joanneum Research
Graz, Austria
stefan.grebien@joanneum.at

7th Maria-Theresia Dvorak
Vienna, Austria
mariathdvorak@gmail.com

***Abstract*—In the domain of railway infrastructure, monitoring and securing the operational stability remains a significant problem. Vandalism, trespassing, sabotage and theft are constant threats, endangering the safety and integrity of the entire system. At the same time monitoring of these systems is becoming harder and harder as the systems grow and the amount of data produced by the surveillance equipment scales accordingly. Additionally, since specific sensor modalities can have weaknesses in detecting one kind of threat, it is often necessary to install different sensors to get a better understanding of the situation. In this paper we present a fusion model based on Probabilistic Occupancy Maps (POM) and Bayesian Inference for environmental mapping of critical events such as vandalism and trespassing in the vicinity of railway infrastructure. We show that this approach helps to increase accuracy, while simultaneously decreasing the amount of false alarms generated by a system.**

***Index Terms*—Bayesian inference, Multi-modal information fusion, critical infrastructure protection, trust in fusion systems**

I. Introduction

In the domain of railway security, which generally consists of large scale systems, monitoring and ensuring operational stability remains a significant challenge. Vandalism, trespassing, sabotage and theft are permanent threats, endangering the safety and integrity of this critical infrastructure [1]–[3]. At the same time permanent, seamless monitoring of the railway infrastructure is becoming harder and harder as the systems grow and the amount of data produced by the surveillance equipment scales accordingly. Additionally, since specific sensor modalities can have weaknesses in detecting one kind of threat, it is often necessary to install many sensors to achieve a complete coverage and improve the situational awareness of the security operators. On the contrary, more sensors lead to increased overall false alarms without any further processing mechanisms. Consequently, an increased amount of false alarms wastes resources in examining them and reduces trust in the overall system.

Several papers have recently tried to tackle this problem including [4]–[6]. Their approaches rely however on a single type of sensor; cameras in the case of [4] and [5] and infrared sensors for [6]. An example of using multiple sensors, namely radar and LiDAR is given in [7].

Using multiple sensor technologies has its own challenges as different sensors produce complementary and redundant data in different time frames. Spatio-temporal alignment of the sensor data is a challenge on its own. Data fusion can be applied at different steps of the signal processing pipeline and while different classifications exist, generally, a distinction between data level, feature level and decision level fusion is useful [8], [9]. Data level fusion is usually applied between similar sensor modalities (e.g. different microphones for denoising), while feature and decision level fusion can be used for different types of sensors with the goal to increase the overall performance of a sensor system with respect to accuracy and reduction to false alarms.

One example of an application of data fusion of different sensors for public surveillance is given in [10]. The authors

The work presented in this article has received funding from the Mobility of the Future programme. Mobility of the Future is a research, technology and innovation funding programme of the Republic of Austria, Ministry for Climate Action. The Austrian Research Promotion Agency (FFG) has been authorised for the programme management.

have used a combination of audio and video sensors for the detection of security relevant event in public areas. Another recent example of a similar system developed with NATO is given in [11]. Both are examples of feature level fusion, where the first focuses on identifying and refining the static location of an incident and the second applies fusion to the tracking of subjects.

In this paper we present our fusion model based on Probabilistic Occupancy Maps (POM) and a Bayesian approach for map fusion for geographical mapping of critical events. We show that this approach helps to increase accuracy while simultaneously decreasing false alarms in various scenarios.

II. Methodology

The main focus of this section lies in the description of the fusion model. For completeness we also briefly address the methodologies used for the detection systems.

1) Thermal Imaging - Person Detection: The You Only Look Once (YOLO) DNN-based detector has been widely adopted for its real-time capabilities and effectiveness in detecting different object classes [12]. In this work it was used for detecting people at a distance (up to 150m) using thermal imaging cameras. Thermal images present unique challenges due to lower quality, reduced contrast, fewer discernible features and increased noise compared to RGB images. The same model was applied on the images of two thermal cameras with different focal lengths (wide angle and telephoto) and overlapping fields of view to cover both short and long distances with appropriate image resolution. As a result bounding boxes of detected persons are provided by the detector.

