Assessment Procedure with Specific ROC Curves for Comparison of Fusion Engines

Timo Sartor, Norbert Scherer-Negenborn, Eckart Michaelsen and Klaus Jäger
Department Object Recognition
Fraunhofer IOSB
Ettlingen, Germany
timo.sartor@iosb.fraunhofer.de, norbert.scherer@iosb.fraunhofer.de,
eckart.michaelsen@iosb.fraunhofer.de, klaus.jaeger@iosb.fraunhofer.de

Abstract - Fusion Engines in general have to handle manifold types of information from different sensors. In particular, in urban terrain such diverse sensor systems as e.g. electro-optical cameras, thermal cameras, small ground radars, acoustical sensors, and chemical, biological, radiological and nuclear (CBRN) sensors can contribute information. Several approaches to the fusion of such possibly contradictory or affirming information are known. This contribution evaluates such fusion results by estimating the gain of information comparing it to the individual sensor results. Specific ROC curves are used as evaluation criterion. This procedure opens the way for a comparison of Fusion Engines in general. To this end the evaluation procedure has to be capable of handling the diverse interfaces for the ground truth and sensor data, as well as for the fusion results. Common to all of these interfaces is that all information has to be labeled by geographical location and time, but they can have additional variables such as e.g. features or classification results. Input data can be synthetic or real data. Different quantitative measures are derived for the ranking of such systems.

Keywords: Fusion Engine, track fusion, performance evaluation, assessment procedure, ROC curve.

1 Introduction

To obtain situation awareness in complex situations especially in urban terrain the use of many different sensor systems is desired. The information given by these systems can differ in many ways.

There are simple sensors that detect only events at one point in time (e.g. acoustical sensor), sensors that detect activity over a certain time span (e.g. CBRN, COMINT), and sensors that send detections and updates in a certain fixed frequency i.e. fixed framerate (e.g. cameras, radar, …). The position data given can range from 1-dimensional (e.g. COMINT direction finder) or 2-dimensional (e.g. camera) to 3-dimensional data (e.g. radar). Additional features (e.g. velocity from doppler radar) or classification labels (e.g. from CBRN) can be part of the sensor output.

The sheer amount of data combined with the diversity of information leads to the necessity of data fusion.

In literature there exist several approaches to the fusion of such information that (from the nature of its sources) can be affirming or contradictory information, containing diffusion in measurements and false measurements. The fusion of data from different sensors or the fusion of sensor data and knowledge can be processed by artificial intelligence methods such as semantic nets, production systems, etc. [4][5]. A good example for the fusion of IR and RADAR evidence using a knowledge-based method is [6]. While being flexible and applicable to diverse data and situations such approaches suffer from two downsides: 1) Heuristic modeling will often be sub-optimal in the presence of uncertainty; 2) automatic learning from examples would be preferable in a world where interrelationships are very complex and drifting. Therefore much work on the fusion issue concentrates on probabilistic formulations. In particular the fusion of information from sensors for location and mapping has been addressed probabilistically in robotics [7]. Probabilistic tracking of point targets and also of dense clusters of such objects under inclusion of diverse constraints such as kinematics or road maps etc. has been treated thoroughly [8].

This leads to the necessity of assessment of the performance of fusion algorithms or systems assembled within a Fusion Engine (FE) to make comparisons between the different approaches and the Fusion Engine as a whole.

The benefits that are generated by a Fusion Engine are on the one hand additional information that reaches beyond the data that can be provided by a single sensor. That can be e.g. 3D position from data of 2D cameras, alarms according to specific rules or risks, decision support, or other high level information, which can be used to the fusion process itself (compare [3]).

On the other hand the benefit lies in the improvement of the input sensor data (e.g. by sensor fusion, by sensor selection, …). In this contribution only the second type of improvement will be considered.

To quantify this improvement it is necessary to evaluate the single sensors as well as the Fusion Engine and to compare those results. For this comparison we use ROC
curves as assessment measures for Fusion Engines. Such ROC manifolds are estimated from training data. The issue how representative this may be for the real classification task is treated rigorously in [9].

In section 2 evaluation measures will be defined to identify the improvement in the Fusion Engine output. In section 3 the assessment procedure as a whole will be outlined. Section 4 contains the description of the practical implementation as well as examples. Some final remarks are made in section 5.

2 Evaluation Measures

The general criteria used to score the output data of the Fusion Engine as well as the output data of the single sensors will be defined and explained, the performance measures will be derived from these scores. The output data consist of tracks, which are built from detections, as well as classifications. To score the performance, the detections and classifications have to be compared with the positions and classes of the actually existing targets known from the ground truth data. In order to score a detection as hit or miss and to relate these detections to the existing targets, the targets will be represented by an area (e.g. bounding box or polygon shape) and the detections, lying inside the boundaries of that area, will be reported as an alarm for the respective target.

