D2.3 – Metrics and Experiments for V & V of the driver, vehicle and situation models in the 2nd cycle

Project Number:	690705
Classification	Public
Deliverable No.:	D2.3
Work Package(s):	WP2
Document Version:	Vs. 8
Issue Date:	31.10.2017
Document Timescale:	Project Start Date: September 1, 2016
Start of the Document:	Month 12
Final version due:	Month 14
Compiled by:	Sebastian Feuerstack, OFF
Authors:	Mark Eilers, HMT Sebastian Feuerstack,OFF Alain Giralt, CAF Maximilian Graf, ULM David Käthner, DLR Adam Knapp, BIT Paulin Pekezou Fouopi, DLR Jürgen Pichen, ULM Fabio Tango, CRF
Technical Approval:	Fabio Tango, CRF
Issue Authorisation:	Andreas Lüdtke, OFF

DISTRIBUTION LIST		
Copy type[1]	Company and Location	Recipient
T	AutoMate Consortium	all AutoMate Partners

[1] Copy types: E=Email, C=Controlled copy (paper), D=electronic copy on Disk or other medium, T=Team site (Sharepoint)

RECORD OF REVISION		
Date	Status Description	Author
01.08.2017	Deliverable structure	S.Feuerstack (OFF)
14.09.2017	Integration of first contributions sections 2.2+2.3	S.Feuerstack (OFF)
29.09.2017	Integration of contributions from CAF and BIT	S.Feuerstack (OFF)
11.10.2017	Integration of contribution from ULM	S.Suck (OFF)
23.10.2017	Integration of contributions from ULM, CAF, and CRF	S.Feuerstack (OFF)
31.10.2017	Final Version	S.Feuerstack (OFF)

Table of Contents

1 Introduction

WP 2 is concerned with monitoring, understanding, assessing and anticipating the driver, the vehicle and the traffic situation. For doing so, a large variety of models are being designed and built by the project partners in this WP. Model development is organized within three cycles. This Deliverable initiates the second cycle for verification and validation (V&V) of the models. For each cycle metrics are identified or refined that can be used to measure the quality and progress of the model constriction. Whereas for the first cycle we focused on individual component testing (partners therefore performed internal testing), for the second cycle also initial component integration will be performed to evaluate and assess the current state of the overall automate approach.

Due to the wide variety of models in WP2, different approaches to V&V are required. Thus, some models are validated based on divided training data sets, others are validated in driving simulator experiments, or by functional software testing.

The upcoming main chapter of this document therefore focuses on the metric definition for each component: First the basic component functionality is presented, followed by a description of the evaluation method. Finally, the individual component metrics are discussed and a brief outlook is given to the current state of the experiment and integration planning.

2 Metric Definition

In general, every model which is developed in AutoMate is validated with empirical or generated data which is compared to the data predicted by the model. In AutoMate several models are developed, which have been described in section 1 of deliverable 2.1.

In AutoMate a four-step process of the model validation is used:

1. Step: Describe what the goal of the model is (c.f. D2.1, section 1,).
2. Step: Define general metrics, which are related to the goals of the models (c.f. D2.1, section 3).
3. Step: Description of the purpose of the experiments with regards to the model validation.
4. Step: Specify criteria for deciding whether the validation was successful or not.

This chapter therefore focuses on the detailed metrics for each of the three enablers of WP2. All are targeted to Objective 2 of the project.

2.1 Driver State Model and V2X Communication

2.1.1 Driver State Model

The driver's state model is a SW module, which provides the following the models: Drowsiness, Visual attention, Visual distraction, and Cognitive distraction (possibly, under investigation). The main input are provided by a camera based system, which scans the driver's face to extract information such as eyelid movements, head gaze, eye gaze, and facial patterns.

2.1.1.1 Drowsiness

2.1.1.1.1 Drowsiness model overview

The model output are the following:

- **Drowsiness states:**

States	Rule of detection	Description
Alert	The driver has very few long blinks and very few very long blinks;	Driver is alert; no sign of drowsiness; KSS: 1 to 5
Slightly drowsy	The driver has few long blinks and very few very long blinks.	First signs of drowsiness; Driver should only be informed; KSS: 6 to 7
Drowsy	The driver could have some long blinks and few very long blinks or simply some very long blinks.	Driver is drowsy; Fighting sleep; Degradation of his/her driving performances; Driver must stop and take a rest. KSS: 8
Sleepy	The driver has some sleepy blinks	Driver is almost falling asleep; Critical state; Driver must stop urgently. KSS: 9
Sleeping	Eye closed for a long duration when the driver is DROWSY or SLEEPY	Driver falls asleep; requires an instantaneous wake up action. KSS: 10

- **Drowsiness Quality:**
  The confidence rate of the Drowsiness level [0 1].
- **Eye closed status:**
  True when both eyes are detected closed for more than 100ms.
- **Eye closed status Quality:**
  The confidence rate of the Eye closed status [0 1].