Additionally, a Global Navigation Satellite System (GNSS) receiver for providing accurate camera position and orientation information was used. Based on the extrinsic (position and orientation) and intrinsic camera parameters it was possible to transform bounding boxes to a location (geo-referenced polygons) in world coordinates using a pinhole model. In Fig. 2 we see the resulting polygons of the geo-referenced polygons resulting from the used algorithm and the applied pinhole model.The softmax scores of the YOLO were used to provide an estimate for the confidence, which is necessary in the fusion model.

2) Acoustic - Vandalism and Person Detection: The acoustic sensors are equipped with a 64-ary microphone array, a GNSS module, a 9 degree-of-freedom orientation sensor, a single-board-computer, a battery and a LTE modem for communication. The signal from a single microphone is used as input for a detection stage to classify the incoming audio with a convolutional recurrent neural network [13] that has been trained with data from a previous measurement campaign. The neural network uses an 80-band log-mel spectrogram as input and consists of three convolutional and one recurrent layer. The last layer uses a sigmoid activation function to output the probability of a detection. If an event is detected, the signals of the 64 MEMS microphones are fed to an angle-of-arrival estimator. The 64 MEMS microphones are arranged as 4 concentric circles on a horizontal plane with diameters of $\{7.2, 10.5, 13.7, 17\}$ cm, respectively. For positioning we employ a Bartlett beamformer that generates angle-of-arrival estimates every 2 ms, combined with a k-means algorithm for clustering and variance estimation that outputs a mean and variance of the angle-of-arrival every 2 s in a local reference frame. Using data from the orientation sensor and the GNSS module, the angle-of-arrival estimation is transformed to a global reference frame. We decided to use a triangular-polygon to describe the possible position of the detected event with the estimated mean and variance describing the direction of the triangle and opening angle of the triangle, respectively (see Fig. 2). However, the acoustic measurements do not allow to estimate the distance of the detected event from the acoustic sensor. Thus, we set the maximum distance to 30 m and scale the triangle polygon accordingly. The probability for the fusion model is given by the output of the neural network.

A. Multi-Sensor Data Fusion

Occupancy grid mapping is one of the most popular approaches for geographical mapping. Its usage is prominent in the domain of autonomous driving for mapping multiple sensor information such as LiDAR, Radar and cameras to the surrounding of the vehicle. Our approach is based on the methodologies described in [14]. They have been adapted for the purpose of geographical mapping of critical events in the domain of critical infrastructure, though. In contrast to automotive applications, we assume that the area of interest (surveyed area) changes due to continuous sensor observations (e.g person detection, or spraying) reported in the act of a person committing vandalism or trespassing. For completeness we introduce the basic concept of our fusion model and its derivation from [14].

*1) **Probabilistic Occupancy Maps - POM**:* We use POMs as basis for spatio-temporal mapping of sensor observations. Let a POM be described as:

$$\mathbf{m} = \{m_{ij}, 1 \leq i \leq N_H, 1 \leq j \leq N_W\}, \tag{1}$$

where N_H and N_W denotes the number of rows and columns of the map that represents the area of interest. Thus, m_{ij} represents one cell of a POM. We denote z_t^k as a sensor observation of the k-th sensors observed at the time t. For each cell m_{ij} the posterior probability of the cells occupancy is defined as:

$$p(m_{ij}|z_t^k) \in (0, 1), \quad 1 \leq k \leq K \tag{2}$$

where K is the total number of sensors in use. Thus, each cell m_{ij} holds the probability of occupancy estimated by the received sensor observation at a specific time t. A probability near 0 means that the occurrence of a critical event is highly unlikely and vice versa. If no sensor information exists the actual state of a cell is *unknown*. This is also how a map $\mathbf{m}$ is initialized if we assume that no prior information about the

area of interest is available. The *unknown* state is characterized as follows:

