Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 151 (2019) 92–99

Procedia Computer Science

www.elsevier.com/locate/procedia

The 10th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 29 - May 2, 2019, Leuven, Belgium

Predicting Traffic Phases from Car Sensor Data using Machine Learning

E. Heyns[a], S. Uniyal[a], E. Dugundji[b], F. Tillema[a], C. Huijboom[a]

[a]*HAN University of Applied Sciences Automotive Research, Ruitenberglaan 29, 6826CC Arnhem, the Netherlands*
[b]*Vrije Universiteit Amsterdam, Faculty of Science, Mathematics, De Boelelaan 1105, 1081 HV Amsterdam, the Netherlands*

Abstract

This research is an explorative study to look for the potential to predict traffic density from driver behaviour using signals collected from the Controller Area Network (CAN) bus. The hypothesis is that driver behaviour is influenced by traffic density in such a way that an approximation of the traffic density can be determined from changes in the driver behaviour. Machine learning will be employed to correlate a selection of commonly available sensors on cars to the traffic density. Challenges in the processing of the data for this purpose will be outlined. The data for this study is collected from five passenger cars and nineteen trucks driving on the A28 highway in Utrecht region in the Netherlands. This study is restricted to straight roads in order to isolate the steering behaviour attributable to the traffic state influences rather than following the curve in the road. The results are encouraging that the correlation between driver behaviour and traffic density can be established. An overall accuracy of over 95% is achieved with a precision of 92%. The recall rate however is low most likely caused by over-fitting due to the unbalanced data set. The results still look promising and more training data should improve the results. This research is part of the broader project VIA NOVA which aims to investigate the use of car-sensor data for traffic and road asset management.

Keywords: Congestion; Traffic Density; Driver Behavior; CAN Bus; Probe Vehicle Data; Supervised Machine Learning;

1. Introduction

Traffic jams and road safety problems are rising with the increasing number of vehicles on the road. Traffic congestion leads to negative health impact, waste of fuel and unproductive hours [1], and road authorities are under increasing pressure to manage the effects. In 2017, traffic congestion cost an estimated total of €80 billion in the European Union

* Corresponding author. Tel.: +31-6-4683-6010
E-mail address: emiliano.heyns@han.nl

1877-0509
10.1016/j.procs.2019.04.016

and $305 billion in the United States. So, tackling traffic congestion is imperative to reduce wastage of energy and resources. With the advance of technology, we can now detect jams effectively. It would be better however to predict them in advance, so preventative measures can be deployed. This research, part of the broader VIA NOVA project investigating the use of car-sensor data for traffic and road asset management, presents the results of using the driver behaviour signals obtained via commonly available vehicle sensors to detect the traffic phase that is likely to precede congestion. Driver behaviour here refers to direct driver input to the vehicle. The driver behaviour would be speed changes, operation of throttle and brake pedals and steering angle changes. The driver behaviour is hypothesized to change or be influenced with change in traffic density. Under the three-phase traffic theory [6], traffic flow knows three recognizable phases of traffic density (free flow, synchronised flow and wide-moving jam) which typically follow each other as the traffic density rises. If a detectable behaviour correlates with these ranges in the traffic density, a detection of behaviour associated with the synchronized flow could function as an early warning for the wide-moving jam phase. This study intends to use driver behaviour changes to predict traffic density. For this study the road under consideration would be a highway. There has been significant research in the field of driver behaviour [4, 3, 10, 14, 7]. Most of this research aims at improving driver behaviour and increasing safety. Driver behaviour is influenced by traffic conditions and surroundings [1, 10, 5, 8]. Highway driving is characterized by both longitudinal and lateral manoeuvres [7]. The lateral behaviour, mainly high steering rate, can be a reflection of gap-seeking behaviour indicative of rising traffic densities. Ito and Kaneyasu [5] use neural networks to predict traffic congestion from driving input signals (driver behaviour) with an average accuracy of 81.65% in detecting traffic phases using data from driving simulators. In this study, we will use naturalistic driving data to test a diversity of machine learning algorithms to see which different approaches yield better detection rates. The traffic density to estimate traffic phase is obtained using openly available traffic information data from NDW (Dutch National Data Warehouse for Traffic Information). Driver behaviour is of course not the only factor in the occurrence of congestion. Traffic is also affected by rush hours, type of road, and the weather. To isolate the driver behaviour component, for this study we have chosen a stretch of highway with few on/off ramps and no curves. This helps in minimizing the effect of road interference and intersections and minimizes the assumption errors arising due to instantaneous changes in traffic density due to on/off ramps.