2.1.1.1.2 Evaluation method

The evaluation is done by computing performance indicators comparing model outputs to sleepiness expert ratings considered as ground truth. The ratings are done using the Karolinska Sleepiness Scale (KSS), ranging from 1-10.

The expert ratings are done by an expert monitoring constantly the driver. The expert annotates the driver's KSS level every 5 minutes. The drivers also evaluated themselves on the KSS scale every 10 minutes. There are several situations where the driver rates himself higher compared to the expert. It is difficult for the expert to rate the driver's state perfectly as an external observer. An additional expertise based on the recorded video could be used to refine the expert ratings.

The method compares the drowsiness model output data to ground truth data. The evaluation of diagnostic and eye-closure are independent where the diagnostic evaluation is based on log files and eye-closure on visual annotation. The model availability is the time where the Drowsiness confidence was superior to a define threshold. For all other situations the DLWE drowsiness diagnostic was considered to be Alert. The database is divided in events. Expert Events is the period of time between two expert ratings. An expert events lasts about 5 minutes as expert rates every 5 minutes.

A DSS event is a period of time, rated in one of the following events: Drowsy, Sleepy or Sleeping. SDrowsy events are not included in this category. A DSS event begins when a drowsy or higher states is detected and its quality is above a defined threshold while the ground truth was rated KSS 7 or below. The event ends when the model returns to SDrowsy/Alert or if the ground truth rating changes to KSS 8 or above.

In order to compute statistical figures a confusion matrix of the model output versus the expert ratings is created. As expert events last for about 5 minutes, the maximum drowsiness level during the events was compared to the ground truth. The events are stored in the confusion matrix where, for each expert event, the cell that corresponds to (MAX_DLWE, GroundTruth_KSS) is incremented.

State \KSS	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Alert											
Sdrowsy											
Drowsy											
Sleepy											
Sleeping											

Figure 1, DLWE / Expert Confusion matrix

The database is divided into an ASD database that combines Alert and SDrowsy states (KSS 1 to 7) and the DSS database that combines Drowsy, Sleepy and Sleeping states. This division is done to separate strong drowsy states from weak ones. The driver should be informed of the occurrence of a strong drowsy state while weak ones may only be used for making the detection of the strong ones more robust.

2.1.1.1.3 Metrics

- **Output availability:**
  Ratio of time the quality of model output is above a define quality threshold
- **Number of False DSS events per hour:**
  Numbers of events diagnosed Drowsy, Sleepy or Sleeping per hour of the ASD Database (non DSS), including questionable expert ratings.
- **Detection rate of DSS events:**
  Detected DSS events divided by number of ground truth DSS events.
- **Detection rate of Drowsy (KSS 8) events:**
  Detected Drowsy events divided by number of ground truth Drowsy events.
- **Detection rate of Sleepy (KSS 9) events:**
  Detected Sleepy events divided by number of ground truth Sleepy events.
- **Detection rate of Sleeping (KSS 10) events:**
  Detected Sleeping events divided by number of ground truth Sleeping events.

2.1.1.2 Visual attention/distraction

2.1.1.2.1 Visual attention/distraction model overview

Visual distraction is a type of distraction which occurs when driver take their eyes off the road. Typically this occurs when the driver looks away from the road to engage in a secondary activity either inside (e.g. central display, radio, smartphone on his knees, kids on the rear seat etc.) or outside (e.g. through the lateral window etc.) of the vehicle.

The function provides a primary output of the ID of the area the driver is looking at. Based on the distribution over time of these primary data to the module derives the level of attention and finally an attention state declined in three states: attentive, partly attentive and distracted.
The figure below show examples of different head poses performed while driving.

Figure 1: Head positions while driving

2.1.1.2.2 Evaluation method

The primary output shall be tested according to a defined protocol in which the driver will be asked to look at specific area of the vehicle (e.g. rear view mirror, radio, lateral rear view mirror etc.). The driver will also be asked to look at various "on road" areas in order to check the algorithm robustness to false alarms. This protocol will be performed first in laboratory conditions on a significant set of representative drivers in various light conditions, then in driving ones on a reduced set due to the complexity to carry on such test in safe conditions.

The assessment of the "Attention state" performance is more complex to evaluate as it is a subjective notion, which is also related with the driving scenario. A subjective scale must be defined in agreement with the demo car owner and human factor teams. A similar scale to the KSS one could be used. As KSS it could be a 10 level scale ranging from fully attentive to fully distracted.

Same as for drowsiness, attention event is a period of time, rated in one of the state: attentive, partly distracted or distracted.

In order to compute statistical figures a confusion matrix of the model output versus the expert/driver ratings is created. The duration of expert/driver event have to be defined. The maximum attention level during the events is compared to the ground truth. The events are stored in the confusion matrix where, for each expert event, the cell that corresponds to (MAX model, Attention_expert_scale) is incremented.

