Video-based fusion of multiple detectors to counter terrorism

Luigi De Dominicis 1, Henri Bouma 2*, Sirra Toivonen 3, Cristiano Stifini 4, Maria Luisa Villani 1, Antonio De Nicola 1, Arthur van Rooijen 2, Jan Baan 2, Johannes Peltola 3, Arttu Lämsä 3, Pauli Räsänen 3.

1 ENEA, Rome, Italy
2 TNO, The Hague, The Netherlands
3 VTT, Tampere, Finland
4 ATAC, Rome, Italy

ABSTRACT

Terrorism is an international security challenge. The early detection of threats (e.g., explosives or firearms) could provide a valuable contribution to the ability to prevent, protect and respond to terrorism. This paper presents a system for the management of a plurality of sensors to improve the threat-detection capabilities without disrupting the flow of passengers. The system improves the prevention capabilities of soft targets (such as airports, undergrounds and railway stations) with a high number of daily commuters. The system architecture consists of three main components. The first component is 2D video tracking and re-identification (Re-ID), which allows the labelling and tracking of commuters in a small area. Thereby, it supports the fusion of sensors at different locations. The Re-ID has a smart training strategy with anonymized snippets to increase flexibility for new environments. The second component is 3D video tracking with a stereo camera, which gives a more accurate location measurement than 2D video. Location prediction is used to compensate for latency in the control of active elements in the threat detection sensor. Recurrent neural networks for location prediction were trained by using real 3D tracking data from a railway station. The performance is evaluated with a ground-truth based on Ultra-Wide Band (UWB) radio positioning and a coordinate conversion method was created to compensate for identified inaccuracies. The third component is Command & Control (C&C), which consists of three submodules: message broker, data-fusion and security client. The message broker is a publish-subscribe middleware layer to enable flexible integration of the various sensors and components. The data-fusion combines outputs of multiple sensors. In case of a suspect person, the security client triggers an alarm and a comprehensive report is sent to the security guards.

Keywords: Fusion, re-identification, surveillance, command and control, threat detection, counter-terrorism, explosives, firearms.

1. INTRODUCTION

Terrorism is a persistent global threat that knows no border or nationality. The early detection of threats could provide a valuable contribution to the ability to prevent, protect and respond to terrorism. This paper presents a real-time surveillance system for the fusion and management of a plurality of sensors to improve the threat-detection capabilities in public spaces without disrupting the flow of passengers. The system improves the prevention capabilities of soft targets (such as airports, undergrounds and railway stations) with a high number of daily commuters.

The system for the fusion of sensors is called the ‘INSTEAD system’ because it was developed in the INSTEAD project, which is part of the DEXTER program, which was launched in 2019 [NATO-SPS, 2018]. Other projects in this program

* Henri.Bouma@tno.nl; phone +31 888 66 4054; www.tno.nl

https://doi.org/10.1117/12.2598147

Copyright © 2021 Society of Photo-Optical Instrumentation Engineers (SPIE). One print or electronic copy may be made for personal use only. Systematic reproduction and distribution, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited.
focus on the development of sensors for detection of firearms (MIC project) and explosives (EXTRAS project). The complete integrated DEXTER system is expected to be live-tested in a metro station in Rome in 2022.

The INSTEAD system consists of 2D and 3D video tracking and Command & Control (C&C). The 2D video tracking – including re-identification (Re-ID) – allows the tracking of commuters from one sensor to another to support the fusion of sensors at different locations. The main novelty of this part is the smart Re-ID training strategy with anonymized snippets to increase flexibility for new environments. The 3D video tracking allows a more accurate location estimation to facilitate the aiming of threat-detection sensors (such as EXTRAS). The main novelty of this part is the location prediction to compensate for latency in the control of active elements in the threat detection sensor. Both 2D and 3D focus on unbiased location estimates. The C&C facilitates flexible integration of various sensors with a new message broker and generation of a comprehensive report for security guards.

The outline of this paper is as follows. The overall system architecture is described in Section 2. Video tracking is described in Section 3, which contains a description of the methods for 2D video tracking and 3D video tracking, and a comparison between results of 2D and 3D. Command and control is described in Section 4. Finally, the conclusions are summarized in Section 5.