$$p(m_{ij}) = \frac{1}{2}, \quad \forall i, j. \tag{3}$$

*2) **Bayesian Updating**:* Since sensor observations are continuously generated, POMs need to be updated over time accordingly. For this we use an updating process that combines Bayesian Inference in Log-Odds form with an exponential decay to model the effect of sensor information aging over time (i.e older events have less impact). We use the Bayesian formula in log odds to estimate the posterior $p(m_{ij}|z^k_{1:t_n})$ that infers all sensor observations of prior updating steps $1...t_n$. The Log-Odds ratio $l^k_{t_n}(m_{ij})$ at a cell m_{ij} is defined as:

$$\begin{aligned} l^k_{t_n}(m_{ij}) &= \log \frac{p(m_{ij}|z^k_{1:t_n})}{1 - p(m_{ij}|z^k_{1:t_n})} \\ &= l^k_{t_{n-1}}(m_{ij}) \cdot e^{-\frac{\Delta t}{\tau}} + \\ &\quad \log \frac{p(m_{ij}|z^k_{t_n})}{1 - p(m_{ij}|z^k_{t_n})} - \log \frac{p(m_{ij})}{1 - p(m_{ij})} \end{aligned} \tag{4}$$

Three terms are involved in (4). First, the previous state of the map $l^k_{t_{n-1}}(m_{ij})$ which is reduced by the forgetting factor described in [14]. The second term involves the log-odds ratio of the probability distribution $p(m_{ij}|z^k_{t_n})$. It represents the probability of each cell in the map given the current sensor observation $z^k_{t_n}$. This is the step where the current estimate of the map is updated with a new sensor observation. Finally, the third term represents the *prior probability* of the map, which will normally be $p(m_{ij}) = 0.5$ since the map is *unknown* a priory. In the case prior information of the map (e.g blind spots where an observation is physically impossible) is available, it can be incorporated via the *prior probability* of the map. In this work no prior information about the map was assumed, therefore (3) holds.

*3) **Forgetting Factor**:* Since in our approach we decided to use the log-odds form (4), we introduced the forgetting factor in the log-odds set:

$$l^k_{t_n}(m_{ij}) \leftarrow l^k_{t_n}(m_{ij}) \cdot e^{-\frac{\Delta t}{\tau}}, \quad \Delta t = t_n - t_{n-1} \tag{5}$$

keeping the same characteristics:

$$\lim_{l^k_{t_n}(m_{ij}) \to 0} \frac{1}{1 + \exp^{-l^k_{t_n}(m_{ij})}} = \frac{1}{2} \tag{6}$$

Thus, we ensure that the decay converges to the *unknown* state. This decay is applied before each updating process. The decay factor τ plays an important role in modelling the impact of past information. It usually is estimated empirically. As a rule of thumb in this work it is parameterized proportional to the frequency of the sensor observations provided. E.g $\tau = \frac{1}{20}$ in case sensor observations from the video detector are sent at 20Hz.

*4) **Bayesian Fusion**:* In order to integrate all available sensor observations, $z^k_{t_n}$, multiple maps are defined and fused according to Bayes' theorem. This approach is analogous to the updating process previously defined. There is one restriction that needs to be made, though. In the updating process the priors of the map $p(m_{ij})$ are used to infer information about the map. For the fusion process we omit this information, so that it won´t be inferred k times. That is why we assume (3) when calculating the fusion of all maps using (4). The formula for the fusion of k sensors in each cell m_{ij} is then defined as:

$$l^{1:K}_{t_n}(m_{ij}) = \sum_{k=0}^{K} l^k_{t_n}(m_{ij}) \tag{7}$$

To transform back into the probabilistic form we use the following function:

$$p(m_{ij}|z^{1:k}_{1:t_n}) = \frac{1}{1 + \exp^{-l^{1:K}_{t_n}(m_{ij})}} \tag{8}$$