State \Attention scale	0	1	2	3	4	5	6	7	8	9	10
Attentive											
Partly distracted											
Distracted											

Confusion matrix

2.1.1.2.2 Metrics

Primary output (area ID)

The used metrics are the classical ones:

- **Availability per hour:**
  Ratio of time the quality of model output is above a define quality threshold.
- **Detection rate per ID:**
  Detection rate will be evaluated for each tested area (e.g. lateral mirror, radio, smartphone, etc.).
- **Number of False detection per hour:**
  Numbers of events diagnosed with the incorrect ID per hour.

Attention state

The used metrics are the classical ones:

- **Output availability:**
  Ratio of time the quality of model output is above a define quality threshold.
- **Number of False “Partly Distracted” and “Distracted” events per hour:**
  Numbers of events diagnosed “partly distracted” or “distracted” while expert/driver rating says “Attentive”.
- **Detection rate of “Distracted” events:**
  Detected “distracted” events divided by number of expert/driver “partly distracted” rated events.

2.1.1.3 Upcoming Experiment/Study/Test (2nd Cycle)

CAF proposes a hardware and a software components for the distraction and the drowsiness detection. The aim is to test the detector from CAF on the VED demonstrators. The software version will be installed on the Smarteye hardware that is connected to the VED simulator. The software exploits the smart eye cameras and outputs in order to detect the two factors. The second version will be integrated in the VED demo car.

The tests will be operated on more than twenty (20) participants. These persons will be involved in the experiment on the demo car on open road and on the simulator on highways scenarios.

With reference to the verification/validation of the CAF component in REL/CRF demo, this will happen in the third cycle of the project, during the integration phase of this component, in correspondence to the HMI tests (described in WP4). In particular, these tests consider the distraction aspect and the following figure can show a sketch of a possible experimental configuration:

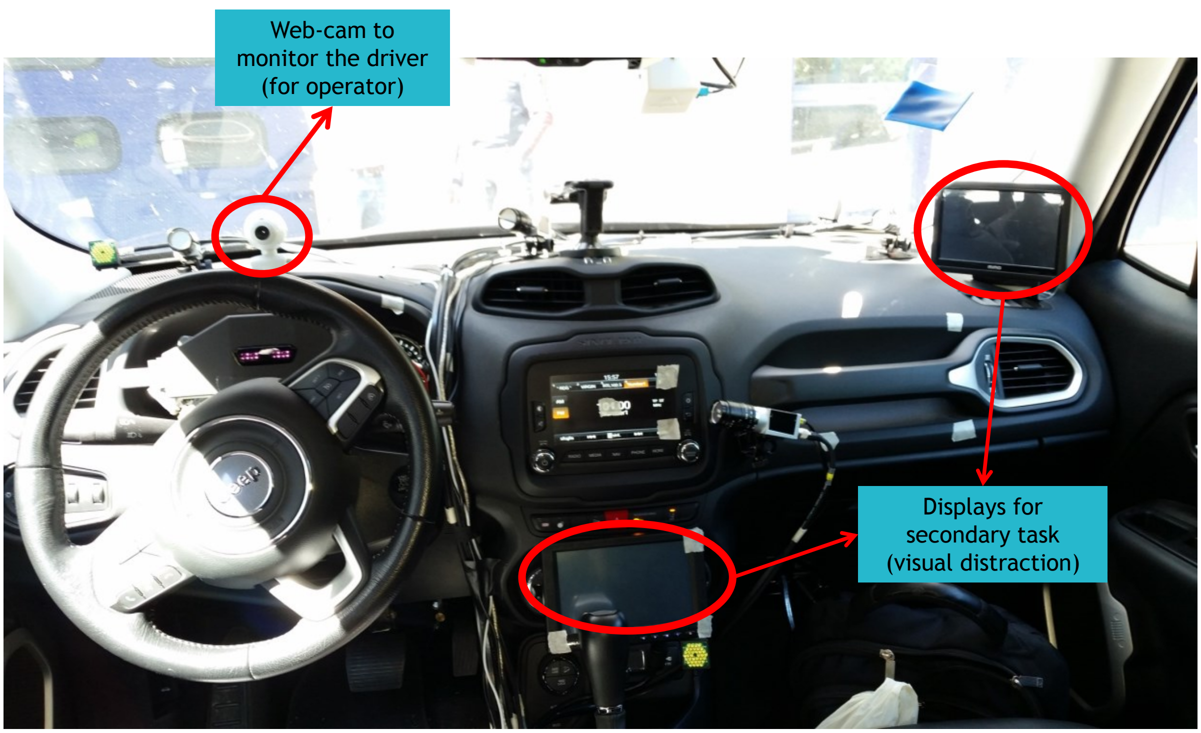

Figure 2: cockpit experimental installation

ULM will conduct experiments to validate the driver state model developed by CAF. CAF and ULM will work together to include the driver state model algorithms into ULM's driving simulator. The hardware component of the driver state recognition will not be installed in the simulator – the already built-in static eye tracking system (Smart Eye) will provide the necessary data for the algorithms, which then will provide the driver state to the simulation software (SILAB). To validate the driver state model, the test subjects will drive on a test track where they drive through use-cases of the PETER scenario. Throughout the experiment the test track will be designed more and more "boring" for the test drivers. This "boringness" should lead to a drowsier or more distracted driver state, which will be captured by the driver state model algorithms and recorded as well. It is expected that the more distracted or drowsy the driver is, the worse his or her reaction time and other recorded parameters will be. These parameters also include driving mistakes and crashes.