2. OVERALL SYSTEM ARCHITECTURE

The INSTEAD system consists of several components (Figure 1). The first component is 2D video tracking and re-identification (Re-ID), which allows the labelling and tracking of commuters. Thereby, it supports the fusion of sensors at different locations [Bouma, 2013]. The second component is 3D video tracking and prediction with a stereo camera, where prediction is used to compensate for the latency in the control of active elements. The third component is Command & Control (C&C), which consists of three submodules: message broker, data-fusion and security client. The message broker is a publish-subscribe middleware layer to enable flexible integration of the various sensors and components. The data-fusion combines outputs of multiple sensors. The security client generates a comprehensive report for each commuter in a timely manner to trigger alarm management by the security guards.

![Figure 1: Overall system architecture](image)

3. VIDEO TRACKING

3.1 Method for 2D-video tracking and re-identification

The 2D-video tracking system is developed by TNO and it aims to track people from one camera to neighboring cameras in a local environment. The architecture consists of person detection, anonymization, signature computation, tracking, and matching (Figure 2). Person detection is performed using the SSD detector [Liu, 2016] with a ResNet50 backbone [He, 2016]. Typically cameras produce a video stream of 25 frames per second, full HD resolution (1080p) and compression (H.264). The person detection is applied at a lower frame rate (5 frames per second) and the other frames are skipped so...
that all tasks can be performed in real-time on a single computer. Anonymization was performed by applying a median filter to the top 25% of the person [Rooijen, 2020]. For re-identification, we used the “strong baseline” implementation of Luo et al. [Luo, 2019]. The signature is computed for every detection. Tracking combines multiple snippets in subsequent frames in the same camera over time and thus generates a single-camera track [Marck, 2014].

Person re-identification is commonly used as an interactive system for forensic search, where the user is supported by the system to find a suspect. The forensic system sorts candidates by similarity and the user manually decides which candidate matches the suspect. The aim of the surveillance system in this paper is to fuse multiple sensor signals in real-time [Bouma, 2013]. This requires full-automatic assignment to the best matching candidate. All detections and tracks of each camera are continuously stored in a database. When a detection appears in a region of interest (ROI) near the second sensor (shown in Figure 1), it is matched with detections in the database to retrieve a similar result. The search space is limited in space (only detections near the first sensor) and limited in time (only detection of the last minute).

The common way to use state-of-the-art deep-learning technology for person Re-ID is to train and test it on independent parts of the same dataset. However, if it is trained on one large public dataset it does not generalize well to other environments, which makes it less suitable for practical applications. The collection and annotation of a new in-domain dataset can be difficult and time consuming. Therefore we used a new strategy for rapid Re-ID retraining in new environments. In order to train Re-ID from the first to the last camera, we create and annotate a training set by using multiple intermediate cameras (Figure 3). In a situation with multiple (almost overlapping) cameras, tracking is a reliable technique to automatically assign unique labels to each person. In this way, pairs with the same unique label become available in the first and the last camera, which can be used for training of Re-ID.

The cameras are manually calibrated to support the conversion of a location in pixel coordinates to a location in world coordinates [Hollander, 2015]. The manual calibration relies on a set of points on the ground plane that are indicated by the user in the camera view (i.e. pixel coordinates), and for which the world coordinates are also known. These points are used to optimize the calibration parameters. The basic approach to approximate the location of people is to take the bottom-center of the bounding box as an estimate of the location of the feet. Near the edges of the camera view, this may result in a biased location estimate because the orientation of people is not exactly vertical (blue dot in Figure 6). An new unbiased estimate of the location is developed by TNO and it uses two steps. First the orientation of the person is estimated projecting the feet location to the world coordinates, going upward to 1.0 meter height, and projecting this point back to pixel coordinates. The line through these two points gives a better estimate of the orientation of the person in pixel coordinates. Second, this orientation is used to compensate for the bias by taking the intersection between one line with the same orientation through the center of the bounding box (green line in Figure 6) and another line through the bottom of the bounding box. This intersection point (green dot in Figure 6) is converted from pixel to world coordinates to obtain an unbiased estimate of the person location.