*5) **Decision**:* Finally, a decision is made to trigger an alarm if a certain threshold $\kappa \in [0, 1]$ is exceeded for an individual cell $(i, j) \in m$. The alarm resulting from the decision process is localized at the cells of the resulting map after the fusion process, where the threshold κ is exceeded. This set of cells is denoted as $\{(i, j) \in m : p(m_{ij})(t) > \kappa\}$. As each cell m_{ij} represents a location in space (area of interest), as a result a geo-referenced alarm is generated. Adjusting κ allows us to parameterize the sensitivity of the fusion model. It gives a means of how much information is needed to trigger an alarm. In our work we chose this threshold based on end-user requirements $\kappa = 0.8$. It is important to note, that this influences the sensitivity of the fusion model. Hence, the performance evaluation as well. In practice a trade off needs to be found when choosing the parameters of the fusion model.

III. Data Description

In this section we describe the setting and location of the data recording of real-life scenarios as well as the format of the data, which is used in the evaluation. For the data recording two main use-cases have been defined. Namely, trespassing and vandalism. For each of the use-cases three scenarios have been defined, that represent typical critical events in the context of railway infrastructure security. A playbook was written, in which the course of the scenarios is described. Within the recording sessions all the scenarios were acted according to the definitions in the playbook. Furthermore, all the actors where instructed to act according to the definitions of particular scenarios. For example in the scenario *graffiti on train person group*, the persons were chatting as well as rattling of the spraying cans was done. In Table I a summary of the use-cases/scenarios and the recorded data in terms of acted scenarios and duration is depicted. Approximately 58 minutes of recording were used for the evaluation of our fusion model in total.

TABLE I
SUMMARY OF THE RECORDED DATA IN TERMS OF NUMBER OF SCENARIOS AND DURATION RECORDED

use-case / scenario	number of scenarios	duration
trespassing	15	00:26:06
person crossing track	5	00:06:29
person lingering on track	5	00:08:54
group crossing track	5	00:10:43
vandalism	15	00:32:22
grafiti on train one person	5	00:12:30
graffiti on train person group	5	00:13:26
smashing window	5	00:06:26
total	30	00:58:28

TABLE II
OVERVIEW OF THE OBSERVATIONS USED FOR EVALUATION

	observations			
use-case / scenario	**acoustic**	**thermal**	**fusion**	**gt**
trespassing	252	39682	30402	479
person crossing track	58	6728	5322	87
person lingering on track	4	7862	5852	124
group crossing track	190	25092	19228	268
vandalism	2095	52351	36513	722
graffiti on train one person	327	11468	9750	309
graffiti on train person group	477	25636	20362	295
smashing window	214	15247	6401	118
total	1270	92033	66915	1201

*1) **Data Acquisition**:* The recording took place in the vicinity of railway depot in Austria. The surveilled area covers three parallel rail (25m) tracks over a length of roughly 150m, (also including old parked wagons - for graffiti). We further call this area of interest, which approximately covers an area of 25m x 150m. In Fig. 1 a screenshot of the area of interest and the sensor placement is shown. Two thermal cameras mounted on a mast 7.5m above the ground and 3 acoustic sensors placed at ground level were used. The sensors were placed to exploit complementary sensor observation (e.g., person detection + rattling of spraying can). Thus, leading to an overlapping detection area of the thermal and acoustic sensors/detectors. The placement and field of view (FoV) of the thermal cameras is shown in Fig. 1 - in orange. For the cameras two different optics where used. One with a short focal length and wide FoV for the near field, to get a better resolution near the mast. For the second camera a far field optic with a long focal length was used for detection at greater distances (100m-150m). The acoustic sensor placement was chosen uniformly over the length of the area of interest. This setup was specifically chosen for maximum coverage of the area of interest. All the scenarios were acted within this area of interest.

*2) **Sensor Observation**:* A dedicated detector (classifier) was deployed for each sensor, which provides continuous sensor observations while monitoring the area of interest. The sensor observation were recorded during the acting of the scenarios. In Table II an overview of all the recorded sensor observations is given. A sensor observation comprises the location (i.e. geo-reference), the detected object, a timestamp, the confidence level and a label. The labels assigned by the video detector only consists of person, i.e., a person detector based on thermal images. The labels assigned by the acoustic sensors were: speech, rattle and vandalism (e.g glass brake). This data represents the required input for the fusion model.