It is to worth to mention here that such an experimental configuration is valid independently if the tests are carried out on driving simulator or on a real-vehicle.
The total number of subjects will be between 20 and 30 users, which will be asked to drive on a dedicated driving session, in (real) traffic situations, while completing a secondary task session. This consists in reading a sequence of random letters, displayed on one of the two secondary screen (lateral or bottom, selected randomly). The necessary time to read the letters sequence is about 2s (as defined by literature), in such a way that subject's eyes were out of the road for 2s.

2.1.2 V2X Communication

V2X communication is one of the key enablers of future transport systems. It is well standardized, but it is still in emerging phase. V2X promises more efficient traffic and reduced accident rate by utilizing the cooperation of vehicles (V2V communication) and the infrastructure (V2I communication). From other point of view V2X can be handled as another sensor of the vehicle, which provide many input data for the car, but it senses different kind of object and from different distances as LIDAR, RADAR etc. The information from V2X communication are used for enhancing the decisions of the embedded intelligence.

The V2X communication is technically based on Mobile Ad-hoc Network (MANET). Therefore, the continuity of connection is very various meaning that it does not provide reliable and permanent communication channel. This issue is handled by the IEEE 802.11p standard, which is the physical and MAC layer of the ETSI ITS protocol stack. The network layer includes the GeoNetworking protocol, which was developed to fulfil specific requirements of vehicular environment. It includes such important feature like broadcasting messages in a designated geo-area. This functionality utilized by the facility layer of the protocol stack, where standardized messages for carrying specific information are available. For the scenarios of AutoMate project two facility layer messages are important:

- the Cooperative Awareness Message (CAM) provides by means of periodic sending of status data (basic status, position, current speed etc.) a cooperative awareness to neighbouring nodes [1];
- Decentralized Environmental Notification Message (DENM) allows for broadcasting useful information related to road traffic conditions. DENM serves as input for Road Hazard Warning (RHW) application, which is an active road safety application that is distributed among vehicles ITS station and roadside ITS stations [2].

The following figure depicts the ETSI ITS protocol stack:

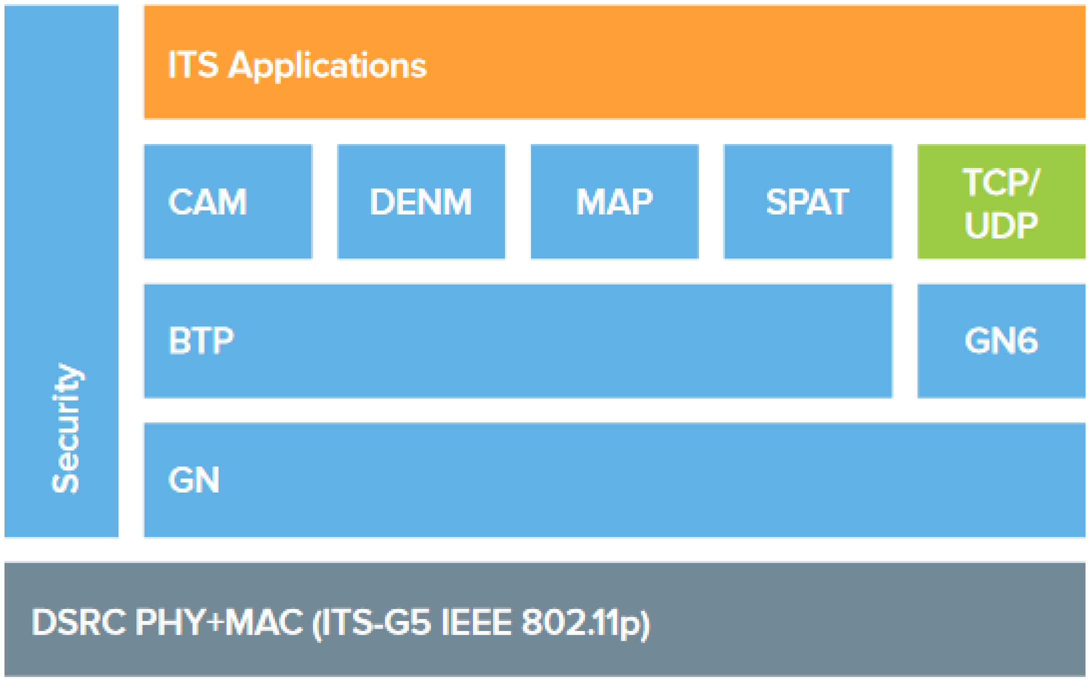

Figure 3. ETSI ITS protocol stack

Since V2X communication is already standardized, and 3rd party tools are available in the market, which are conform with these standards, we focus on high-level test cases and experiments to validate the benefits of the technology for the AutoMate project.