Proc. SPIE, Vol. 11869, p. 3
3.2 Method for 3D video tracking and prediction

The 3D position tracking is based on a commercial StereoLabs ZED2 stereo camera [StereoLabs, 2021]. For 3D position data, the ZED2 application software has been configured to operate in world-coordinate frames in positional tracking mode, to detect humans as monitored objects and to provide the estimated 3D center position of head of observed humans. The ZED2 application software deploys a vendor provided proprietary algorithm taking into account camera installation height and angle to estimate 3D object position in a world coordinate system, and includes vendor provided camera unit-specific calibration to compensate optical lens distortions. The location accuracy of the 3D tracking has a bias in different regions in the field of view. For compensating this bias, a simple and computationally efficient affinity transformation was developed by VTT. A robot has been utilized in capturing a large data set about the systematic distortion characteristics. All points in the robot dataset have been utilized in computing the conversion matrix from the biased estimate to the unbiased one. A robot-based calibration allows fast and accurate installation of 3D tracking system in new environments.

When camera-based tracking is applied, the latest position information obtained from the system does not necessarily match the actual current position of the tracked person. This is due to the processing delay in the system. If the person is moving while the system is executing computational or mechanical pan-tilt-zoom (PTZ) operations to obtain the latest position, a spatial error will occur between the actual and estimated coordinate. To be able to use the coordinate information, e.g. to make measurements of the object remotely in optical means, the spatial error must be compensated. The compensation can be done by using the collected position information from a person (i.e. trajectory) to forecast his/her position at a fixed short time offset in the future. Two prediction approaches were developed by VTT. The first approach is a recurrent neural network (RNN) and the second is a linear regression (LR) as a reference estimator. The intuition for using RNNs is that they can learn the relations of coordinates in temporal dimension and by doing so, forecast the coordinate to a forward-looking time point. The RNN model consists of one LSTM layer with 256 neurons. The output is obtained through a fully connected layer with two neurons (output for x and y). The training data is preprocessed by removing short or otherwise void tracks. Since the goal is to make a forecasting model to be independent of the surroundings where it will be deployed, we wanted to avoid a situation where the model learns features based on actual coordinates. For this reason, absolute coordinate values of a trajectory are not used as an input for the model. Instead, the first-order derivatives of x and y coordinates of the trajectory are used as an input, and the derivative values are normalized into range [0, 1] (limits inclusive). The output of the model is also made relative by formulating it as a normalized offset from the last known x and y coordinates. The second approach with the LR-model is trained for each track and coordinate axis separately, with many of the same pre-processing steps but with absolute world-coordinates as an input.

3.3 Experimental setup to compare 2D and 3D video tracking

The positioning accuracy of 2D and 3D tracking is important, when system makes a handover of a followed person between the tracking systems. In order to evaluate performance of 2D video and 3D video tracking systems, there has been set up a controlled test environment of 10x8 meter (Figure 4). A commercial Ultra-Wide Band (UWB) radio positioning system based on Decawave DWM1001 modules [Decawave, 2017] (4 x anchors and 1-32 tags) provides the best-effort ground-truth reference data at sample rate of 10Hz and operates in real-world coordinate system. 2D video and 3D positioning data have been captured using a commercial UniFi G3 surveillance camera [Ubiquiti, 2019] and ZED2 stereo camera [StereoLabs, 2021]. The tracked UWB tag has been placed on top of the test object, which is as close as possible to the human head center.
In order to evaluate accuracy of the UWB system to ground-truth values in real-world coordinate system and to justify UWB feasibility for a reference system, there has been completed a measurement at 25 static ground-truth test points defined at ±0.5cm accuracy. The UWB system manufacturer reports an accuracy within ±10cm. Our measurements indicate that the L2-error is less than 20cm in 90% of the location estimates (Table 1).