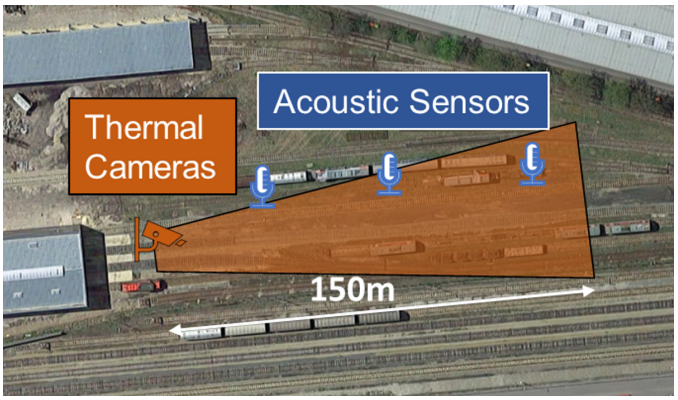

Fig. 1. Screenshot of the area of interest and the sensor placement

*3) **Fusion Observation**:* One of the main tasks when applying our fusion model is to select suitable configuration for specific use-cases. This means that while configuring the fusion model, we need to decide which observation are actually are fed into the fusion model. That is why it is crucial, for what purpose the fusion model is used. For example if the use-case is to detect graffiti on trains committed by a person, we consider person detection (thermal sensor) and rattling of can (audio sensor) as complementary sensor information. This way a fused observation always depends on the selected configuration and the model parameters (τ, κ). Analog to the sensor observation a fusion observation also comprises the location and the timestamp. The confidence is derived from (8) if threshold κ is exceeded. The label is derived from the used configuration. E.g vandalism or trespassing.

*4) **Ground Truth**:* In order to determine the performance of the overall system, it is essential to collect the ground truth. The ground truth consists of the time and location of the actors who enacted the scenarios of the use-cases. Mobile phones were used for this purpose. An app was installed that recorded the coordinates and timestamps of the actors during the scenarios. This ensured that a ground truth in the form of location and time was measured at all times during the scenarios. It should be noted that determining the GNSS coordinate is also a measurement procedure and is therefore subject to measurement errors (+-5m). This was taken into account during the evaluation. For this reason, the geo-reference of the ground truth is not a point, but a circle with a radius of 5 metres. The ground truth was recorded in the same format as the sensor observations. In Table II the total number of ground truth, sensor and fused observations collected is shown. In Fig. 2 a schematic example of the

recorded data of the geo-referenced sensor observation is depicted.

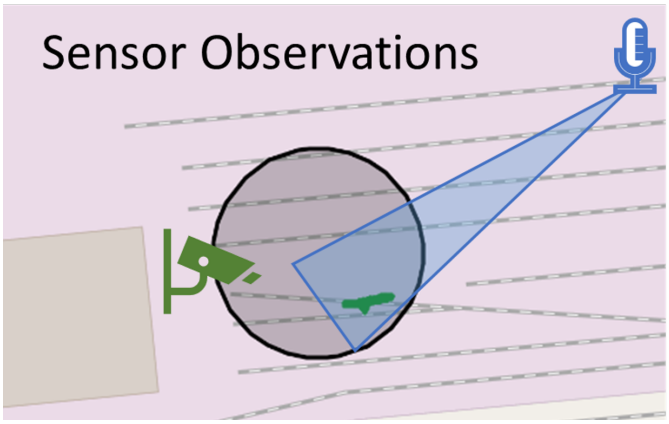

Fig. 2. Example of recorded sensor observations. Green - person detection by camera; Blue - speech detection by acoustic sensor; Black - Ground Truth recorded using GPS coordinates.