2.1.2.1 Verification of V2X communication

- The communication is established between two V2X capable components. The transmitted information is received by the vehicle.
- Fulfils the following requirements: V2X COMMUNICATION, V2X CAPABLE PARTNERS
- Metric: yes/no

2.1.2.2 Verification of DENM message communication

- The communication is established between two V2X capable components. The transmitted DENM messages from the Road Side Unit (RSU) is received by the vehicle's On-Board Unit (OBU). The message have to be assembled properly.
- Fulfils the following requirements: V2X COMMUNICATION, V2X CAPABLE PARTNERS, RELEVANT V2X INFORMATION
- Metric: yes/no

2.1.2.3 Verification of CAM message communication

- The communication is established between two V2X capable components. The transmitted CAM messages is received by the vehicle's On-Board Unit (OBU). The message have to be assembled properly.
- Fulfils the following requirements: V2X COMMUNICATION, V2X CAPABLE PARTNERS, RELEVANT V2X INFORMATION
- Metric: yes/no

2.1.2.4 Verification of DENM messages

- The received DENM messages are relevant information about the traffic situation, i.e. it contains information about roadworks ahead.
- Fulfils the following requirements: V2X COMMUNICATION, V2X CAPABLE PARTNERS, RELEVANT V2X INFORMATION
- Metric: yes/no

2.1.2.5 Verification of CAM messages

- The received CAM messages are relevant information about the vehicle status, i.e. it contains information about vehicle's current position, speed and heading.
- Fulfils the following requirements: V2X COMMUNICATION, V2X CAPABLE PARTNERS, RELEVANT V2X INFORMATION, DATA FOR CAM
- Metric: yes/no

2.1.2.6 Validation of DENM message reception

- The DENM messages are received properly:
  - without loss within 200 meters, i.e. all transmitted DENM message are received;
  - only with low-level of jitter, i.e. the differences of reception times are below 10% of the transmission period
- Metric:
  - number of received and sent message
  - percentage rate of jitter

2.1.2.7 Validation of CAM message reception

- The CAM messages are received properly:
  - without loss within 200 meters, i.e. all transmitted DENM message are received;
  - only with low-level of jitter, i.e. the differences of reception times are below 10% of the transmission period
- Metric:
  - number of received and sent message
  - percentage rate of jitter

2.2 Probabilistic Driver Modelling and Learning

2.2.1 Probabilistic Driver Models for Intention Recognition and Behaviour Prediction

The probabilistic driver model for intention recognition and behavior prediction is a Dynamic Bayesian Network that models the causal and statistical relations between the driver's intentions, driving maneuvers resp. behaviors, the lateral and longitudinal vehicle controls and the situational context, as observable by the TeamMate vehicle's sensor platform.

In the second cycle of AutoMate, the development of the model will be focused on the Peter scenario, dealing with overtaking scenarios on rural roads. The purpose of the model in the Peter scenario is to constantly provide the TeamMate vehicle with an online recognition of the current intentions of the driver and show driving behaviors of the TeamMate vehicle. If the driver is in control of the TeamMate vehicle (manual driving), the information provided by the model will be used to assess the safety of the intended driving maneuver. If the automation is in control of the TeamMate vehicle (autonomous driving), the information provided by the model will serve as a mechanism to propose appropriate maneuvers to the automation that reflect the potential intentions of the driver for the given situation.

For modelling purposes, the model will attempt to recognize the intended target lane, i.e., whether the driver wants to travel on the right or left lane, represented by a discrete random variable I and a set of driving maneuvers resp. behaviors, represented by a discrete random variable B. Given evidence about the lane the TeamMate vehicle is currently travelling in, represented by a discrete random variable L, these target lane intentions can be transferred to actual overtaking or lane change intentions. For example, if the TeamMate vehicle is currently travelling on the right lane, the intention to travel on the left lane implies an intention to change to the left lane, resp. to overtake.

At each point in time t, the model will be used to infer a probability distribution over the target lane intention $P(I^t|l^{1:t},\boldsymbol{o}^{1:t})$ and driving maneuvers resp. behaviours $P(B^t|l^{1:t},\boldsymbol{o}^{1:t})$, given all available sensory input obtained thus far. In this context, validation of this model will rely on comparisons of the model output with empirical data in form of multivariate time series.