For scoring the classification of an object, there are in general only three possibilities: the object is classified correctly, wrongly or is not classified. In the following sections, we will explain the scoring values and the performance measures for detections, tracking and classification.

Detections and targets are not considered at each point in time but are combined to alarm tracks and target object tracks by means of tracking information. An alarm track consists of multiple temporally sequential detections generally holding one detection at each point in time. (trackbased measures)

• DT (Detected Targets)
  Number of target objects over the total sequence that are reported at any time in any alarm track. It is considered that a target can be reported by different alarm tracks at different times. No target object is counted more than once.

• RD (Redundant Detections)
  Number of alarm tracks that have reported at least one target during their lifespan minus the number of detected targets (DT).

• FD (False Detections)
  Number of alarm tracks that have reported no target during their lifespan.

• MT (Missed Targets)
  Number of target objects that are never reported in any alarm track.

In a multi-sensor scenario there can be some sensors providing a classification for objects in the scenario. The classifications are compared to the ground truth being either correct or false. Other objects have not been classified at all, since a classifier can have the option to refuse to classify an ambiguous object and not all objects are presented to a classifying sensor in any scenario.

• CC (Correctly Classified Targets)
  Number of targets that have been assigned a classification that coincides with the correct classification from ground truth.

• FC (Falsely Classified Targets)
  Number of targets that have been assigned a classification that does not coincide with the correct classification from ground truth.

• notC (Not Classified Targets)
  Number of targets that have not been assigned any classification.

The Fusion Engine gathers the information of the different sensors, which can contain distorted, false or redundant data. The challenge to the Fusion process is that it has to lead to an improvement in the data in some ways.

In the Fusion Engine the assembly of the detections from each single sensor should lead to a higher probability of detection of the objects in the scene.

On the other hand false detections by each single sensor are accumulated in the Fusion Engine’s input and must be reduced during the fusion process.

From one or especially many sensors the Fusion Engine receives different tracks that follow the same target, the number of these tracks should be reduced by the Fusion Engine.

Further the tracks following a certain target should be combined by the fusion process to lead to a single track following one target for a longer time.

Classification labels assigned to the target objects by the different sensors can be affirming or contradictory. These conflicts have to be solved by fusion and reduction to at most one classification label per target.

From those considerations the following measures were chosen to indicate improvement of the data by the Fusion Engine.

Detection Probability: The number of detected objects divided by the number of objects from ground truth.
\[ P_d = \frac{DT}{DT + MT} \]  

**False Alarm Rate:** The number of detections that do not coincide with any object from ground truth in a certain amount of time.  
\[ FAR = \frac{FD}{time} \]  

**ID-change:** For each object in ground truth the number of different track IDs that coincide for at least one point in time with the ground truth.  
\[ ID(Obj) := \{ id \mid track[id] \rightarrow Obj \} \]  
\[ IDC(Obj) = \# ID(Obj) \]  

**Longest track segment:** For each object in the ground truth the track is chosen that coincides the longest time with the ground truth (possibly with interruptions). The time where track and ground truth coincide is measured. Meaning the time where the track of a sensor and the ground truth coincide is measured as difference of the first and the last detection, where the trackpoint lies inside the ground truth area, subtracting eventual times, where the trackpoint lies not inside the ground truth area (gaps or spatial discrepancy).  
\[ LTS(Obj) = \max_{id \in ID(Obj)} \{ Time(track[id] \rightarrow Obj) \} \]  

**Correct Classification Probability:** The number of the correctly classified objects divided by the number of all objects.  
\[ P_{cc} = \frac{CC}{CC + FC + notC} \]  

### 3 Assessment procedure

The performance evaluation is measured by a comparison with ground truth data. The evaluation is conducted based on a general Performance Evaluation Procedure by NATO Task Group SET-070/RTG-38 [1]. This procedure will lead to values for the quality of the data produced by FE compared to the ground truth data. The performance evaluation procedure for a Fusion Engine is shown schematically in Fig.1. The cycle in this procedure is described below.

First, the role of the FE system has to be determined, the requirements, applications and application fields of the system have to be identified (**Define Application / Requirement**). Then the criteria for the assessment of the FE system have to be defined. In this step, the performance measures have to be described and the criteria for scoring the Fusion Engine system have to be established (**Define Assessment**). In the next step realistic scenarios have to be assembled (**Design Collection**) and collected this is either done in a simulation or in an (**Execute Collection of Data and Ground Truth**). After partitioning the data in test and training data (**Partition Measured Data**), the algorithms of the Fusion Engine can be optimized with the training data and afterwards tested with the test data (**Test Fusion Engine**). During the test, the values of the scoring variables (e.g. number of detected targets) will be computed (**Score Fusion Engine**). The performance measures can be calculated from these variables (**Calculate Performance Measures**). The measures include values and ratios describing the quality of the test results such as Probability of Detection, False Alarm Rate, Probability of Classification, etc. The last step is to conduct an analysis of the data with respect to these measures to identify the results of the test and assess the performance of the system (**Conduct Analysis / Assessment**).

