FOT-E

Project report

Rev: 20 November 2023

FFI/Vinnova Diarienr: 2019-03095

Contents

Acknowledgements 4
Executive summary 5
1. Introduction 6
2. FOT-e Dataset 7
3. FOT-e Algorithm 10
4. Methodology 12
4.1. Physical infrastructure 12
4.2. Algorithm development 12
4.2.1. General description of the iterative process 12
4.2.2. Manual labelling: How? 13
4.2.3. Manual labelling: What? 14
4.2.4. Manual labeling: How many FOT-e trips? 15
4.2.5. Manual labelling: How many FOT-e frames per feature? 15
4.2.6. Auto-labelling: development results 16
4.2.7. Testing 18
5. Results 19
5.1. Some statistics of the auto-labeling results 19
5.2. Analysis of the selected auto-labelled features 22
5.2.1. Stage 1: Real base rate 22
5.2.2. Stage 2: Total amount of labeled frames and Real performance metrics for the balanced set 23
5.2.3. Stage 3: Estimated performance metrics for the full auto-labelled dataset 26
6. An example of application: Head kinematics during braking in naturalistic driving 30
6.1. Introduction 30
6.2. Methods 30
6.3. Results 32
6.4. Discussion and conclusions 34
7. General project discussion and conclusions 35
References 36
Appendix A: Auto-labelled features description 37

Project coordinator
John-Fredrik Grönvall
Chalmers Tekniska Högskola (SAFER)
john-fredrik.gronvall@chalmers.se
Phone +46 768-504 107

Authors
Svitlana Finér *[Main editor]* (Smart Eye AB)
Ara Jafarzadeh (Smart Eye AB)
Henrik Lind (Smart Eye AB)
Giulio Bianchi Piccinini (Chalmers/Aalborg University)
Alberto Morando (Autoliv Development AB)
Erik Svanberg (SAFER/ Svanberg&Svanberg)
John-Fredrik Grönvall (SAFER / Chalmers Industriteknik)

Contributors
Johan Karlsson (Autoliv Development AB)
Thomas Streubel (Volvo Car Corporation)

Acknowledgements

This work was funded by the Sweden´s Innovation Agency VINNOVA under the program FFI, Strategic Vehicle Research and Innovation (Grant No 2019-03095, https://www.vinnova.se).

We thank Krystoffer Mroz, Ekant Mishra, and Nils Lubbe at Autoliv Development AB for their suggestions on Chapter 4 and to Carol Flannagan for useful suggestions on the evaluation of the algorithm in Chapter 5.

Executive summary

Driver impairment, including distraction and inattention, has a major impact on traffic safety and contributes to most crashes in Europe. Automation is expected to have a positive effect on safety as the driver is taken out of the driving task. However, for first implementations in SAE level 3, the driver is still expected to serve as fall back which requires a fast switch of attention towards the driving task. So, detecting the driver state and developing a better understanding of driver impairment is a key enabler to enhance existing advanced driver assistance systems (ADAS) as well as to identify new safety relevant factors to be considered in the development of new systems up to automated driving functions.

The recent development in machine learning (ML) shows promising potential in recognizing different driver states (such as distraction), various secondary task engagements (such as talking on the phone), and driver posture (such as out of position). These algorithms are data-driven and require large amounts of labelled images for training and testing.

The SAFER Naturalistic Driving Data (NDD) platform has been developed over ten years, at the vehicle and traffic safety centre hosted by Chalmers University of Technology. The datasets cover 7.5 million km of real-world driving in different contexts, countries, and vehicle types. In particular, the dataset includes big datasets of videos collected from cameras pointed at the driver.

The project goal was to enrich the existing datasets at SAFER by extracting features from video data. The vision was to create a world class vehicle and traffic safety dataset for research and development of active and passive vehicle safety systems. Further, the project validated several new functions for attention warning on the enriched dataset.

The project built on competence and advancements in image processing by applying and optimizing state-of-the-art ML algorithms on the existing large-scale naturalistic driving dataset. The industrial purpose was to validate and enhance existing algorithms in smart camera technology and in-vehicle ADAS systems.

The main research goal was to enable new ground-breaking traffic-safety analyses on driver state, secondary task engagement, and posture, which have never been possible before because existing datasets are too small and pure manual labelling is costly and time consuming.

SAFER/Chalmers was the main applicant leading the project and in addition contributing to data management preparation and performing analysis, both on active and passive safety systems, based on the output of the enriched data. Smart Eye, experts in smart camera technology, did the main work on feature extraction and used the enriched dataset to validate new functions. Autoliv focused on passive safety features and provided access to data from the Eyes-On-Road project. Volvo Cars and Volvo Group supported the project by providing access to data.

The project was originally three years starting on November 2019, but extended to four years due to COVID-pandemic issues. The total project budget of 8,5 MSEK where the consortia received VINNOVA/FFI funding of 4,2 MSEK.

1. Introduction

This project started with two important inputs: the SAFER NDD database, and an improved version of the Smart Eye driver and cabin monitoring algorithms. The improved algorithm though needed to be adapted to the specifics of the available data. This adaptation, i.e., training of the algorithm, needed manually labelled data (from manual video labeling done by students at SAFER) for generating a model. This model was tested on selected parts of the SAFER NDD database, and the result was validated. The time-consuming labelling and validation effort had to be repeated multiple times, and the size of the tested data increased for each step. This iterative approach generated the result of an improved dataset of two groups of features: labelled data of the Driving Monitoring System (DMS) including gaze, head and eye tracking, and features generated from Cabin Monitoring System (CMS), covering driver posture from detection of limps.

In previous projects, data was only partially annotated and analyzed, as it was done in relation to certain events of interest of those projects. Since the data was collected and certain features were auto labeled by Smart Eye the first time, the Smart Eye software has been further developed, and now of this project had not only more possibilities to analyze head and eyes of the driver, but also had new possibilities to analyze full upper body. As a result, this project created an opportunity to annotate more trips and more features.

In general, the numbers presented in this project for both CM and DM are impressive.
In this project **over two billion images were auto labelled** where the frame quality is ok.

2. FOT-e Dataset

The SAFER naturalistic driving data (NDD) platform contains data collected mainly in three large European projects: EuroFOT, Drive C2X and UDRIVE. The data includes various videos (see Fig. 2.1 for examples of videos) and logs collected on public roads in different contexts, countries, and vehicle types covering 7.5 million km for 130 000 hours. During the data collection projects, all drivers have signed consent forms, that they agree that the images and videos are ok for research and publication. Nevertheless, all videos are only available at SAFER in special locked FOT rooms, by authorized personnel.

The data from EuroFOT and DriveC2X was using similar logging equipment, camera positions and head and eye tracking system, which motivated the choice of these two datasets for the project, collectively called further on as the "FOT-e dataset". In total, the FOT-e dataset represents 2 440 million kilometers during 52 400 hours of real-world driving on public roads.

The data in EuroFOT was collected in 2010-2011 and DriveC2X in 2013. The data recorders were installed in two different vehicle types: Volvo V70 and Volvo XC70. The interior in these vehicles is similar but it is important to understand that the conditions for recording data vary. Since the vehicles are used in normal day-to-day conditions, the position of the steering wheel, as well as the position of the driver seat, varies in height (x), depth (z), and for the seat, also the angle. The drivers were short and tall, men and women, with or without glasses. All these aspects add each a level of complexity for the algorithms to detect different features. Also, data was collected in different seasons which means different clothing, and use of sunglasses etc. In addition, data recorded nighttime in dark conditions is quite different from daytime, however some images were too bright due to sunlight. This is expected beforehand, nevertheless a true challenge.

a) b)

Figure 2.1 Example of videos from the cameras. a) driver's head view example; b) right driver's upper body view example.

The files are organized based on the concept of a "trip". A trip is defined from when the engine ignition is on until the engine ignition is off, and all data recorded during this time is attached to it. Any trip shorter than 90 seconds has already been discarded, and the average trip times are about 20 minutes long. However, a trip can be more than 6 hours, in its extreme.

Each trip has following datafiles associated with it (see Figure 2.2):

- five .avi video files of different views including:
  - front view outside of the car,
  - rear view outside of the car,
  - driver's upper body view (see Figure 2.1),
  - driver's head view (see Figure 2.1),
  - driver's feet view,
- oDBdata.mat (Matlab file that include in-vehicle signals, GPS position, kinematics),
- oDBdataET.mat (Matlab files that include head- and eye-tracking data produced by the earlier versions of the Smart Eye software processing driver's head view videos).
-

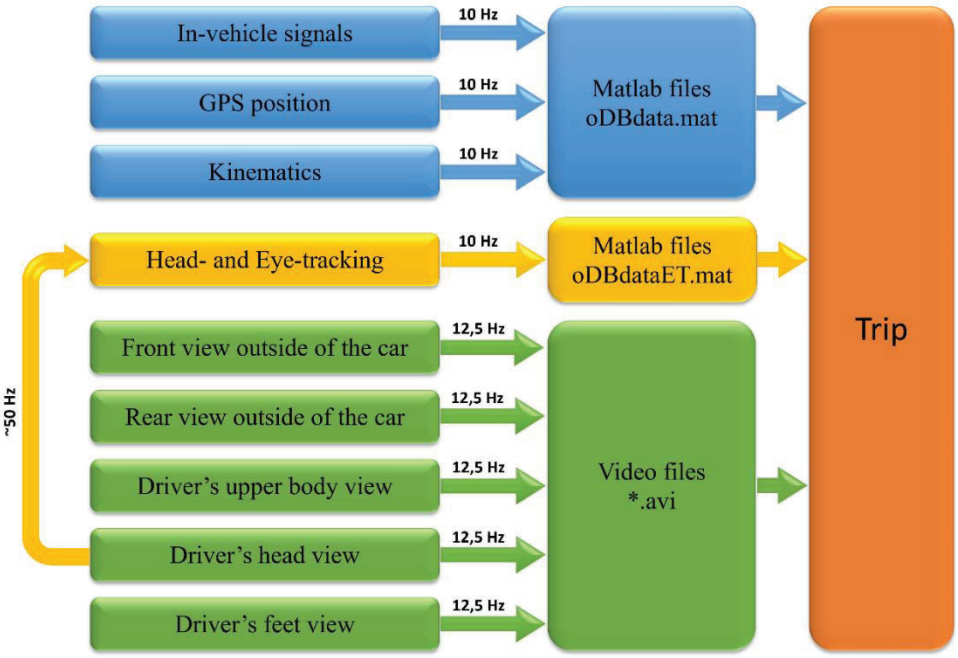

Figure 2.2 Schematic representation of the data contained in a "trip", in the FOT-e dataset.