2.2.1.1 Metric 1

The model will be validated on a test set D_{Test}. Let D_{Test} be composed by a number of m trials, where each trial is a time-series consisting of a number of $n_j, j = 1, \ldots, m$ samples $d_j^k = \left(i_j^k, b_j^k, \boldsymbol{o}_j^k\right), k = 1, \ldots, n_j$ data samples, prior annotated by experts with the assumed correct intention i_j^k and behaviour b_j^k. For each sample d_j^k, we use the model to infer a probability distribution over the intentions $P\left(I_j^k | \boldsymbol{o}_j^{1:k}\right)$ and behaviours $P\left(B_j^k | \boldsymbol{o}_j^{1:k}\right)$ given all available sensory input in the resp. time-series up to the sample. The output of the model is then defined as the most probable target lane intention

$$i_{j,out}^k = \arg\max_{\mathrm{i}} P\left(I_j^k = i | \boldsymbol{o}_j^{1:k}\right)$$

and behaviour

$$b_{j,out}^k = \arg\max_{\mathrm{b}} P\left(B_j^k = b | \boldsymbol{o}_j^{1:k}\right).$$

For the assessment of intention recognition, the (annotated) "true" and predicted target lane intentions i_j^k and $i_{j,out}^k$ are first mapped onto actual lane change intentions (in that a lane change intention is present if the current lane and the target lane intentions differ) by defining $\hat{\imath}_j^k = \mathbf{1}\left(l_j^k \neq i_j^k\right)$ and $\hat{\imath}_{j,out}^k = \mathbf{1}\left(l_j^k \neq i_{j,out}^k\right)$, where $\mathbf{1}$ denotes the indicator function. Interpreting the existence of a lane change intention as positive and the absence as negative, we can construct a binary confusion matrix as shown in Figure 4.

		Ground Truth	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Figure 4: Binary confusion matrix. In the case of driver models for intention and behavior recognition in AutoMate, the ground truth is based on a manual annotation of test data.

The actual metric used is called the accuracy and is defined as:

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}.$$

For the overall assessment of the behavior recognition, we interpret a correctly recognized behavior $b_j^k = b_{j,out}^k$ as positive and an incorrectly recognized behavior $b_j^k \neq b_{j,out}^k$ as negative to obtain a binary confusion matrix and subsequently calculate the accuracy.

The accuracy can be used to verify that the requirements R_EN2_model1.10 for intention recognition and R_EN2_model1.11 for behavior recognition are fulfilled. More specifically, we regard R_EN2_model1.10 as fulfilled if the accuracy for intention recognition is above 80% and resp. regard R_EN2_model1.11 as fulfilled if the accuracy for behavior recognition is above 80%.

2.2.1.2 Upcoming Experiment (2nd Cycle)

The test set D_{Test} will be obtained from the experimental data conducted for training and evaluation of the probabilistic driver models in the first cycle. More specifically, the experimental data will be split into a training set D_{Train}, including approx. 70% of the experimental data, and a test set D_{Test}, including the remaining experimental data. The driver model will be learned exclusively using the training data D_{Train} and subsequently validated on the test set D_{Test}. For this, both data sets need to be manually inspected by experts to annotate each data sample with the currently assumed intention and shown driving manoeuvre resp. behaviour of the driver.

Within the second cycle, the test procedure can be repeated for a number of different data sets provided by the demonstrator owners to evaluate the generality of the model.

2.2.2 Task Model for Driving

DriveGOMS is a task analytic approach which aims at modelling driver behaviour in safety critical situations. It does so by breaking down the driving task into smallest behavioural units, which is called operators. It is based on the principles of the GOMS framework (Card, Moran & Newell, 1983). By combining the operators according to goals present in a given situation, driver behaviour can be sketched on the level of an individual operator. This allows for estimating task execution times as well as a nominal attention and gaze allocation. Within AutoMate, we will use this approach to model the empirical studies conducted within the Peter use case.

The validity of the resulting models can be established mainly

- by predicting the gaze allocation to a specific area of interest (AOI) for a specific situation, or over a series of events,
- by applying expert judgement,
- and by comparing the sequence of operators with empirical results from the driving simulator experiments.

Appropriate metrics for the model validity therefore are:

- accuracy of predicted fixations to AOIs,
- correlation of predicted sequences of operators with empirical operator sequences,
- occurrence of predicted operators in a given situation.

2.3 Probabilistic Vehicle and Situation Modelling

The Vehicle and Situation Modelling Module (VSMM) represents and predicts the situation around the vehicle, including dynamic objects. Provided with an environment model from the perception layer, it enriches that model semantically and infers the permissible driving manoeuvres in a given situation with respect to the traffic rules. The environment can be represented by a grid map, which is also common in the robotic domain. In this case, space is divided into cells, each cell having a probability of occupancy. In addition, if a cell is occupied by a dynamic object, there is also information about its speed etc. included. The grid may be used to plan paths through the environment, but other information may be extracted therefrom too. Another possibility of the environment representation is the so-called multi-object tracking. A list of dynamic objects is provided. Each of these objects is first assigned to a certain class (vehicles, pedestrians, etc.). Subsequently, the movement of each object is predicted and combined with the corresponding sensor measurements. This procedure can, for example, be carried out with aid of a Kalman filter.
In a second step, the VSMM predicts the future evolution of the situation based on the enriched situation model and the dynamic of detected objects.

2.3.1 Semantic Enrichment Sub-Module (SESM)

2.3.1.1 Description

In this section, the metrics used to determine the quality of the semantic enrichment sub-module (SESM) are described. To understand the meaning of these metrics, let us sketch the purpose and the architecture of the SESM.