<table>
<thead>
<tr>
<th>Percentile</th>
<th>L2-error [m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0.03</td>
</tr>
<tr>
<td>50%</td>
<td>0.11</td>
</tr>
<tr>
<td>90%</td>
<td>0.19</td>
</tr>
<tr>
<td>100%</td>
<td>0.21</td>
</tr>
</tbody>
</table>

3.4 Results of the comparison of 2D and 3D video tracking

In the experiments we compare 2D video (on UniFi and ZED2 video, with and without bias) and 3D video positioning (ZED2, with and without bias) with the UWB ground truth. The ZED2 timestamps were compensated for a latency of 200ms. The results are shown in Table 2. The results show that Unbiased 2D video tracking on both sensors (ZED2 and UniFi) results in the most accurate estimates, both reaching the average accuracy better than 30cm and a median accuracy better than 25cm.

Figure 5 presents an example trajectory of a person walking in different regions in our test setup. Unifi and ZED2 cameras are located in the bottom right corner of the area. The picture contains all measurement points listed in the Table 2. It can be seen that positioning accuracy is good in the center of the field of view but reduces on sides of the view as well as when distance from the camera increases. Bias correction can significantly increase the accuracy especially on the sides of the view. Figure 6 illustrates that due to optical distortion, the person inside the bounding box is not vertically properly oriented, that needs to be taken into account when performance is estimated against UWB ground truth.

Figure 7 depicts the L2 error between the best 2D and 3D tracking systems, including Unbiased 2D video for UniFi and Unbiased 3D tracking for ZED2. The results show that a handover between 2D and 3D tracking systems is achievable when the bias correction for both have been done for UWB based common coordinate system. The left upper corner is most far from the both cameras and this region especially problematic for ZED2 distance estimation routines.
Table 2: Comparison of 2D and 3D video tracking

<table>
<thead>
<tr>
<th>Sensor 1</th>
<th>Sensor 2</th>
<th>Average R [m]</th>
<th>Median R [m]</th>
<th>Std R [m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZED2 + 3D (VTT)</td>
<td>UWB</td>
<td>0.80</td>
<td>0.81</td>
<td>0.25</td>
</tr>
<tr>
<td>ZED2 + 3D unbiased (VTT)</td>
<td>UWB</td>
<td>0.34</td>
<td>0.29</td>
<td>0.25</td>
</tr>
<tr>
<td>ZED2 + 2D (TNO)</td>
<td>UWB</td>
<td>0.40</td>
<td>0.41</td>
<td>0.20</td>
</tr>
<tr>
<td>ZED2 + 2D unbiased (TNO)</td>
<td>UWB</td>
<td>0.25</td>
<td><strong>0.23</strong></td>
<td>0.13</td>
</tr>
<tr>
<td>UniFi + 2D (TNO)</td>
<td>UWB</td>
<td>0.33</td>
<td>0.29</td>
<td>0.19</td>
</tr>
<tr>
<td>UniFi + 2D unbiased (TNO)</td>
<td>UWB</td>
<td>0.28</td>
<td><strong>0.23</strong></td>
<td>0.20</td>
</tr>
</tbody>
</table>

Figure 5: Visualization of trajectory and L2 error to the reference in assessed 2D and 3D systems.
Figure 6: Example the ground truth (red), biased video estimate (blue) and unbiased estimate (green).

Figure 7: Visualization of L2 error between the best 2D and 3D tracking methods.

3.5 Experiments and results of location prediction

A dataset collected from a railway station consisting of human trajectories from a period of seven days was used to train and evaluate machine learning models for predicting the location to a fixed forward-looking time offset. The dataset was collected by using 3D depth cameras and transforming the depth camera data into trajectories. The total number of
trajectories in the dataset was approximately 1.1 million. Six days of data was used for training and one day was used for testing. The input to the model consisted of five seconds of coordinate data. The quantitative results for this experiment are shown in Table 3. The prediction error is calculated as an L2-error between the predicted and the corresponding detected coordinate points. The results show that the recurrent neural network (RNN) model performs better than linear regression (LR). An example on actual and predicted locations for the ZED2 is shown in Figure 8. Note that the prediction includes large errors next to all the rapid changes in the trajectory path, like those at the corners of the test area while straight and slightly curly track parts have acceptable error in general. Secondly, the current data model outputs the first prediction after a track data buffering delay of five seconds.