In-vehicle signals, GPS position, kinematics, head- and eye-tracking data were recorded at 10 Hz, while the videos were recorded at 12.5 Hz. The videos were recorded in black and white. When driving in darker conditions, IR was used. There was no control over the illumination conditions, so there were multiple cases of over and underexposure (see Figure 2.3). To reduce consumed space, the video files were processed through a compression algorithm (h264). To conclude, the video data is of low quality, but this limitation is compensated for the unique data on driver behavior included in the dataset.

Figure 2.3 Examples of over- and under-exposed video images.

It is important to note that video with the view of the driver's head was recorded with the Smart Eye System, while the other 4 videos were recorded using another equipment. Because of this, videos with the view of the driver's head have different total amount of frames then the rest of the videos.

3. FOT-e Algorithm

The main idea behind this part of the project is the development of an algorithm which can auto-label certain features by applying Smart Eye software to the videos of the driver.

Since the videos are stored in 12,5 Hz frequency and the driver monitoring data are stored in 10 Hz frequency, the algorithm is split into two stages. During stage 1, features are auto labeled for each frame of the video, and, during the stage 2, the label frequency is adjusted from 12,5 Hz to 10 Hz (see Figure 3.1). The two stages are described below in more detail.

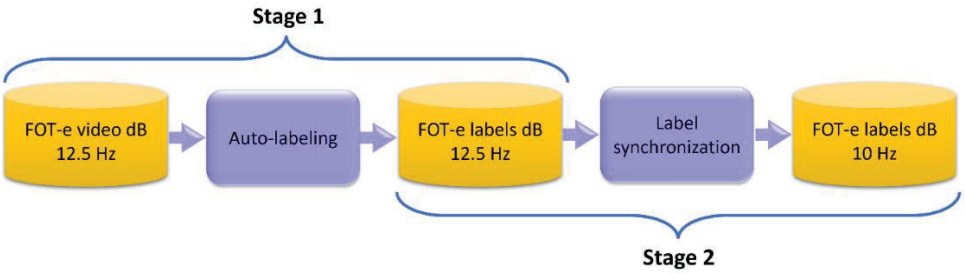

Figure 3.1 Schematic presentation of the two stages of the algorithm.

Stage 1. The Smart Eye software is divided into two modules: Driver monitoring (DM) module and Cabin monitoring (CM) module. DM module used as input videos with the view of the driver's head (see Figure 2.1) and CM module used as input videos with the view of the driver's upper body (see Figure 2.1). At first, videos are processed extracting all available Smart Eye features and then those features are post-processed to extract labels defined in this project (see Figure 3.2).

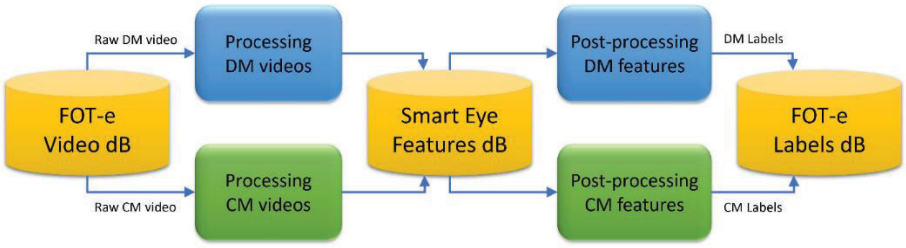

Figure 3.2 Schematic presentation of the stage 1 of the algorithm.

In the algorithm two quality levels were present, frame quality and prediction quality (see Figure 3.3 and Figure 3.4). Frame quality was introduced due to the poor quality of the videos and prediction quality was present in the original Smart Eye software.

"Frame quality" output is a Boolean; however, it is defined differently for DM and CM videos. For DM the "Frame quality" had a value of 1 if the algorithm has been able to predict the head bounding box. For CM

the “Frame quality” had a value of 1 if the algorithm has been able to identify the object on the driver seat as a person.

“Prediction quality” output is a value in the range between 0 and 1, indicating how confident the algorithm is in that prediction. Filtering relative to the “Prediction quality” may be done using different threshold values, depending on the user needs. In some cases that filtering is done inside of the algorithm and in other cases it is output together with the feature label.

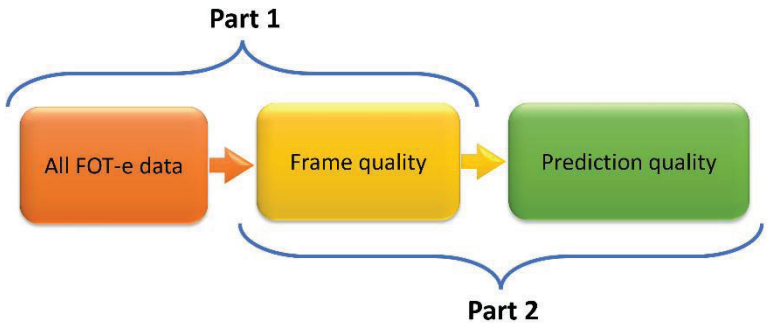

Figure 3.3 Schematic representation of the quality filtering steps.

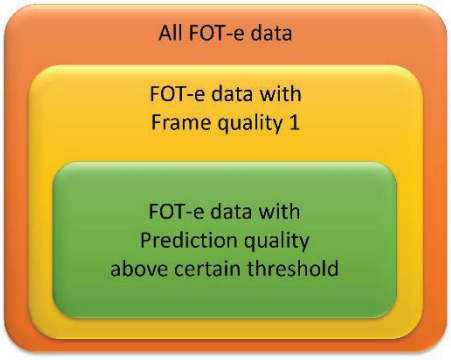

Figure 3.4 Schematic representation of the data relative to the quality levels.

Stage 2. In the label synchronization stage, each video frame in the oDBdata file is associated with the data from the auto-labeled data. In all, this means that five frames of the auto-labeled data are skipped every two seconds. Since the data was quite noisy and still with high details, no sampling techniques are used to compensate for the excluded frames. The principle is described in the example of Figure 3.5 where the frames 4, 9, 14, 19 and 24 are skipped.

Time	1 sec												2 sec											
	0	0.1	0.2		0.3	0.4	0.5	0.6		0.7	0.8	0.9	0		0.1	0.2	0.3	0.4		0.5	0.6	0.7	0.8		0.9
Frame ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
oDBdata																									
videoFrames																									

Figure 3.5 Schematic presentation of the stage 2 of the algorithm.

4. Methodology

4.1. Physical infrastructure

The videos included personal data, and their use was restricted by the study participants' consent. This stipulated that the data had to be accessed at the SAFER premises. The SAFER NDD infrastructure included four dedicated FOT rooms. One of those rooms was used by the Smart Eye team for the algorithm development and processing of the data. For that, Smart Eye AB provided computers that were placed in that room. Computers got password protection to restrict access to only the members of the project. In addition to that, computers were not allowed to connect to the Internet. Once these security measures were implemented, the FOT-e dataset was copied to those computers occupying approximately 4 TB of the disc space.

4.2. Algorithm development

4.2.1. General description of the iterative process

At first, existing Smart Eye software was used to auto-label the data and was applied to few selected videos. It was found that the video files from the FOT-e dataset had a very low resolution and were partly corrupted by overexposure. In addition to that, videos had different fields of view between different vehicles, due to slight differences in the cameras' position resulting from the use of the vehicle. Therefore, the existing Smart Eye software needed to be adapted to be used on the FOT-e data. This adaptation of the Smart Eye software took place through an iterative process and at every iteration went through roughly the same steps (see Figure 4.1).

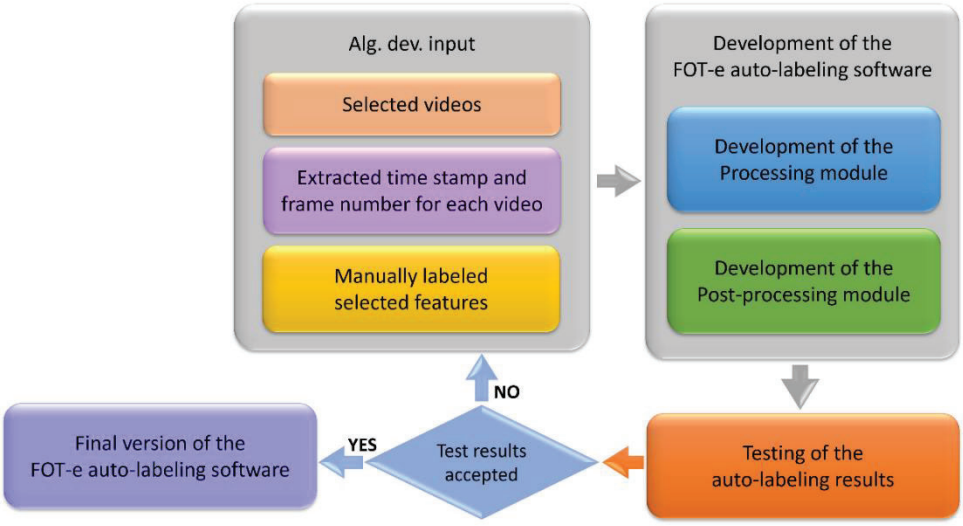

Figure 4.1 Algorithm development iterative process overview.

At first, a set of videos was selected as well as the features for auto-labelling. The choice of the features was guided by the condition that these features could also be manually labeled by annotators Manually labeled features are stored at 10 Hz, similarly to the oDBdata.mat, while the auto-labeled features are stored at 12.5 Hz, similarly to the videos. To create a match between auto and manually annotated labels, the internal timestamp (time index) of the manually labeled features and the corresponding frame number of the auto-labeled features were extracted and stored in a separate file (see Figure 4.2). The manually labeled data with

corrected frequency was used as an input for the development and adaptation of the Smart Eye software to the FOT-e data (further called “FOT-e auto-labeling software”). Once the development and adaptation were finished, the resulting software was applied on the test video set and the results were compared to the manual labeling where possible. If the results were not acceptable, new iteration was initiated. If the results were acceptable the resulting software was finalized and applied to the whole FOT-e dataset.

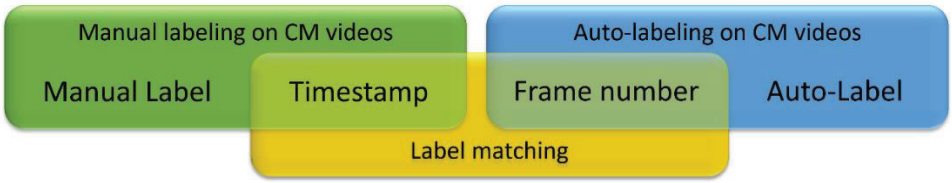

Figure 4.2 Schematic presentation of the manual and auto-label matching for the CM videos.