The perception layer of the TeamMate car provides the dynamic objects of the environment, e.g. in the form of a list. Further, from a map the topology of the situation will be available. In a sense, this provides a parsed environment. However, such a parsed environment does not contain meaningful relations between its objects. As human observers, we are used to have an instant access to the semantics of a situation, but this is just because we are unaware of the continuous semantic enrichment our cognition provides.

An important aspect of such semantics is permissible actions of traffic participants, such as if it is allowed to cross a red traffic light. Another aspect is physical constraints: A bicyclist will not instantly turn around 180° and head into the opposite direction. Knowledge of allowed or even possible actions greatly diminishes the set of future situation developments that must be considered in an attempt to predict the future situation. It is therefore an indispensable component of an efficient situation prediction.

This is what the SESM provides: a semantic for the traffic situation. This comprises:

- relations between scene elements, e.g. the relation assigning a vehicle to a lane or a traffic signal to a lane;
- inferences regarding traffic participants manoeuvres with respect to traffic rules.

2.3.1.2 Evaluation metric

Testing the SESM is part of the normal development cycle. Generally, it will consist of a stream of environment data which the sub-module reads and processes. The set of permissible driving manoeuvres will be the output. Determining if the output is correct or not will be done essentially by expert judgement. The metric used will be therefore the ratio of correctly identified manoeuvres called “true positive rate”. An acceptable ratio will have to be determined on the component level, but should be in the range of over 90% as specified in the requirement “*R_EN3_model1.3*”.

The experiments for the SESM will be determined by the requirements for the test data. Two approaches will be taken:

1. Use of a constructed data set: To retain control over the data set, in a first step a data set with known objects and object relations will be constructed.
2. Use of a data set from the field: To determine the error free functioning of the SESM, the second step will consist of data from the environment in which the module will be used.

Both approaches imply the challenge to label the manoeuvre predictions as correct or incorrect. However, the final decision on how to conduct this labelling has not been made yet.

2.3.1.3 Upcoming Experiment

To measure the metric mentioned in sub-section 0, two experiments will be done. The first experiment will deal with manually generated test data. In the second experiment, data from the field will be used for testing the sub-module. For integration purpose, the sub-module interface will be specified and implemented. The SESM sub-module provides an input interface to the perception layer module and an output interface to the situation prediction sub-module. For this cycle, we propose a client/server based interface, where the server interface provides output to other modules. The client interface reads the output of the server interface and provides it as input to the corresponding module. Interfaces exchange date among each other using Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) packets. The plan for the next steps is described in the following table.

Task description	Due Date	Responsible
First version of the interface specification	30.09.2017	DLR/ULM/HMT
Description and collection of data needed for the first experiment	30.09.2017	DLR/ULM
Implementation, integration and test of the	31.10.2017	DLR/ULM

first interface version as well as evaluation of the SESM sub-module on first experiment data		
Second version of the interface specification	15.11.2017	DLR/ULM/HMT
Description and collection of data needed for the second experiment	15.11.2017	DLR/ULM
Implementation, integration and test of the second interface version as well as evaluation of the SESM sub-module based on the second experiment data	15.12.2017	DLR/ULM/HMT
Documentation in D2.4	30.12.2017	DLR/ULM/HMT

2.3.2 Prediction *of the future evolution of the situation*

The purpose of the prediction of the future evolution of the traffic situation is to provide the TeamMate vehicles with a temporal-spatial prediction of the state of other traffic participants. Such a prediction is required by the online risk assessment to derive a safety corridor in which the TeamMate vehicle can maneuver safely and to assess the safety of potential maneuvers.
Let $\boldsymbol{V} = \{v_1, \dots, v_{n_V}\}$ denote a set of objects (usually traffic participants) detected by the sensor platform of the TeamMate vehicle. For the prediction of the future evolution of the traffic situation, we assume that at each point in time t, the sensor platform provides a belief state $p(\boldsymbol{X}_v^t|\boldsymbol{o}^{1:t})$ for each $v \in \boldsymbol{V}$, estimated from the sensor observations received up to the current point in time $\boldsymbol{o}^{1:t}$, where $\boldsymbol{X}_v^t = \{X_v^t, Y_v^t, \Theta_v^t, V_v^t, A_v^t, W_v^t, S_{L_v}^t, S_{W_v}^t, E_v^t, C_v^t, L_v^t\}$, as described in Table 1.

Table 1: Description of variables for the representation of an object $v \in \boldsymbol{V}$ in the vicinity of the TeamMate vehicle considered for the first cycle.