IV. Evaluation

To determine the performance of the sensors and the overall system (including data fusion) a confusion matrix is generated. Since the system deals with geo-referenced sensor observations, both the location and the time of the observations must be taken into account. To do so, an epoch is introduced for all data including sensor observations, fusion observations and ground truth. The epoch defines a time frame for an observation (and ground truth) in which it is valid. Usually, a duration of an epoch is selected based on the frequency of the observation. E.g a sensor observation is valid for 0.5 seconds if the observation frequency is 2Hz. Vice versa a ground truth might be valid for 5 seconds if the frequency of the GPS signal is 0.2Hz. Thus, we can characterize the classes of the confusion matrix as follows:

- **True Positive (TP):** A sensor observation is classified as TP if its epoch intersects with the epoch of an existing ground truth and the location of the sensor observation and the ground truth intersect as well.

- **True Negative (TN):** This is a special case. If one looks at the full duration of a scenario all epochs of sensor observation and ground truth will not cover it fully. The resulting gaps can therefore also be categorized in terms of epochs. This way it is possible to characterize TN as all epochs, for which neither an observation nor a GT exists

- **False Positive (FP):** A sensor observation is classified as FP if its epoch does not intersect any epoch of any existing ground truth or if the location does not intersect any ground truth present during that epoch.

- **False Negative (FN):** A ground truth is considered as a FN if there is no sensor observation whose epoch and location intersects with the epoch and location of the ground truth.

Evaluation Methodology: For the evaluation we use the following approach. First the sensor observations (thermal, acoustic) are evaluated without the fusion component in order to determine the performance of the system with detectors only. This means that the collected sensor observations from the acoustic and thermal detectors are evaluated against the ground truth. Next, the fusion model is applied to the same observations and the fusion results are evaluated with respect to its performance as well. Finally, the results of the detectors (thermal, acoustic) without fusion and the results of the fusion model are compared with respect to the selected metrics:

Accuracy The accuracy describes a measure of how many of all classes predicted (positive and negative) are actually correct:

$$Acc = \frac{TP + TN}{TP + TN + FP + TN} \tag{9}$$

F1-Score The F1-Score combines precision and the recall in the form of a harmonic mean. The F-measure is therefore very well suited to characterising a compromise between the accuracy and hit rate of the overall system. It is determined as follows:

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{10}$$

Relative False Positives It is crucial to specify the proportion of reported alarms that are considered false alarms, particularly for end-users. For this purpose we define the Relative False Positives as the proportion of False Positives among all existing observations.

$$RFP = \frac{FP}{N_O}, \tag{11}$$

whereas N_O is the total amount of sensor observations.

V. Results and Discussion

Table III shows the results of the evaluation by comparing the system without fusion model (Sensors) and system with our fusion model (Fusion) with respect to selected metrics. According to Table I each scenario includes data (sensor observations and ground truth) of five independent run through. For each scenarios all data was collected and consequently the confusion matrix and the metrics were calculated.

Table III shows very promising aspects. In two out of six scenarios we observed a positive tendency in all of the evaluated metrics. This means that by applying our fusion model it is possible to increase the accuracy and F1-score while simultaneously decreasing the relative amount of false positives. E.g for the scenario *smashing windows* this means that the accuracy increased from 73.7% (acoustic), 85.9% (thermal) to 95.9% (fusion), the F1-score from 41.6% (acoustic), 93.8% (thermal) to 97.6% (fusion) while the relative false

TABLE III
RESULTS OF EVALUATION. COMPARISON OF SYSTEM WITHOUT FUSION MODEL (SENSORS) AND SYSTEM WITH FUSION MODEL (FUSION) WITH RESPECT TO SELECTED METRICS.