4.2.2. Manual labelling: How?

The manual labelling was done using the software FOTware (Dozza et al., 2010), developed in the previous Field Operational Tests (FOT) SemiFOT and EuroFOT, and continuously updated in the follow-up projects. FOTware allows to play multiple videos collected from the different cameras described in Figure 2.2, and to show signals collected from different sensors (see Figure 2.2). The software also provides a graphical user interface for manual labelling of the data (see Figure 4.3).

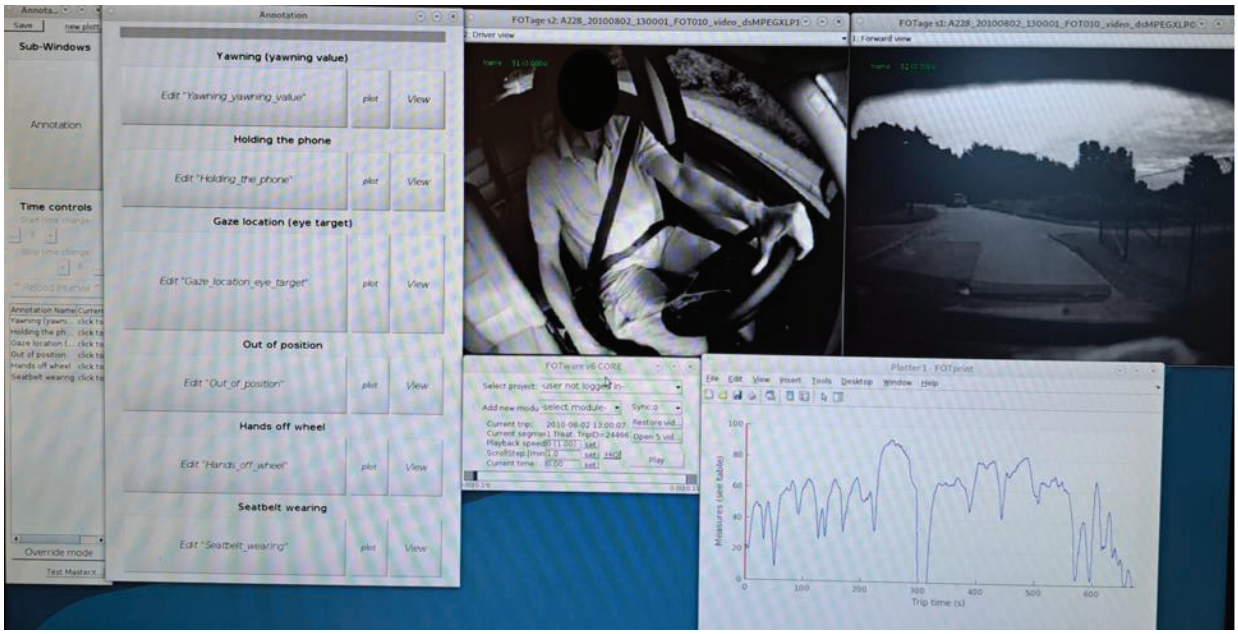

Figure 4.3 Screenshot of the FOTware software used for manual labeling.

The data was manually labeled by student assistants (hereafter called ‘annotators’) who joined the project for different periods of time. The annotators were supervised by the Chalmers’ researchers, who provided training on how to perform manual labeling, supported the students when needed and checked the quality of the manual labelling process by randomly sampling the labelled data.

4.2.3. Manual labelling: What?

The choice of the variables to manually label was jointly taken in the project, considering temporal constraints due to the time-consuming labeling process, and feasibility related to the likelihood of being successful in auto-labelling the specific features. The definition and the categories for each feature were proposed based on the researcher dictionaries c developed in the SHRP2 (SHRP2, 2016) and UDRIVE projects (Bargman et al., 2017). Overall, the following features were selected for the initial manual labeling:

- "Hands on wheel"
- "Secondary task"
- "Phone usage"
- "Drowsiness"
- "Out of position"

A short description of each feature is given in Table 4.1.

Table 4.1 Short description of the features chosen for the initial manual labeling. Features that were not auto-labeled by the final FOT-e algorithm are highlighted in grey.

#	Feature name	Short feature description	Possible categories
1	Drowsiness	Visual indications for drowsiness are long blinks, yawning and posture changes. Long periods of closed eyes (micro sleep), head nods and startles are strong indications for a high level of drowsiness.	1. No drowsiness visible (baseline) 2. Both eyes closed 3. One eye closed / another eye open 4. Yawning. 5. Startle
2	Phone usage	Handheld use of phone for talking and/or texting.	1. No phone usage 2. Talking/listening with left hand 3. Talking/listening with right hand 4. Texting with left hand 5. Texting with right hand
3	Out of position	Estimated offset laterally and longitudinally from the head rest.	1. Central position (baseline) 2. Lateral out-of-position 3. Longitudinal out-of-position 4. Lateral and longitudinal out-of-position
4	Hands on wheel	Which hands of the driver are touching the steering wheel.	1. No hands-on wheel 2. Left hand on / right hand off. 3. Left hand off / right hand on 4. Both hands on wheel 5. Unknown
5	Secondary task	Attempts to capture distraction activities (other than phone usage) which may influence driver performance. Focus on distraction from inside the vehicle that are visible in the actions of the driver (video detection) and exclude phone use.	1. No activities 2. Interaction with passenger in adjacent seat 3. Talking/singing/whistling 4. Reaching for an object 5. Interaction with center stack 6. Eating/drinking 7. Smoking 8. Hands-face-interaction 9. Reading

4.2.4. Manual labeling: How many FOT-e trips?

The first manually annotated dataset contained features that were selected based on the convenience sampling, i.e., sampling, which relied on the previously labelled data within the original projects DriveC2X and EuroFOT. The convenience sampling allowed identifying parts of the data where some of the features (e.g., phone usage) were already existing, without randomly scanning the whole dataset for finding occurrences of the feature. The first manual annotations based on convenience sampling resulted in the dataset A1, including 1032 segments (i.e., period for which the data was annotated) whose duration was approximately 28 seconds (see Table 4.2). This first dataset did not include enough information for the development of the algorithm for one of the features (phone usage), so additional manual labelling was performed. An additional dataset—named dataset A2—was therefore manually labeled. The dataset aimed to mitigate two concerns associated to dataset A1: a) the lack of sufficient labeled data about participants' phone use; b) the bias associated with the small number of participants whose videos were manually labeled. The dataset A2 included data from two participants known for their extensive use of the phone, and from participants whose videos were not previously manually labelled. The resulting dataset A2 included a smaller number of segments (i.e., 37) but with the duration extended to 140 seconds (see Table 4.2). Then, the total dataset encompassed 1069 segments, each one having an average duration of 32 seconds as shown in Table 4.2. The corresponding number of trips manually labelled trips became 900 out of the total number of approximately 135 000 trips. Some statistics of the manually labelled data are presented in Table 4.2.

Table 4.2 Selected statistics for the results for manual labelling.

Category	Dataset A1	Dataset A2	Dataset A=A1+A2
Number of trips	880	20	900
Number of drivers	127	4	131
Number of segments	1032	37	1069
Number of segments by project: DriveC2X EuroFOT	 396 636	 0 37	 396
Average segment length	28 seconds	141 seconds	32 seconds
Overall duration	Approximately 8 hours	Approximately 1 hour and 30 minutes	Approximately 9 hours and 30 min

4.2.5. Manual labelling: How many FOT-e frames per feature?

In this section, we present the number of frames manually labelled for the features phone usage, hands on wheel and secondary tasks for the 900 trips included in datasets A1 and A2 (see Table 4.3). Table 4.3 also reports the number of frames annotated for the categories belonging to each feature.

Table 4.3 Selected statistics for manually annotated features.

#	Feature	Category	Frames	%
2	Phone usage	1. No phone usage 2. Talking/listening with left hand 3. Talking/listening with right hand 4. Texting with left hand 5. Texting with right hand **Total**	681 262 58550 86203 11198 25807 **863020**	79 % 7 % 10 % 1 % 3 % 100 %
4	Hands on wheel	1. No hands on wheel 2. Left hand on / right hand off 3. Left hand off / right hand on 4. Both hands on wheel 5. Unknown **Total**	28443 514544 177263 146866 8449 **875565**	3 % 59 % 20 % 17 % 1 % 100 %
5	Secondary task	1. No activities 2. Interaction with passenger in adjacent seat 3. Talking/singing/whistling 4. Reaching for an object 5. Interaction with center stack 6. Eating/drinking 7. Smoking 8. Hands-face-interaction 9. Reading **Total**	381973 3635 36599 30781 12943 8426 0 63401 2978 **540736**	71 % 1 % 7 % 6 % 2 % 2 % 0 % 12 % 1 % 100 %

4.2.6. Auto-labelling: development results

During the algorithm development, some features were added (due to the Smart Eye software advancement), and some features were removed or re-defined (due to poor quality of the videos or lack of time). The final list of auto-labeled features became:

- “Eye openness”
- “Yawning”
- “Viewing targets”
- “Head pose”
- “Phone usage”
- “Body key-points”
- “Out of position”
- “Hands off wheel”
- “Frame quality”

Throughout the algorithm development process, reliability was prioritized over availability. Due to the abundance and low quality of FOT-e videos, uncertain predictions were discarded, and only confident outputs were reported to provide high quality data for future research. It is important to note that all FOT-e features required some finetuning and adjustments of the latest Smart Eye software. In addition to that, neural networks used for “Phone usage” were re-trained to fit FOT-e data, as well as classical statistical

models used for the “Yawning” and “Out of position”. For that, re-training manual labelling presented in the manual labelling chapter was used. In addition to manual labelling of “Yawning”, “Phone usage” and “Out of position”, manual labelling for “Hands off wheel” was also used for adjusting the algorithm. A short description of each label is given in Table 4.4. A more detailed description can be found in Appendix A.

Table 4.4 Short description of the auto-labeled FOT-e features. Features finetuned in this project using manual labeling are highlighted in green.