Table 3: Prediction error of RNN and LR with the railway station dataset.

<table>
<thead>
<tr>
<th>Forecast Length (sec)</th>
<th>RNN Average L2 Error [m]</th>
<th>LR Average L2 Error [m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.151</td>
<td>0.466</td>
</tr>
<tr>
<td>1.0</td>
<td>0.245</td>
<td>0.634</td>
</tr>
<tr>
<td>2.0</td>
<td>0.494</td>
<td>1.010</td>
</tr>
</tbody>
</table>

Figure 8: Example visualization on L2 error between actual and predicted trajectory with recurrent neural network (forecast of 1 seconds).

4. COMMAND AND CONTROL

The Command and Control is developed by ENEA and it is a distributed software component with the roles of orchestrating the functions of the technological components for threat detection and of supporting security agents in the alarm management that may follow.

More precisely, with reference to the architecture of Figure 1, the responsibilities of the Command and Control are:
• **Management of components communication.** This function concerns managing the messages exchange among all the other technological components of the architecture to enable real-time situation awareness and tracking of the commuters who will have to be inspected by the security agents. A Time-synchronized message broker component is devoted to handling the messages exchange.

• **Real-time data fusion and prioritization of messages based on detections.** This function is realized by the Data fusion component, which combines data from the video and sensor-based systems as soon as they are transmitted, and timely notifies the components of the subsequent data flow about results related to suspect people. A data-stream integrator server hosts the Data fusion component.

• **Persistent storage of data to enable post-processing.** This function concerns data storage of all the messages received and sent by the Command and Control, including partial and complete sensor and video results, to the aim of the overall system monitoring and to enable ex-post analysis of data for decision making. A Database Management System (DBMS) is used for such a purpose.

• **Alarm communication.** This function concerns dispatching all the resulting alerts to several sites, including local security agents and a security room. A Security client component is devoted to collecting alarms and dispatching them to the various security management applications that may be registered to this aim.

A detailed description of the software components realizing the functions above follows.

### 4.1 Message broker

The Message Broker provides the sole interface by which all the other components send and receive application-specific data, with a timestamp synchronized with respect to the time of a common NTP server integrated in the architecture.

The communication model is according to a publish-subscribe paradigm, which is especially effective for distributed systems with highly decoupled components and provides a flexible means to propagate messages to multiple components that are interested in them. In the system set up phase, the publish interface allows a component to register itself as a provider of some type of message and the subscribe interface as a consumer of message types being published by others. All communications are asynchronous.

To realize the message flow of the system illustrated in Figure 1, the following configuration is used, where the topics of the messages are strings indicating the type and the function of the message, in case they are different, e.g., sensor/Sensor1, or sensor/2DVideo/Reid.

1. Sensor-1 publishes a message whenever it elaborates a result for its observed property higher than some threshold.
2. The 2D Video system publishes a message from CAM-1 containing label and image of each intercepted commuter in front of Sensor-1. Moreover, the 2DVideo subscribes to receiving Sensor-1 messages to later use them to prioritize tracking and reidentification of suspect commuters.
3. The C&C Data fusion component subscribes to all messages by both sensors and the 2DVideo system, and publishes prioritized messages with current position of each commuter reaching the area covered by Sensor-2.
4. The 3D Video component subscribes to receiving messages with commuter reidentification data and priority indication, sent by the Data fusion component. These messages enable handover tracking of commuters from the 2D Video CAM2 to 3D Video camera, and support Sensor-2 detections prioritizing suspect commuters.
5. Sensor-2 publishes messages with all its detection results.

To best accomplish interoperability and security requirements, the broker interface is designed to use widely accepted and mature standards, so that a variety of off-the-shelf technologies and libraries are available to building a compliant system.

Essentially, the publish/subscribe interface is utilized by transmitting/receiving messages formatted in JavaScript Object Notation (JSON) and using the MQTT (MQ Telemetry Transport) [MQTT] application-layer publish/subscribe protocol.