The results of the above procedure alone will heavily depend on the quality of the sensor data given as input to the Fusion Engine. For example, if all the aligned sensors miss a target, the Fusion Engine is unable to detect it as well. We don’t want to measure the influence of the input data on the Fusion Engine result (performed in [2]), but we want to measure the gain of information generated by the Fusion Engine to the given data. Therefore we conduct an Assessment procedure in three steps.

![Fig. 1: Schematic overview of the performance evaluation procedure of a Fusion Engine.](image-url)
The three steps of the Assessments Procedure depicted in Fig. 2: Evaluation of the single sensor output (that is used as input to the Fusion Engine), Evaluation of the Fusion Engine output (comparable to each single sensor), Comparison of the results in a ROC curve.

3.1 Evaluation of single sensor output

Each single sensor is evaluated individually. The ground truth data to all objects in the scenario has to be available in sensor coordinates, even at times, where the sensor cannot detect the object (e.g., object behind a building for an EO camera). This data is necessary, because the FE output could contain the data, thus representing an improvement.

3.2 Evaluation of the Fusion Engine output

Fusion Engines usually present the output in a common 3-dimensional coordinate system. To make the comparison of the results of the Fusion Engine to the (single sensor) input data it would not be fair to use the (common unified 3D) coordinate system of the Fusion Engine to compare it to the ground truth data. Accurate 2D position as input from an optical sensor can lead to inaccurate 3D position output in the Fusion Engine.

Therefore the output of the Fusion Engine has to be projected into the specific coordinate system of each sensor.

The comparison to the ground truth is then performed in the same way as for the single sensor output.

3.3 Comparison of the results in a ROC curve

The results of the Fusion Engine (FE) evaluation and the single sensor evaluation lead to 2-tuples: [sensor result, FE result] for each: Detection Probability, False Alarm Rate, target ID-Change, target Longest Track Segment, and Correct Classification Probability. The values are displayed in a specific ROC curve that has the sensor result as the first axis value and the FE result as the second axis value (see Fig. 3).

Fig. 2: Sketch of the three steps of the Assessment procedure

Fig. 3: ROC curve illustrating the evaluation result of the single sensor and of the fusion engine.

For an assessment all the sensors have to be evaluated, and more test runs have to be performed.

4 Implementation of the procedure

The assessment procedure will be implemented on a multi-sensor Fusion Engine that is currently developed by a European consortium within the project “Distributed and Adaptive multisensor Fusion Engine” (DAFNE), financed by the European Defense Agency (EDA). For further details on the Fusion Engine developed within this project see [12]-[17].

A tool was developed to compute the evaluation measures for single sensors from the sensor output and corresponding ground truth data. The same tool is used to compute the measures for Fusion Engine given the FE output projected to the sensor coordinates and the same ground truth data as for single sensor evaluation.

The sensor output, Fusion Engine output, and ground truth data is given in an XML format.

4.1 Input data

The data generated by the sensors is assembled in an XML format. The structure was chosen to adapt easily to the different data fields that can be generated by different sensor types. Each observation consists of a timestamp and a location in sensor coordinates, further it has to contain information about the sensors position and orientation in a common unified 3D coordinate system.

The format is applicable for both simulated and real data. For reliable results of the performance of a Fusion Engine the use of real data has a major relevance. On the other hand the acquisition of real data and especially of the corresponding ground truth data is elaborate and
costly. Therefore, mostly in the early development stages of a Fusion Engine, simulated data is used.

Example 1: A scenario with sensors of similar type and with similar parameters:
The evaluation results for the individual sensors are very similar and can be averaged. By changing the parameters of all sensors the results generate a classical ROC curve (see Fig. 5).

Example 2: Different scenarios with the same sensors:
The evaluation results vary in the individual testing scenarios. The statistical distribution is depicted as ROC area with mean value and variance (see Fig. 6).

Example 3: Sensors with different evaluation results in one scenario:
The different results of all sensors are plotted individually. Since the false alarm rate (FAR) of the Fusion Engine is the same in one scenario, the points belonging to all the sensors in one scenario lie on one horizontal line in the ROC space (see Fig. 7).
Final remarks

We have presented a procedure to give an assessment to a Fusion Engine. This procedure is intended for assessing different Fusion Engines and making a comparison between the performances of these Fusion Engines.

This contribution is part of an ongoing project and ongoing research. We are about to execute our assessment procedure on the Fusion Engine developed within the DAFNE project.

Acknowledgements

The authors would like to thank the European Defense Agency (EDA) for their financing of the DAFNE project and thus this research.

Further we want to thank the partners in the DAFNE project: IDS, University of Udine, TNO, Warsaw University of Technology and VOP-026 Šternberk for their collaboration and fruitful discussions.

References