#	Feature name	Short feature description	Output description	Alg. based on neural network or statistical model	Testing based on visual inspection and ground truth comparison
1	Eye openness	Eye openness or Eyelid opening, describes how open the driver eyes are in a frame.	It is a floating-point value in the range [0.0, 1.0].	NN	VI
2	Yawning	Yawning output stating whether the driver is yawning.	Boolean	SM	GTC
3	Viewing targets	The viewing targets module provides information on what the driver is currently looking at. Viewing targets are geometrical shapes that represent the environment around the driver. In this case 4 planes were used: left, right, forward, and down.	For every target plane the output is a Boolean, where a value equal to one represents an intersection point between the viewing direction and a viewing target. Only intersections that have a higher probability than 0.5 are counted. In absence of intersection with the defined planes or high-quality intersection *eye_target_unknown* is reported.	NN	VI
4	Head pose	Head pose output includes head position and orientation.	The head position (x,y,z) is expressed in reference coordinate system and the head orientation is expressed in heading, pitch, and roll.	NN	VI
5	Phone usage	Phone usage attribute states whether driver is holding phone to the ear with the left or right hand.	Boolean	NN	GTC
6	Body key-points	Driver posture is described using 15 core body key-points: Nose, Two Eyes, Two Ears, Two Shoulders, Two Elbows, Two Wrists, Two Hips, Two Knees.	Each body key-point is represented in pixel coordinates with the origin in the top-left corner.	NN	VI
7	Out of position	Out of position indicates whether a person is out of position compared to the neutral seating position relative to the head rest.	Boolean	SM	GTC
8	Hands off wheel	All cars used in the dataset have the steering wheel on the left side. Steering wheel polygon region was drawn, and hands-off-wheel were predicted using that region. If the wrist point’s distance to the steering wheel polygon was more than the size of a hand, the hand was set as off the wheel. If it was inside the region or closer than a hand-size to the region, the hands-off-wheel prediction was set to unknown for both hands.	1. Both hands off wheel 2. Left hand off wheel. 3. Right hand off wheel 4. Unknown	NN	GTC

4.2.7. Testing

The output labels are categorized into two types, regression-type, and classification-type. The regression-type labels, as well as some of the classification-type labels, have been evaluated by visual inspection as no manual labels have been available for those auto-labeled features. Such labels include “Eye target”, “Eye openness”, “Head pose”, and “Body key-points”. For some of those labels, some of the algorithm output results were overlayed with the corresponding videos, which were further inspected visually (see Figure 4.4).

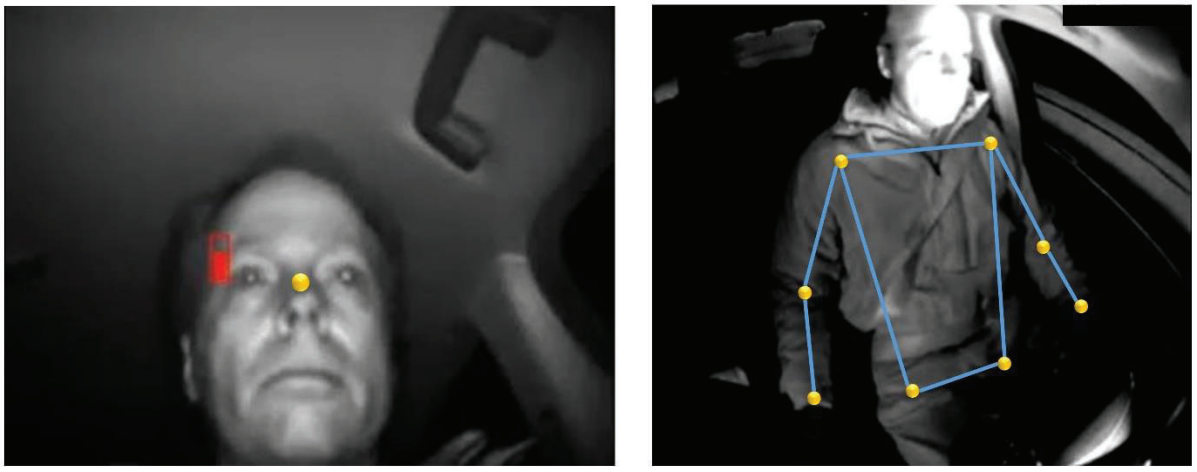

Figure 4.4 Test results for Head pose, Eye openness (to the left) and Body key-points (to the right).

Auto-labeled features of classification-type for which manual labels have been available, were compared to the manual labels available from the project (see sections 4.2.4 and 4.2.5). In this section, testing results are presented for “Yawning”, “Phone usage” and “Out of position”. The normalized confusion matrices are presented in Table 4.5

Table 4.5 Confusion matrix formula.

Confusion matrix formula		Auto-labelled	
		0	1
Manually labelled	0	True Negative	False Positive
	1	False Positive	True Positive

Table 4.6 Normalized confusion matrix for “Out of position” label

Confusion matrix for “Out of position”		Auto-labelled	
		0	1
Manually labelled	0	0,89	0,11
	1	0,07	0,93

Table 4.7 Normalized confusion matrix for "Yawning" label.

Confusion matrix for “Yawning”		Auto-labelled	
		0	1
Manually labelled	0	0,99	0,01
	1	0	1

Table 4.8 Normalized confusion matrix for "Phone usage" label.

Confusion matrix for “Phone usage”		Auto-labelled	
		0	1
Manually labelled	0	1	0
	1	0,12	0,88

5. Results

5.1. Some statistics of the auto-labeling results

This section presents some statistics of the auto-labeling results. It is organized by video type, DM vs. CM (see Table 5.1, as well as Part 1 in Figure 3.3) and by label name (see Table 5.2 , as well as Part 2 in Figure 3.3) as well as by quality level, “Frame quality” and “Prediction quality”.

Table 5.1 Selected statistics of the results relative to the video type and “Frame quality”.

#	Data	Trips	Frames	Time (hours)	Percentage relative to the Total data	Percentage relative to the DM/CM data
1	Total data	149 556	1 530 695 794	42 519	100 % (1)	-
2	Total DM data	130 582	1 238 707 530	34 408	81 % (1)	100 % (2)
	Total DM data with “Frame quality”=1	129 293	979 931 860	27 220	-	79% (2)
3	Total CM data	135 361	1 379 085 315	38 300	90 % (1)	100 % (3)
	Total CM data with “Frame quality”=1	135 340	1 162 833 816	32 301	-	84% (3)

Table 5.2 Selected statistics of the results relative to the feature label and quality.

#	Label	Frames with "Frame quality"=1 and "Prediction quality" above 0 or included in the algorithm	Frames with "Frame quality"=1 and "Prediction quality" above 0.25	% of total DM data with "Frame quality"=1	% of total CM data with "Frame quality"=1
1	Eye openness	838 267 462	688 294 828	86 %	
2	Yawning	685 125	n/a	0.07 %	
3	Viewing targets *				
	Forward	794 146 852	n/a	81 %	
	Right	100 573 849	n/a	10 %	
	Left	67 359 520	n/a	7 %	
	Down	19 563 235	n/a	2 %	
	Unknown	53 382 023	n/a	5 %	
4	Head pose				
	Head position X Head position Y Head position Z	979 931 861	968 158 059	99 %	
5	Phone usage (total labelled either 1 or 0)	979 931 860	962 983 729	98 %	
	Phone Usage equal to 1	11 659 857		1.19 %	
6	Body key-points				
	Eye Left		1 127 315 238		97 %
	Eye Right		1 123 903 539		97 %
	Nose		1 112 966 034		96 %
	Ear Left		1 111 155 067		96 %
	Ear Right		1 118 939 324		96 %
	Shoulder Left		1 057 058 634		91 %
	Shoulder Right		1 044 355 143		90 %
	Elbow Left		924 919 523		80 %
	Elbow Right		967 384 851		83 %
	Wrist Left		953 587 082		82 %
	Wrist Right		988 535 501		85 %
	Hip Left		660 067 984		57 %
	Hip Right		667 846 748		57 %
	Knee Left		300 352 085		26 %
	Knee Right		391 164 901		34 %
7	Out of position		259 089 873		22 %
8	Hands off wheel		277 322 925		24 %

* Note that these features are set individually why multiple areas could be detected in the same frame (thus leading to a total of 105 %)

A project collecting data with a naturalistic approach imply that some data cannot be used for various reasons. The total amount of data (1) is thus just an indication of the overall availability of any data from the data logger standpoint. Collecting data for a period of up to three months, means that cameras can stop working, or other issues with the logger arise. To prevent major dropouts, the projects developed an online monitoring tool used to visually inspect the data collection phase. For each trip, one image for each camera view, minor subset of in-vehicle data, sample from the accelerometer and indication from the GPS, was uploaded to this online tool. Since only generating a snapshot, the complete trip was not checked, and for time-to-time cameras stopped working. If issues were detected, it was not all easy to get the problems fixed at a workshop. Also, there might be video files that cannot be read due to inconsistencies in the data format, or having issues related to the data compression algorithm.

The more interesting figure to look at is the relation between the total labelled frames for either DM(2) or CM(3) and so called good quality frames, frames with “Frame quality”=1. For DM videos, this discards any frames where a head bounding box is not found, which often could be due to the position of the camera in relation to the steering wheel or the position of the hand of the driver. It is seen in the data, that some vehicles have an overall higher quality than others. For the DM videos there are 38 vehicles that have “Frame quality” =1 in more than 80% of the frames, and in more than 50% of their trips. This is a very strict threshold and will help define new trip-based quality metrics to decide which data to use in future research projects. It is notable that 105 066 trips have more than 100 seconds of good frame quality.

In general, the numbers presented in this project for both CM and DM are quite impressive. In this project **over two billion images were auto labelled** where the frame quality is ok. As an additional quality metric, many features have their own quality indicator, “prediction quality”, that is available separately or was considered in the feature prediction value.

Looking at the numbers grouped per vehicle it is easy to see over-representation of “Phone usage”. This is expected due to behavioural patterns of the drivers. It should be worth stating that using phones was not illegal for the course of the EuroFOT project (2010-2011), and only in part of the Drive C2X project (2013). However, the public opinion on using phones when driving could have had an effect. At a glance, it was seen a decrease in phone usage by -20% in the data collected in 2013 compared to 2010-2011. When excluding outliers (6 vehicles in EuroFOT and 1 vehicle in Drive C2X), this effect is even more clear (-70%). “Phone usage” is labelled in a total of 9 508 trips.

Yawning is labelled in 19 530 trips and are usually happen more than one time in the same trip. This could support future research on drowsiness.

For the CM videos, the detected values of body key points are logical; it is easier to detect the face and upper body, than the knee or hip. The right side often has a higher detection rate. This should be explained by the values are seen from the driver perspective looking forward, and thus the right side is closest to the camera. Why the shoulder has a slightly lower value on the right side is difficult to explain, however the rate is fairly high to start with (right side 90 % / left side 91 %).

The indication of “Out of position” need further analysis. A first look indicate that this value is set to true where different trips stand out. This feature is over-represented in longer journeys with more than one person in the vehicle. It is yet unclear why these trips are over-represented and further analysis is needed.