Variable	Type	Unit	Description
X_v	Continuous	[m]	X-coordinate of the center of the object $v \in \boldsymbol{V}$ in a two-dimensional spatial coordinate system relative to the position of the TeamMate vehicle
Y_v	Continuous	[m]	Y-coordinate of the center of the object $v \in \boldsymbol{V}$ in a two-dimensional spatial coordinate system relative to the position of the TeamMate vehicle
Θ_v	Continuous	[rad]	Yaw-angle relative to a reference axis
V_v	Continuous	[m/s]	Longitudinal velocity along the objects heading
A_v	Continuous	[m/s²]	Longitudinal acceleration
W_v	Continuous	[rad/s]	Yaw-rate

S_{L_v}	Continuous	[m]	Length (along the x-axis)
S_{W_v}	Continuous	[m]	Width (along the y-axis)
E_v	Binary	$\{\text{true,false}\}$	Binary flag, whether the object $v \in \boldsymbol{V}$ exists in the current traffic scene.
C_v	Discrete	$\{0, \ldots, \lfloor C_v \rfloor\}$	Classification of the object $v \in \boldsymbol{V}$, e.g. PKW, LKW, VRU etc.
L_v	Discrete	$\{0, \ldots, \lfloor L_v \rfloor\}$	The lane, the object $v \in \boldsymbol{V}$ is currently located in, e.g. fast or slow lane on a two-lane road

For the actual prediction, let $\boldsymbol{S}_v = \{X_v, Y_v, \Theta_v, V_v, A_v, W_v\}$ denote a six-dimensional state for any $v \in \boldsymbol{V}$. At each point in time t, the VSSM will be used to infer a sequence of future states $p(\boldsymbol{S}_v^{t+i\Delta t} | E_v = true, \boldsymbol{o}^{1:t}), i = 1, \ldots, n$, with n and Δt being use-case dependent parameters. For online risk assessment, these predictions will be used to derive a region that encompasses the probable future location of the object v, in respect to its position, dimension, and orientation, with a probability of $(100 - \alpha)\%$, with α being an additional use-case dependent parameter.

2.3.2.1 Metric 1

Concerning the validation of the prediction of the evolution of the traffic situation, it is most important that the predicted regions actually encompass the true future location of the vehicle. As a metric to validate the performance, we therefore choose the "correct classification rate" as the ratio of correct predictions and the number of total predictions.

More specifically, validation will be performed on a set of independent test data D_{Test}, representing ground truth time-series of traffic situations. Let D_{Test} be composed by a number of m trials, where each trial is a time-series consisting of a number of $n_j, j = 1, \ldots, m$ data samples $d_j^k = \left(\boldsymbol{X}_{v_1}^k, \ldots, \boldsymbol{X}_{v_{n_V}}^k\right), k = 1, \ldots, n_j$. For each sample d_j^k, and each object $v \in \boldsymbol{V}$, we will predict a sequence of future states $p(\boldsymbol{S}_{j,v}^{k+i\Delta t} | E_{j,v}^k = true, \boldsymbol{o}^{1:k}), i = 1, \ldots, n$, and derive the region that includes the expected position of the vehicle with a probability of $(100 - \alpha)\%$. Based on test data, we will then check, whether the actual vehicle at time step $t + i\Delta t$ is within or outside of this region. We will perform this validation process for different levels of $\alpha = 50$, $\alpha = 75$, $\alpha = 90$, and $\alpha = 95$.

The metric can be used to assess the fulfillment of requirements R_EN3_model1.6 and R_EN3_model1.7, stating that the "integrated model must predict possible evolutions of the traffic situation in respect to potential interventions of the driver" (R_EN3_model1.6), resp. "[…] potential interventions of the automation" (R_EN3_model1.7) with a correct rate of the prediction of above 90% to be fulfilled. For the second cycle, we will abstract from the potential interventions of the driver and automation. For a 100% correct prediction and a region that encloses $(100-\alpha)\%$ of the probability mass, we would expect a failure-rate of $\alpha\%$. As such, we will treat the requirement as fulfilled, if the ratio of correct predictions is above $0.9(100-\alpha)\%$ for each prediction horizon $i\Delta t$ and level of α independently.

2.3.2.2 Upcoming Experiment/Study/Test (2nd Cycle)

Experimental data for the evaluation will be obtained from the experiments conducted in the first cycle for obtaining data for the driver models for intention recognition, experiments conducted in the second cycle for the semantic enrichment sub-module and where available, additional real data sets provided by the demonstrator owners.

The VED demo-car will be equipped with most of the components of the project. In the case of the WP2 the VED car will integrate the driver monitoring, the driver modeling systems and the will communicate with the OBU of BIT. The integration of the previous modules will be done in during the period month 18 – month 20.

First, the drowsiness and distraction module will be installed and tested on the VED simulator, in this case a software component will be interfaced with smart eye and an experiment will be realized on 20 persons. Once the VED baseline car will be ready (month 20) the hardware component can be integrated and tested, the tests will be done with 20 other participants.

In the second step we will make a feasibility of the V2X communication between the VED OBU and the RSU of BIT for the exchanging DENM messages.

3 Conclusion

This deliverable presented for each model that we are currently building to understand, assess and predict a driver, the vehicle and traffic situations a set of metrics. These metrics will be measured by experiments, studies, functional testing on an individual component level (Milestone 3). Later on in the upcoming year, these components will get integrated in driving simulators and real vehicles (WP5) for a first overall evaluation in WP6 (Milestone 4).