	accuracy			**f1-score**			**relative false alarms**		
	sensors		fusion	sensors		fusion	sensors		fusion
	acoustic	thermal		acoustic	thermal		acoustic	thermal	
trespassing									
person crossing track	0.758	0.886	0.955	0.416	0.938	0.976	0.466	0.115	0.045
person lingering on track	0.992	0.842	0.923	0.049	0.913	0.959	0.250	0.158	0.077
group crossing track	0.912	0.799	0.850	0.694	0.885	0.915	0.142	0.202	0.151
vandalism									
graffiti on train one person	0.807	0.534	0.632	0.690	0.685	0.769	0.266	0.472	0.371
graffiti on train person group	0.625	0.810	0.775	0.672	0.893	0.871	0.426	0.190	0.226
smashing window	0.737	0.859	0.959	0.612	0.923	0.977	0.374	0.141	0.041

alarms were reduced from 46.6% (acoustic), 11.5% (thermal) to 4.5% (fusion). We observed a similar behaviour for the rest of the scenarios, were the F1-score was increased (compared to the sensors), while the relative false alarms decreased. From the perspective of an operator this reflects a very strong benefit, since the overall performance of the monitoring system is increased while false alarms are reduced at the same time. There is one exception, though. In the scenario *graffiti on train person group* the overall performance of the fusion with respect to all the metrics was in fact reduced.

In Fig. 3 we see a snapshot of one of the recorded run through of the scenario *graffiti on train person group*. Specifically, in this scenario the actors were told to chat during the course of acting the scenario. This was detected by the acoustic sensor whereas the confidence (i.e probability) near the persons (left - acoustic detector) was higher than the confidence of the detector farthest away (right - acoustic detector). Although, there was less confidence at the right acoustic detector, over time, it was sufficient enough confidence inferred into the fusion model to trigger an alarm. Which, in fact represents a valid alarm, since it correctly points to the direction of the people walking (Ground Truth). The reason why it results in a reduction of the performance lies in the geo-reference of the acoustic detectors. As described in the sensor models, an acoustic detector models its geo-reference based on the uncertainty of the detection (beam angle) and a maximum estimated distance. By doing this, geo-references are generated that do not intersect the ground truth, although the time is correct. Thus, leading to a False Positive in the evaluation. As a result, the accuracy and the F1-score deteriorate and the relative false positives increase. A similar case also occurred in the scenario *person lingering on track*. In Table III the acoustic sensor yielded a F1-score of 4.9% which represents a realistic value considering the absence of talking or other characteristics that can be detected by it. Although, the accuracy of the fusion was reduced (compared to acoustic) the f1-score was in creased while simultaneously decreasing the relative false alarms. With these example we also identified a challenging topic. On the one hand, it can be addressed by investigating the modelling of the geo-reference of an observation. But on the other hand it is also possible to adapt the performance metrics to be less restrictive in terms of location. We plan to address this challenge and the variation of the fusion parameters and its impact on the selected metrics in future work.

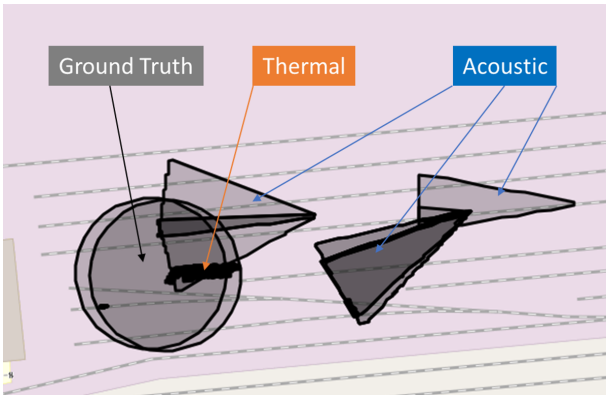

Fig. 3. Example data of graffiti on train group. Big Circles - Ground Truth, small circles - video observations, triangles - acoustic observations

VI. CONCLUSION

In this paper we presented a Bayesian approach for data fusion of audio and thermal observations for the purpose of robust detection of vandalism and trespassing in the context of railway security. The proposed fusion model was evaluated on data recording of six different scenarios dedicated to vandalism and trespassing. The scenarios were enacted at a railway depot in Austria covering an area of roughly (25m x 150m). The employed approach demonstrates that the introduction of our fusion model can reduce the relative number of false alarms while simultaneously increasing accuracy and F1-Score. We observed this behaviour in two out of six scenarios. For the scenario smashing window (vandalism) this means that the performance in terms of accuracy increased by 10,2% while at the same time the number of relative false alarms decreases by 10,3% in comparison to a detector only system (no fusion model). This way, monitoring systems for the purpose of surveillance of critical infrastructure such as the railway infrastructure can be enhanced by deploying our fusion model. This approach reduces false positives and saves

time for operators verifying alarms, while also improving the reliability and trustworthiness of a monitoring system.