It was difficult to detect hands on steering wheel, since sometimes the complete steering wheel was out of sight in the video (again depending on the specific position of the steering wheel). Having two hands on the steering wheel could be of interest to detect in the future. By manual inspection most of the time the driver

has only one hand on the steering wheel and having two hands could potentially indicate a more challenging situation (e.g., driving in winter conditions, in city traffic or when performing an evasive manoeuvre). This is an area where more analysis is needed.

Further work on some features could include interpolation. Depending on the feature characteristics, 0.1 to 0.3 second dropouts could be interpolated to give more stable signals.

Further steps of analysing this data could be to check if different behaviour (phone usage, not looking forward, yawning, as a sign of drowsiness) can be linked to incidents (safety critical events) or increased risk in traffic. Also, research on the driver position in relation to safety critical events, could be of interest. Given the actual posture of the driver in these situations, it could be of interest to analyse the posture in relation to passive safety systems.

5.2. Analysis of the selected auto-labelled features

Due to the extensive time required to manually label and analyze the data, the evaluation of the auto-labeled features was exclusively done for the two features: “Yawning” and “Phone usage”. Both features belong to the classification type, so it was possible to compare auto-labelling results with manual labelling. Within this project, it was impossible to manually label all occurrences for those features in the FOT-e dataset. So, a statistical method was applied to perform the evaluation of the auto-labeled features and to scale up the results to the whole FOT-e dataset. The overall statistical method could be separated into three stages. During the first stage the real base rate, real percentage of positive and percentage of negative labels, is defined. During the second stage, the real performance metrics are calculated for the balanced dataset (Dataset B), as well as the total amount of auto-labelled frames for the full FOT-e dataset. During the third stage, analysis results are upscaled for the full FOT-e dataset to obtain estimated performance metrics.

5.2.1. Stage 1: Real base rate

The base rate for the feature “Yawning” was calculated using the manually labeled frames of the dataset A (900 trips used for the algorithm development) (see Figure 5.1). Out of the total amount of frames, 0.05% were positive labels (i.e., yawning) and 99.95% were negative labels (i.e., no yawning).

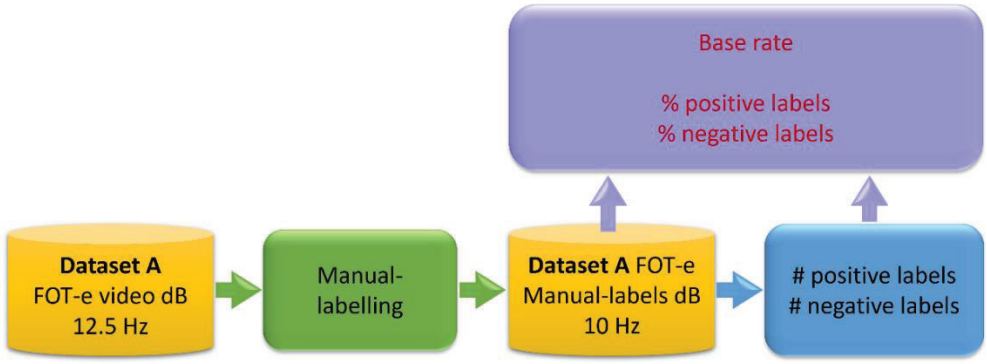

Figure 5.1 Schematic presentation of the Stage 1 of the analysis. Defining real base rate.

For “Phone usage”, the split between positive and negative labels in the dataset A was respectively 6.09% and 93.91%. The selection of the sample for “Phone usage” was biased by the fact that participants with intensive phone activity were selected on purpose, to increase the frame number needed for the algorithm development. Therefore, the base rate for the “Phone usage” obtained from manual labelling of the dataset A could not be used for the upscaling of the results. Instead, it was decided to look in the literature for a surrogate base rate which could be used for the upscaling of the results. Dingus (2014) reported that talking and listening on hand-held phone device ranges from 2.25% in 2013 Cell Phones study with 204 participants to 2.56% in 100-Car with 109 participants. These studies did not separate between hand-held phone use by the ear versus other types of hand-held phone use. Therefore, these base rates are probably an overestimation of the base rate that should be used for the upscaling of this project results. To be conservative, the smaller value (2.25%) among the two studies was taken as a base rate for the upscaling of the “Phone usage” in this project.

5.2.2. Stage 2: Total amount of labeled frames and Real performance metrics for the balanced set

The schematic representation of the stage 2 is shown in Figure 5.2.

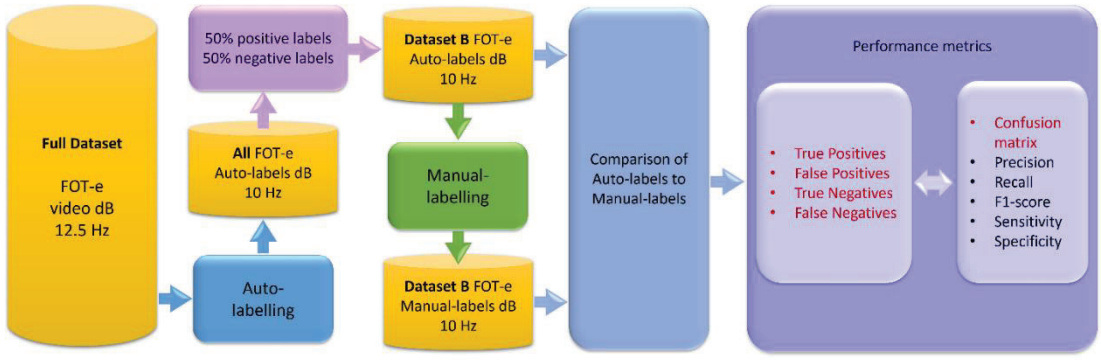

Figure 5.2 Schematic presentation of the Stage 2 of the analysis. Defining total amount of the auto-labeled frames, as well as the real number of TPs, FPs, TNs, FNs for balanced positive and negative labels.

Both “Yawning” and “Phone usage” features were auto labeled using DM videos. The total amount of auto-labeled trips on the full FOT-e dataset became 130 582, which corresponded to 1 238 707 530 frames. Among those frames, a balanced set (Dataset B) with an equal amount of positive and negative labels was selected for each feature, Dataset B1 for “Yawning” and Dataset B2 for “Phone usage”. Balanced sets contained drivers and trips that were not used in the algorithm development, to evaluate the algorithm in the most “demanding” situations (i.e., unbiased data). The size of the balanced set was chosen to be made of 160 frames: the number of frames was selected considering the time-constraints for the manual labelling process and the analysis of the data. This biased selection of an equal number of positive and negative labels was introduced because the rates of positive labels for both “Yawning” and “Phone usage” were low and it was required to obtain a minimum number of entries for each cell of the confusion matrices for “Yawning” and “Phone usage”. Since part of the time for the manual labeling process requires finding and opening files, we tried to take larger advantage of the manual labeling process by manual labeling 10 seconds before and 10 seconds after the extracted frames. So, the aim was to obtain 160 segments of 20 seconds manually labeled for both yawning and phone use.

Out of the 160 segments originally selected for “Yawning”, 11 were excluded due to data quality issues (e.g., it was not possible to play the video). So, the Dataset B1 for the analysis of “Yawning” included 149

segments, 76 segments with central frame auto-labelled as 0 and 73 segments with central frame auto-labelled as 1. The confusion matrix resulting from the comparison of the manual labels performed by two students and the auto-labels made by the algorithm is reported in Table 5.3

Table 5.3 Confusion matrix for "Yawning".

Confusion matrix for "Yawning"		Auto-labelled	
		0	1
Manually labelled	0	74	8
	1	2	65

Table 5.4 Normalized confusion matrix for "Yawning".

Confusion matrix for "Yawning"		Auto-labelled	
		0	1
Manually labelled	0	0.90	0.10
	1	0.03	0.97

Normalized confusion matrix ca be compared to the confusion matrix from development and the results look quite similar, which mean s that the algorithm has stable performance.

The performance metrics for the analysis of "Yawning" are reported below.

$$Precision = \frac{TP^{1}}{TP+FP^{2}} = \frac{65}{65+8} = 89.04\% \quad (1)$$

$$Recall = \frac{TP}{TP+FN^{3}} = \frac{65}{65+2} = 97.01\% \quad (2)$$

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2}{\frac{1}{89.04\%} + \frac{1}{97.01\%}} = 92.85\% \quad (3)$$

$$Sensitivity = \frac{TP}{TP+FN} = \frac{65}{65+2} = 97.01\% \quad (4)$$

$$Specificity = \frac{TN}{FP+TN^{4}} = \frac{74}{8+74} = 90.24\% \quad (5)$$

Out of the 160 segments originally selected for "Phone usage", 11 were excluded due to data quality issues. The Dataset B2 for the analysis of "Phone usage" included 149 segments, 69 segments with central frame auto-labelled as 0 and 80 segments with central frame auto-labelled as 1. Note that the Dataset B1 and B2 were different, because they were obtained from two different samples (the automatic labels for "Yawning" and "Phone usage" respectively). The confusion matrix resulting from the comparison of the manual labels performed by two students and the auto-labels made by the algorithm is reported in Table 5.5 and Table 5.6

[1] TP is the acronym for True Positive
[2] FP is the acronym for False Positive
[3] FP is the acronym for False Negative
[4] TN is the acronym for True Negative

Table 5.5 Confusion matrix for “Phone usage”.

Confusion matrix for “Phone usage”		Auto-labelled	
		0	1
Manually labelled	0	69	30
	1	0	50

Table 5.6 Normalized confusion matrix for “Phone usage”.

Confusion matrix for “Phone usage”		Auto-labelled	
		0	1
Manually labelled	0	0.70	0.30
	1	0.00	1.00

The performance metrics calculated with the formulas (1) to (5) for the analysis of “Phone usage” are reported below.

Precision = 62.50%

Recall = 100.00%

F1 score = 76.92%

Sensitivity = 100.00%

Specificity = 69.69%

5.2.3. Stage 3: Estimated performance metrics for the full auto-labelled dataset

The schematic presentation of the stage 3 is presented in Figure 5.3.