To enhance the architecture flexibility with respect to the type and number of sensors that could be handled, especially in future evolutions, a general message format for sensor result submission has been defined as in Figure 9.

```json
{
    "sensorId": <string>,
    "detectionId": <string>,
    "insteadDetectionId": <string>,
    "detectionTime": <string of date-time>,
    "value": <json array>
}
```
In particular, the “insteadDetectionId” is a string representing a unique identifier attributed by the C&C to a commuter recognized by the system, and it is used by the Data fusion to correlate all messages referring to him/her. Furthermore, the value field is an array of json property objects specific for the sensor, containing one or more measurement results with confidence level (number between 0 and 1). For example, for the 2D Video system reidentification message, “value”: [  
  {“property”: “reidentification”, “type”:“complex”, “result”: {“filenameCrop”: “image_crop001.jpg”, “bbox”:[100, 71, 123, 150]},“confidenceLevel”: 0.85, “threshold”: 0.8}]. Finally, the coordinates refer to the point in the area near the sensor where detection occurs and the dataURL is the common URL path to all other produced data files, such as images and/or videos.

4.2 Data fusion

The data produced by the different sensors is collected by the C&C and processed in real time to enable situation awareness and timely alarm generation. An important task of the Data fusion component is real time association of measurements of physical properties performed by the sensors with the images of the corresponding persons, which are independently produced by the 2D Video camera devices. For the INSTEAD system, this is only the case for Sensor-1 detections, as the 3D Video system and Sensor-2 will already produce combined results correlated with each message received by the C&C.

The data fusion between Sensor-1 and 2D Video CAM-1 messages is performed by matching the values of the space coordinates, corresponding to the position of the examined person, and the detection times supplied by the two components. The space-time correlation can be achieved as all the devices are configured to using the same relative external space coordinate system and the time reference provided by the NTP server. Furthermore, the implemented data fusion function handles the case of concurrent generation of messages issued by Sensor-1 and 2D Video CAM-1, and therefore the possibility of receiving them in different orders and with time delays due, for example, to Sensor-1 higher processing time.

The data acquired by all detectors is processed by means of complex event processing techniques provided by Siddhi technology [Siddhi] in the WSO2 streaming integration environment [WSO2]. The real time requirement is addressed by in-memory data stream queries and usage of temporal windows that may be configured based on the average times for a person to reach the sensors location in the surveilled area, and the response time of the components. Some exceptional scenarios such as sensor off-line or missing data from some component are also intercepted by the C&C and labeled so to be eventually handled directly by the security.

For each detected commuter, the Data fusion will produce a message integrating all the data sent by the various components within the surveilled area and send it to the Security client.

All the messages received and produced by the C&C are persistently stored in a database for post-processing and/or usage by various types of data analytics tools.

4.3 Security client

The Security-client component creates real-time reports of interest by the security team concerning the detected commuters. In case of a suspect person, an alarm is generated and the position and the corresponding image of the person are sent to the various security applications devoted to alarm management. These may be specific mobile applications used from wearable devices by the security guards located in the area, or a desktop application installed in the security control room. Additional data produced by the sensors about the detections that are useful for alarm management can be directly queried to the components owning them.

5. CONCLUSION

This paper presented the INSTEAD system for the fusion and management of a threat-detection sensors without disrupting the flow of passengers. The INSTEAD system consists of components for 2D and 3D video tracking and C&C. The main novelties are a smart Re-ID training strategy to generalize for environments, location prediction to compensate for latency, unbiased location estimates to compensate for systematic errors, and a message broker to flexibly combine outputs of multiple sensors.
The 2D and 3D person tracking accuracy after bias correction reach the median error of less than 25cm. This can be considered as an acceptable accuracy for performing handover between the two systems. The RNN-based location prediction accuracy reaches a level that is better than overall 2D/3D tracking accuracy when the prediction time is less than one second. This indicates that our predictor can be used to increase the accuracy of pointing a correct person if system and other latencies are together less than one second.

ACKNOWLEDGEMENT

The INSTEAD (“INtegrated System for Threats EARly Detection”) project has received funding in the DEXTER (Detection of EXplosives and firearms to counter TERrorism) program from NATO Science for Peace and Security (SPS) under grant agreement number G5605.

REFERENCES