ACKNOWLEDGMENT

We are grateful to the company PKE Holding AG, who generously provided the cameras used for the recording of the data. Their contributions have been instrumental in the success of this study. We are also grateful to the company ÖBB-Operative Services GmbH & Co KG, who generously provided the test location and organized the actors for simulating the scenarios. Their contributions have been instrumental in the success of this study.

REFERENCES

[1] A. Killen, D. S. Coxon, and D. R. Napper, "A Review of the Literature on Mitigation Strategies for Vandalism in Rail Environments," 2017.

[2] S. Grabušić and D. Barić, "A systematic review of railway trespassing: Problems and prevention measures," *Sustainability*, vol. 15, no. 18, 2023.

[3] T. Zhang, W. Aftab, L. Mihaylova, C. Langran-Wheeler, S. Rigby, D. Fletcher, S. Maddock, and G. Bosworth, "Recent Advances in Video Analytics for Rail Network Surveillance for Security, Trespass and Suicide Prevention—A Survey," *Sensors*, vol. 22, p. 4324, Jan. 2022. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[4] Z. Cao, Y. Qin, Z. Xie, Q. Liu, E. Zhang, Z. Wu, and Z. Yu, "An effective railway intrusion detection method using dynamic intrusion region and lightweight neural network," *Measurement*, vol. 191, p. 110564, Mar. 2022.

[5] H. P. Haryono and F. Hidayat, "Trespassing Detection using CCTV and Video Analytics for Safety and Security in Railway Stations," in *2022 International Conference on ICT for Smart Society (ICISS)*, pp. 01–04, Aug. 2022.

[6] D. R. Edla, D. Tripathi, V. Kuppili, and R. Dharavath, "Multilevel Automated Security System for Prevention of Accidents at Unmanned Railway Level Crossings," *Wireless Personal Communications*, vol. 111, pp. 1707–1721, Apr. 2020.

[7] W. Lu, Q. Wang, J. Ding, W. Niu, and J. Sheng, "Rail Track Area Environment Perception Based on Rader Target Gird," in *2022 3rd International Conference on Electronics, Communications and Information Technology (CECIT)*, pp. 236–241, Dec. 2022.

[8] B. Dasarathy, "Sensor fusion potential exploitation-innovative architectures and illustrative applications," *Proceedings of the IEEE*, vol. 85, pp. 24–38, Jan. 1997. Conference Name: Proceedings of the IEEE.

[9] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, "Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges," *Information Fusion*, vol. 35, pp. 68–80, May 2017.

[10] M. Hubner, C. Wiesmeyr, K. Dittrich, B. Kohn, H. Garn, and M. Litzenberger, "Audio-Video Sensor Fusion for the Detection of Security Critical Events in Public Spaces," in *2021 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 1–6, Sept. 2021.

[11] H. Bouma, M. L. Villani, A. van Rooijen, P. Räsänen, J. Peltola, S. Toivonen, A. De Nicola, M. Guarneri, C. Stifini, and L. De Dominicis, "An Integrated Fusion Engine for Early Threat Detection Demonstrated in Public-Space Trials," *Sensors*, vol. 23, p. 440, Jan. 2023. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] E. Cakir and T. Virtanen, "Convolutional Recurrent Neural Networks for Rare Sound Event Detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pp. 27–31, Nov. 2017.

[14] C. Galvez del Postigo Fernandez, "Grid-based multi-sensor fusion for on-road obstacle detection: Application to autonomous driving," Master's thesis, KTH, Computer Vision and Active Perception, CVAP, 2015.