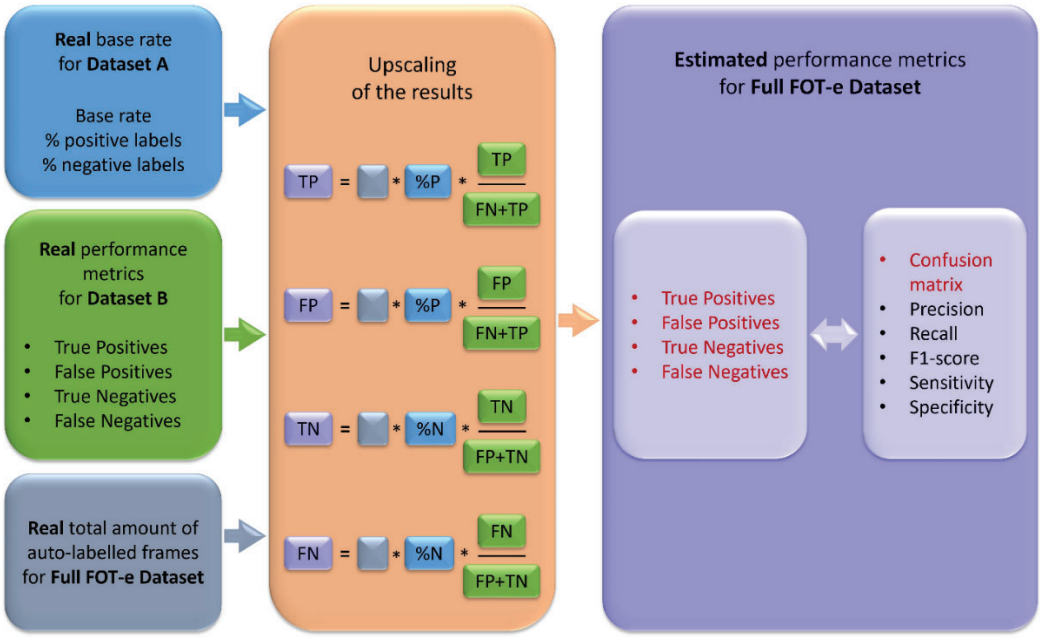

Figure 5.3 Schematic presentation of the stage 3 of the analysis. Upscaling of the performance metrics.

For the "Yawning", the expected number of positive and negative values in the FOT-e dataset is calculated below, considering that there are in total 1 238 707 530 frames in the dataset:

1. The expected number of frames which have a positive label is 1 238 707 530 * 0.05% = 619 354
2. The expected number of frames which have a negative label is 1 238 707 530 * 99.95% = 1 238 088 176

The expected numbers of TN, FP, FN, and TP are:

$$TN = 1\,238\,088\,176 * \frac{74}{82} = 1\,117\,299\,086$$

$$FP = 1\,238\,088\,176 * \frac{8}{82} = 120\,789\,090$$

$$FN = 619\,354 * \frac{2}{67} = 18\,488$$

$$TP = 619\,354 * \frac{65}{67} = 600\,866$$

The expected confusion matrix for the full FOT-e dataset after the upscaling procedure is reported in Table 5.7.

Table 5.7 Confusion matrix for "Yawning" for the whole FOT-e dataset.

Confusion matrix for "Yawning"		Auto-labelled	
		0	1
Manually labelled	0	1 177 299 086	120 789 090
	1	18 488	600 866

The performance metrics – calculated with the formulas (1) to (5) – for the confusion matrix of "Yawning" for the full FOT-e dataset are reported below.

Precision = 0.49%

Recall = 97.01%

F1 score = 0.97%

Sensitivity = 97.01%

Specificity = 90.24%

For the “Phone usage” using the base rate of 2.25%, we obtained the expected number of positive and negative values in the FOT-e dataset with the calculation below.

1. The expected number of frames which have a positive label is 1 238 707 530 * 2.25% = 27 870 919
2. The expected number of frames which have a negative label is 1 238 707 530 * 97.75% = 1 210 836 611

The expected numbers of TN, FP, FN, and TP are:

$$TN = 1\,210\,386\,611 * \frac{69}{99} = 843\,916\,426$$

$$FP = 1\,210\,386\,611 * \frac{30}{99} = 366\,920\,185$$

$$FN = 27\,870\,919 * \frac{0}{50} = 0$$

$$TP = 27\,870\,919 * \frac{50}{50} = 27\,870\,919$$

The expected confusion matrix for the FOT-e dataset for “Phone usage”, after the upscaling procedure is reported in Table 5.8.

Table 5.8 Confusion matrix for “Phone usage” for the whole FOT-e dataset.

Confusion matrix for “Phone usage”		Auto-labelled	
		0	1
Manually labelled	0	843 916 426	366 920 185
	1	0	27 870 919

The performance metrics – calculated with the formulas (1) to (5) – for the confusion matrix of "Phone usage" for the whole FOT-e dataset are reported below.

Precision = 7.06%

Recall = 100.00%

F1 score = 13.19%

Sensitivity = 100.00%

Specificity = 69.69%

6. An example of application: Head kinematics during braking in naturalistic driving

Alberto Morando – alberto.morando@autoliv.com (Autoliv Development AB)

6.1. Introduction

Tests of restraint systems can be done with volunteers and postmortem human subjects (PMHs), or with human surrogates such as anthropomorphic test devices (ATD; also known as crash test dummies) and human body models (HBM). Human surrogates can, in general, accurately mimic human kinematics and injury risk in crashes. At the same time, tests procedures need standards so that they are valid, repeatable, and reproducible. However, while a test configuration may be standard, it may deviate from how people naturally drive their car (Cullen et al., 1996; Manary et al., 1998; Reed, 2017). For example, volunteers, PMHs and ATDs may be positioned upright and at a certain distance to the steering wheel, but a regular driver may often slouch or sit closer to the steering wheel (Reed, 2017). Also, HBMs can simulate muscular response due to occupant bracing, but bracing depends on the specific traffic conflict and other human factors (Östh et al., 2013; Pei et al., 2022).

Naturalistic data and videos can provide insights on how people drive and respond to impending traffic conflicts so that testing of restraint systems is valid (Cullen et al., 1996; Manary et al., 1998; Reed, 2017). Naturalistic driving data can inform on initial conditions (e.g., sitting position) and behavior during pre-crash situations. However, crash tests also require specifications of anthropometric data and vehicle interior dimensions that cannot be retrieved from naturalistic driving data or at the required accuracy (Reed, 2017). Therefore, naturalistic data need to be eventually complemented with other data sources, such as laboratory experiments or test-track tests (Östh et al., 2013).

The goal of this analysis is to compare measurements of head kinematics from naturalistic data with the initial head position in current crash test setups. As a first proof of concept, we focused on frontal crashes; frontal crashes are the most common type of crash and are easier to reliably extract from a large naturalistic database based on vehicle acceleration. The results of this analysis can be used to validate and otherwise improve the bio-fidelity of human body models as well as identify critical scenarios for which countermeasures are needed and shall be developed.

6.2. Methods

Data are drawn from the EuroFOT and DriveC2X naturalistic driving study as processed in the FOT-e project. This analysis is based on a collection of driving segments that consist of 5 s of driving around the onset of a hard braking (safety critical event). The segments were selected according to the following inclusion criteria:

- In the 5 s before the braking onset, driving was on a straight road and in steady-state, free-flow conditions. This meant a turning radius *r* larger than 1000 m (r = speed / yaw rate), speed above 30 km/h, and a longitudinal acceleration between ±1.5 m/s^2;
- A braking that produced a deceleration of at least 8 m/s2 within 1 s after the braking started.

The videos of each event were reviewed to verify that a braking did happen so events were automatically discarded if the videos sources (front, cabin, feet, and head-tracker channels) were not available or not clearly visible. At the time of the analysis, we did not have information on adaptive cruise control (ACC) use or autonomous emergency braking (AEB) interventions.

Driving segments included data about vehicle speed, longitudinal acceleration, and head location over time. All data was recorded at 10 Hz. Vehicle speed and longitudinal acceleration were recorded from the CAN bus. Head location was estimated by a machine learning algorithm that is proprietary to Smart Eye (no extra training was done on the EuroFOT/DriveC2X data). The algorithm yields the location of the nasion (a point on the face in between the eyes in a world *xyz* reference frame with an accuracy of ~3%. The origin of the world reference system is fixed on top of the steering wheel column right in front of the head tracking camera (Figure 6.1). Because we did not have the installation schematics of the camera but just an image from an old report (Selpi et al., 2011), we approximated its location by measuring a similar vehicle (Volvo V70 MY 2008) and estimated where the camera could have been positioned (Figure 6.1). We kept these measurements fixed throughout the analysis even if drivers may have adjusted the settings of the seat and steering wheel. If drivers did make some adjustments, the settings could not be retrieved from the data anyway.

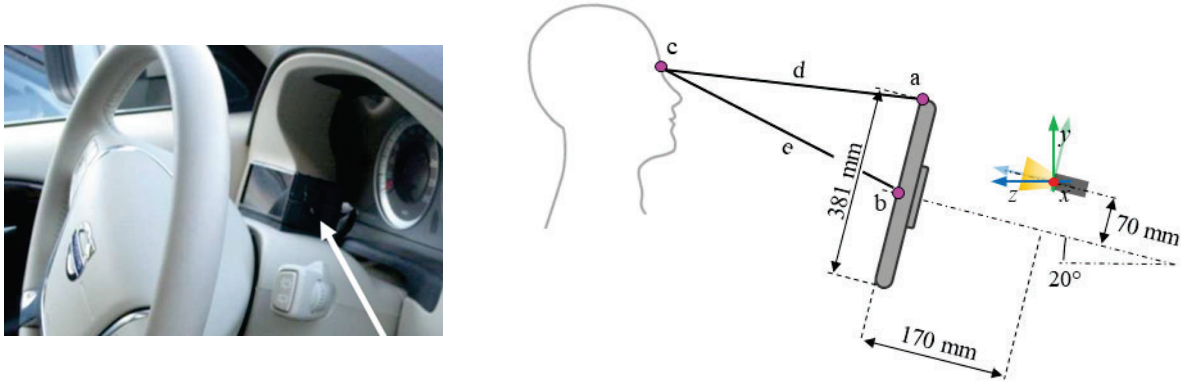

Figure 6.1. (Left) Location of the camera used for the naturalistic data collection from Selpi et al. (2011). Unfortunately, the installation details are unknown. Therefore, the camera location with respect to the steering wheel was guessed based on measurements in a similar vehicle (Volvo V70 MY 2008). (Right) Diagram summarizing the measures used in this paper to retrieve the distance *d* between the nasion and the point *a* on top of the steering wheel. The schematic is not to scale.

We retrieved the distance from the nasion and the steering wheel top and center (respectively the distance *d* and *e* in Figure 6.1) with a transformation of coordinates. Point *a* has coordinates (0, 120.5 mm, 170 mm) in a reference frame *W'* rotated by 20° around the *x* axis of the world frame *W*. Point *b* has coordinates (0, -70 mm, 170 mm) in *W'*. The location of *c* in the reference W, instead, comes directly from the head-tracking. The coordinate transformation resulted in the following equations:

$$^{W}\mathbf{d} = {}^{W}\mathbf{c} - {}^{W}\mathbf{a} \quad (1)$$

$$^{W}\mathbf{e} = {}^{W}\mathbf{c} - {}^{W}\mathbf{b} \quad (2)$$

where

$$^{W}\mathbf{a} = {}^{W}\mathbf{R}_{W'}\, {}^{W'}\mathbf{a} \quad (3)$$

$$^{W}\mathbf{b} = {}^{W}\mathbf{R}_{W'}\, {}^{W'}\mathbf{b} \quad (4)$$

with

$$^{W'}\mathbf{a} = (0, 120.5, 170) \quad (5)$$

$$^{W'}\mathbf{b} = (0, -70, 170) \quad (6)$$

$$^{W}\mathbf{R}_{W'} = \mathbf{R}(\mathbf{x}, 20°) \quad (7)$$

We used the head location data in two distinct ways. First, we used the head location in the -2 to -1 s interval to get an overall location baseline. Second, we normalized each head-location series with respect to its initial

value at the -2 s mark to get the relative head movement during braking. The normalized series were then aggregated and summarized by their median and interquartile range (IQR; 25th—75th quantiles). Quantiles are a more suitable summary statistics for noisy and asymmetric data distribution (Rousselet & Wilcox, 2019). Sometimes, the head tracking resulted in loss of data. Gaps shorter than 0.5 s were interpolated (cubic interpolation via the Matlab *pchip* command), otherwise they were kept as missing data. We excluded driving segments with a proportion of missing data above 50%.

In the end, there were 21 segments that fulfilled the inclusion criteria and had good quality data. The segments were from 17 drivers (3 females).

6.3. Results

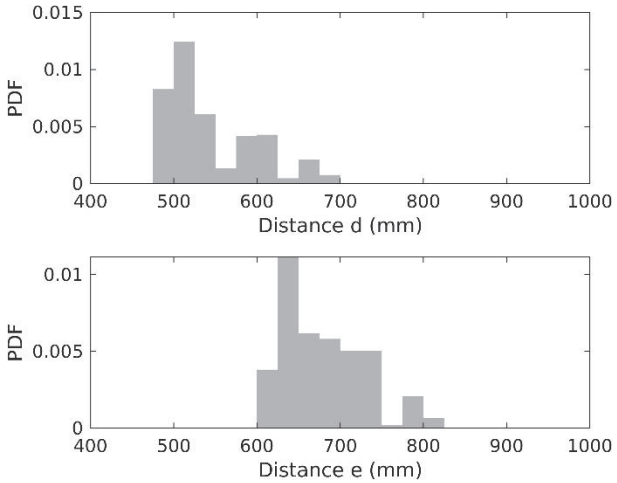

Figure 6.2. Distribution of head distance to steering wheel in the -2 – -1 s interval before the braking. Distance *d* is the distance between the nasion and the steering wheel top. Distance e is the distance between the nasion and steering wheel center.

Most drivers were attentive. In only one event, the driver was engaged in a distracting activity (texting). All drivers kept at least one hand on the steering wheel. All drivers also appeared to sit properly, without egregious deviations from a normal upright posture. In general, the median distance *d* was 520 mm (IQR = 504 – 583 mm); the median distance *e* was 655 mm (IQR = 642 – 715; Figure 6.2).

Videos review indicated that all events were braking maneuvers initiated and controlled by the driver to respond to a sudden change in the traffic environment (e.g., animals crossing, lead vehicle unexpectedly braking or changing lane). In all events, drivers braced against the steering wheel, and this limited head excursion at and after the brake onset; the IQR for distance *d* was within ±9 mm overall, while the IQR for distance *e* was within -8 – 12 mm (Figure 6.3). The time-course of the *xy* head coordinates in the *W* frame shows that the median head excursion in the *y* axis was up to 15 mm, but the respective movement in the z axis was only 0.7 mm (Figure 6.4).

Despite the head excursion being minimal, there was a trend that indicates than in the 1 s after the braking onset, drivers first pushed away and then they got closer to steering wheel as the car decelerated (Figure 6.3). This trend is more noticeable in the event with a distracted driver—a later and more rapid braking compared to the other events was associated with a larger head movement (this event is highlighted in Figure 6.3 and Figure 6.4).

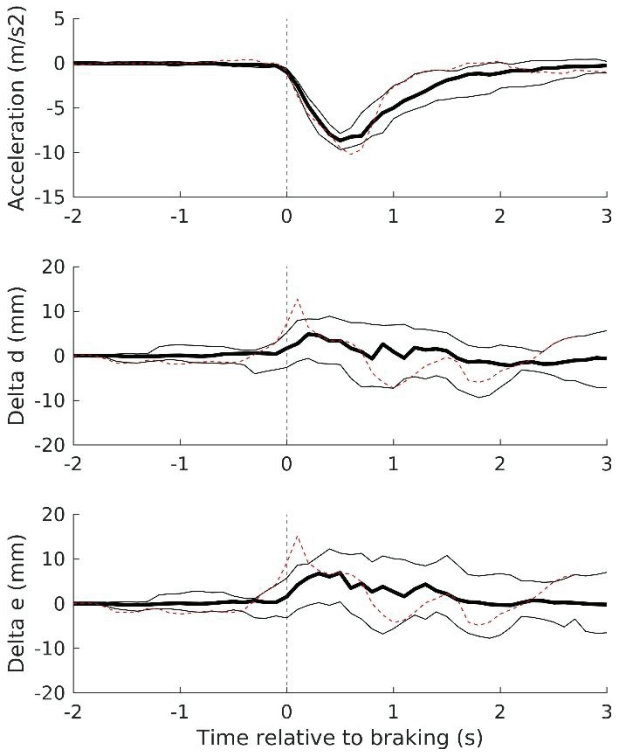

Figure 6.3. Median (thick line) and interquartile range (25th – 75th quantiles; thinner lines) of vehicle acceleration (top panel) and head distance to steering wheel relative to its initial distance (middle and bottom panel). Delta *d* is the displacement of the nasion with respect to the steering wheel top. Delta *e* is the displacement between the nasion and steering wheel center. In each panel, the vertical dashed line in each panel represents the start of the braking. The graphs x-axes indicate the relative time to the braking start. The event involving a distracted driver is highlighted with a red dashed line.

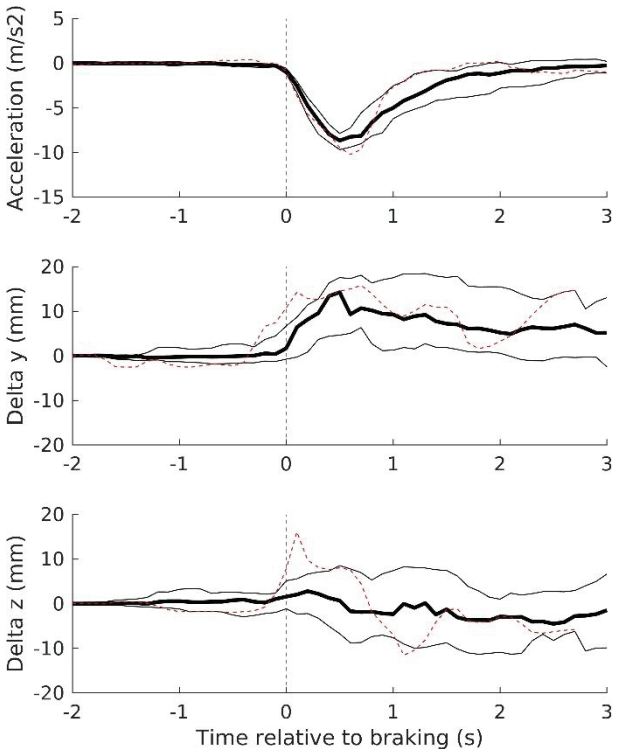

Figure 6.4. Median (thick line) and interquartile range (25th – 75th quantiles; thinner lines) of vehicle acceleration (top panel) and head location in y/z coordinates relative to its initial location (middle and bottom panel). In each panel, the vertical dashed line in each panel represents the start of the braking. The graphs x-axes indicate the relative time to the braking start. The event involving a distracted driver is highlighted with a red dashed line.

6.4. Discussion and conclusions

Observations of natural drivers position found an average separation between the nasion and the steering wheel top (distance *d* in Figure 6.1) of 581 mm (standard deviation = 64 mm), and a median separation between the nasion and the steering wheel center (distance *e* in Figure 6.1) of about 500 mm (Cullen et al., 1996; Parkin et al., 1995). Our measurements deviate from these early observations, and the greater deviation is for the distance *e*. Perhaps, this discrepancy is due to drivers and vehicles type distribution (Cullen et al., 1996; Parkin et al., 1995), which highlights the complexity of defining a minimal set of test procedures that can be representative of a large population of drivers and vehicles.

The typical position of the latest SAFER HBM (Pipkorn et al., 2023) in a frontal crash test is so that d = 500 mm and e = 600 mm. While the head distance in a SAFER HBM deviates from early naturalistic observations (Cullen et al., 1996; Parkin et al., 1995), it is reasonably close to our current findings (Figure 6.3). It is unknown what is the safety impact of a delta in the range of 20 – 55 mm between the ATD's configuration and a natural driver posture, but this could be investigated in future analyses.

Data from a test track experiment in which drivers braked to a full stop with the maximum braking power of the vehicle (about 1 g) showed an average head displacement in the *yz* direction of about -8 mm and 50 mm, respectively (Östh et al., 2013). These results are not consistent with the general trend we found. The displacement along *y* was of similar magnitude, but the opposite sign; the displacement along *z* was larger (Figure 6.4). This discrepancy could be explained by a more pronounced upper body movement (and rotation) due to a stronger (and longer) braking maneuver in the tests by Östh et al. (2013). In general, the deceleration in our events was maintained for less than 1 s (Figure 6.3), whereas the deceleration in the tests by Östh et al. (2013) lasted for about 2 s. Because it is difficult to find a large sample of strong and sustained braking events in our naturalistic data, future work may be required to compare our results to brake pulse tests or simulations.

In the past, researchers found that ATDs were positioned more forward relative to humans (Cullen et al., 1996; Parkin et al., 1995). This discrepancy was problematic because restraint systems are designed based on crash performance predicted from the initial seating positions of ATDs. Our results show that nowadays crash simulations are a much closer representation of real-world driving. However, the naturalistic data we used lacked attributes such as driver anthropometry and cabin geometry. A new naturalistic data collection with accurate calibration, as done by Reed et al. (2017), would be required to obtain such granular measurements. Regardless, naturalistic data not collected for the specific purpose of restraint system design remains a valuable source of information to ensure that crash tests and simulations are close to real world driving.

7. General project discussion and conclusions

At SAFER, an updated FOT database is now available where two features (phone use and yawning) are automatically labelled by the classification algorithm developed and tested within the project. The results obtained from the evaluation of these two features show that the automatic labelling can be successfully used to extract data from the whole FOT-e dataset. The normalized classification matrices show small rates of false positive and false negative for yawning. With respect to phoning, the rate of false positive is higher than for yawning, but this is not considered a limitation given the future use of the classification algorithm (i.e., use the classification algorithm to select relevant data, which need to be manually labelled by annotators). For that, it is rather important to limit the rate of false negative which would cause missing classification of phone use in the FOT-e dataset. The evaluation of the phone use feature shows that this is achieved since the rate of false negative is very low. The automatic labelling of the data conducted within the project will save time in analyses related to yawning and phone use because it will facilitate the identification of these features in the whole dataset. This has resulted in an improved and better dataset through semi-automatic labelling, available for future research at the SAFER community.

In industrial usage, for passive safety positioning benefits, the naturalistic data is the gold standard for driver behaviour analysis and testing of active safety systems. Here, we can use this data source to inform testing and design of restraint systems that are usually carried in lab experiment or well constrained close track tests. Off the shelf algorithms can be used on old data to pull out new features for safety research. That is, the analysis of head distance did not need a new, long, expensive data collection. New features from FOTe have also a great potential to study driver posture in normal driving and in critical situations. However, those features need to be calibrated (e.g., from body key points in the image reference frame to key points in a world coordinates). The calibration could be eased in future study by equipping vehicles with calibrated camera and markers (e.g., to keep track of seat, seatbelt, steering wheel position over time). The information in the collected naturalistic data can be used to improve testing of restraint systems, so that they are close to real world driving situations and they can help identifying needs to develop new safety systems altogether. However, naturalistic data lacked attributes such as driver anthropometry and cabin geometry that are key information to set up standard crash simulations. Naturalistic data collection with accurate calibration of passengers and cabin dimension can be collected, but it is a considerable effort (Reed et al., 2017). Regardless, even naturalistic data not collected for the specific purpose of restraint system design, remains a valuable source of information to ensure that crash tests and simulations are close to real world driving.

For the eye-tracker industry, the result can be used in lower-level automatic annotated features, to develop algorithms for higher level features, such as drowsiness from yawning and eye closure. The improved data can use the derived process/methodology for auto annotation for other datasets, to get more labelled datasets, or other features. It is also possible to use this experience to see what features on low-resolution videos can be reliably and not reliable auto annotated data, and use this experience for limitation of feature annotations, both manual and automatic labelling.

For future research, the partners have suggested to improve the 3D information of body posture of the driver. Regarding camera quality and positioning, future work should focus on improving camera resolution and correct positioning/camera view, also during night-time. There by new positions would increase the possibility to classify hands-on wheel. Future research should also focus on the evaluation of the feature viewing target which is extremely important for research on driver behaviour. The processing is time consuming, so the time proportionally increases with the number of features selected due to the extensive time required for manual labelling. Therefore, efforts should be taken to optimize the algorithms. Future efforts should also evaluate access to streaming data, and how to process data in embedded systems, or in uploaded in real-time.

References

Bärgman, J., van Nes, N., Christoph, M., Jansen, R., Heijne, V., Dotzauer, M, Carsten, O, et al. (2017). UDrive deliverable D41.1: The UDrive dataset and key analysis results. EU FP7 Project UDrive consortium. Brussels, Belgium.

Cullen, E., Stabler, K. M., Mackay, G. M., & Parkin, S. (1996). *How people sit in cars: Implications for driver and passenger safety in frontal collisions—The case for smart restraints*. 77–93.

Dingus, T. A. (2014). Estimates of prevalence and risk associated with inattention and distraction based upon in situ naturalistic data. *Annals of advances in automotive medicine, 58*, 60.

Dozza, M., Moeschlin, F., & Léon-Cano, J. (2010, August). FOTware: A modular, customizable software for analysis of multiple-source field-operational-test data. In *Second international symposium on naturalistic driving research*. August 31-September 2, 2010, Blacksburg, VA.

Manary, M. A., Reed, M., Flannagan, C. A. C., & Schneider, L. W. (1998). *ATD Positioning Based on Driver Posture and Position*. 983163. https://doi.org/10.4271/983163

Östh, J., Ólafsdóttir, J. M., Davidsson, J., & Brolin, K. (2013). *Driver Kinematic and Muscle Responses in Braking Events with Standard and Reversible Pre-tensioned Restraints: Validation Data for Human Models*. 2013-22–0001. https://doi.org/10.4271/2013-22-0001

Parkin, S., Mackay, G. M., & Cooper, A. (1995). How drivers sit in cars. *Accident Analysis & Prevention*, *27*(6), 777–783. https://doi.org/10.1016/0001-4575(95)00035-6

Pei, X., Gan, S., Li, Q., Zhou, Q., Wang, J., & Nie, B. (2022). *An Experimental Framework of Capturing Driver's Pre-Crash Active Behavior under Safety-critical Scenarios*. IRCOBI conference 2022.

Pipkorn, B., Jakobsson, L., Iraeus, J., & Östh, J. (2023). *The SAFER HBM – A Human Body Model for Seamless Integrated Occupant Analysis for All Road Users*. 27th International Technical Conference on the Enhanced Safety Vehicle (ESV), Yokohama, Japan.

Reed, M. P. (2017). *Upper Extremity Postures and Activities in Naturalistic Driving*. University of Michigan, Ann Arbor, Transportation Research Institute. http://deepblue.lib.umich.edu/handle/2027.42/136218

Rousselet, G., & Wilcox, R. (2019). Reaction times and other skewed distributions: Problems with the mean and the median. *Preprint*. https://psyarxiv.com/3y54r/

Selpi, Borgen, S., Bärgman, J., Svanberg, E., Dozza, M., Nisslert, R., Norell, C., Kovaceva, J., Sanchez, D., Saez, M., Val, C., Küfen, J., Benmimoun, M., & Metz, B. (2011). *EuroFOT Deliverable 3.3—Data management in euroFOT*. http://www.eurofot-ip.eu/

SHRP2 (2016). SHRP2 researcher dictionary for video reduction data. Version 3.4. Retrieved from: https://vtechworks.lib.vt.edu/bitstream/handle/10919/56719/V4.1_ResearcherDictionary_for_VideoReductionData_COMPLETE_Oct2015_10-5-15.pdf?sequence=1&isAllowed=y.

Appendix A: Auto-labelled features description

Eye openness

Eye openness or Eyelid opening, describes how open the driver eyes are in a frame. It is a floating-point value in the range [0.0, 1.0]. The value is predicted by a neural network, and then min-max normalized per recording. No extra training has been done for this feature for FOT-e data.

Yawning

Yawning output contains a Boolean stating whether the driver is yawning. The value is predicted using a statistical model, with the inputs *mouth opening value, eye openness value, phoning value,* and *laughing value.*

Viewing target

The viewing targets module provides information on what the driver is currently looking at. Viewing targets are geometrical shapes to represent the environment around the driver. In this case we used 4 planes: left, right, forward, and down (see Figure 0.1). For every target plane the output is a Boolean, where a value equal to one represents an intersection point between the viewing direction and a viewing target. Only intersections that have a higher probability than 0.5 are counted. In absence of intersection with the defined planes or high-quality intersection *eye_target_unknown* is reported. No extra training has been done for this feature for FOT-e data, however, the world model has been updated to fit the recording better.

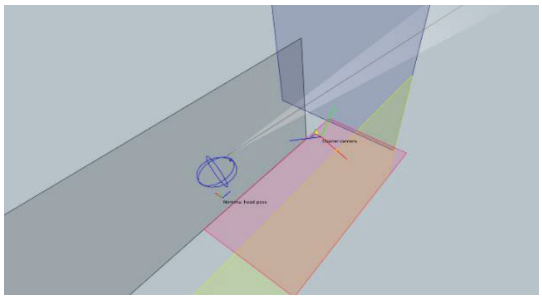

Figure 0.1 Viewing targets as geometrical shapes representing the environment around the driver.

Head Pose

Head pose output includes head position and orientation. The head position(x,y,z) is expressed in reference coordinate system and the head orientation is defined around nominal head pose and expressed in heading, pitch, and roll.

- Rotation around y axis of nominal head pose (heading) represents looking left or right, with positive values representing right.
- Rotation around x axis of nominal head pose (pitch) represents looking up or down, with positive values representing down.
- Rotation around z axis of nominal head pose (roll) represents tilting left or right, with positive values representing right.

The pose is predicted using a neural network and no extra training has been done for FOT-e data.

Phone usage

Phone usage attribute states whether the tracked subject is talking on a phone or holding it to the ear with the left/right hand. The value is predicted using a neural network.

Body key-points

For estimating the driver posture, we use a neural network that estimates human pose, by predicting 15 core body key-points in pixel coordinates with the origin in the top-left corner. The 15 core points are: Nose, Two Eyes, Two Ears, Two Shoulders, Two Elbows, Two Wrists, Two Hips, Two Knees. For each key-point a quality factor is also predicted. No extra training has been done for FOT-e data.

Out of position

Out of position is a binary signal indicating whether a person is out of position compared to the neutral seating position. Situations considered out of position are

1. If the head is shifted by at least 100mm to either side from the centre of the seat headrest.
2. If the head has a distance of at least 100mm to the seat headrest.
3. If the head is lateral shifted by at least 100mm and longitudinal 100mm away from the seat headrest.

For predicting head position, the most important key-points are the torso key-points. As the drivers' body and the camera pose is different across different recordings, normalizing torso points considering the in-position pose in each recording improves the algorithm generalization. One could assume that in a recording most of the time the driver is in-position, so the mean of high-quality torso key-points can be assumed to be the in-position body pose. After the normalization, a statistical classification model (Random Forest) has been trained with the normalized torso points to predict the in-position and out-of-position.

It is note-worthy that for this feature, we only had 10 mins of manually labeled data, so the evaluation results might not be very generalizable.

Hands off steering wheel

As all the cars used in the dataset have the steering wheel on the left side, one could draw a steering wheel polygon region and predict hands-off-wheel using that region and the hands position.

As the camera pose is different across different cars, the steering wheel regions needs to be assumed in a conservative way.

If the wrist point's distance to the steering wheel polygon is more than the size of a hand, the hand is set as off the wheel. If it is inside the region or closer than a hand-size to the region, the hands-off-wheel prediction is set to unknown for both hands. No extra training has been done for this feature for FOT-e data.

Frame quality for DM

The frame quality is defined as if the recording frame quality has been acceptable enough so that algorithm has been able to predict the head bounding box.

Frame quality for CM

The frame quality is defined as if the recording frame quality has been acceptable enough so that algorithm has been able to identify the object on the driver seat as a person.