2. Data Procurement and Analysis

The data for this research has been taken from data collected by the company SD-Insights in the Netherlands. A selection was made for cars driving on a 12 km section of the Dutch A28 highway which was frequented by multiple SD-Insights cars at multiple times. This is a straight road section with just two major exit points in between. We chose this section because the major influence on driving behaviour and traffic density would be the number of cars on the road rather than perturbations from on/off ramps and curve-following. The A28 is one of the major motorways in the Utrecht region of Netherlands. Traffic information is provided as open data by Dutch National Data Warehouse for Traffic Information (NDW).

The trips selected for this research spans January 2018 to March 2018. The selection includes a total of 24 separate vehicles; 5 passenger cars (27 trips across the selected road section) and 19 trucks (100 trips across the selected road section). All data was sampled at 10Hz. The initial goal was to mainly use data from passenger cars, but SD-Insights monitors primarily trucks. The collected data contained sensor data read from the CAN bus but did not contain driver demographics or make/model of the vehicle for reasons of privacy and security. It is known that driver characteristics such as age and gender have statistically significant influence on driving behaviour [2]. While it is our expectation that adding this information to the data set would further increase the accuracy of the flow phase detection, such information is not available as car sensor data, so it is not included. Table 1 shows the available in-car signals and the signals selected for traffic detection in this study.

The selection aims to capture direct driver input without the use of data that represents the same information twice. Brake use is comparatively rare on highways [2]; brake signals are also not available for all vehicles, so brake behaviour is not used in this study. Throttle is selected as throttle use is expected to differ between free flow and jam phases, as we expect the stop-and-go characteristic typical of the jam phase to show up in the data. Also, the throttle signal gives us indications of the use of Active Cruise Control (ACC), commonly in use on highways. The steering angle signal is not available for all vehicles, so yaw rate is selected as an alternate. Wipers and turn signals are not used in this study. Future phases of this study could take weather conditions into account as inferred from the ambient

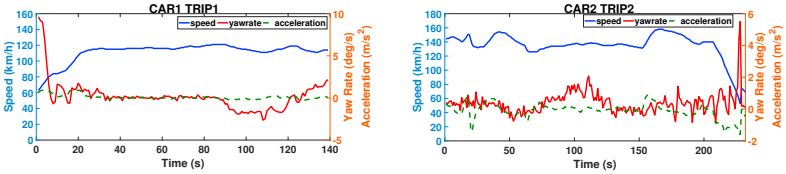

Fig. 1. Speed, Yaw and Acceleration of Cars

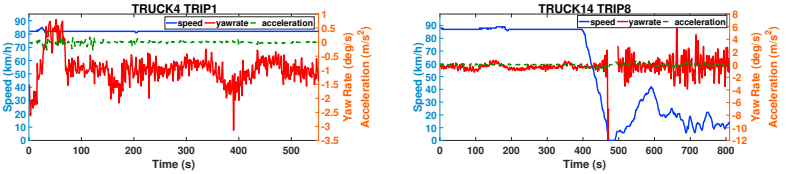

Fig. 2. Speed, Yaw and Acceleration of Trucks

temperature, wipers and fog lamp data available on the CAN (Controller Area Network) bus. Additionally, we feed the model the vehicle type (passenger car vs truck); as we will later see, these exhibit very different behaviours.

A first analysis shows the vehicle data occasionally has gaps where for a period of time, data was either not available or not stored. This could be because of loss of signals or sensor malfunction. These gaps in the data will affect the accuracy of the outcomes, especially since some of these gaps occurred in the density regions of interest (synchronized flow/wide moving jam). Twelve of the 127 trips have multiple gaps. The gaps range from a few seconds up to one hundred seconds. This could lead to potential over-fitting as some vital information about phase changes could be missing.

Table 1. In-Car signals available in data

Signal	Selected for traffic detection	Signal	Selected for traffic detection
Timestamps	Yes	Acceleration	Yes
GPS	Yes	Wipers	No
Speed	Yes	Steering Angle	No
Throttle	Yes	Brake	No
Yaw Rate	Yes	Turn Indicator	No

Figures 1 and 2 show the speed, acceleration and yaw rate for a selection of trips (shown here to highlight different types of behaviour) of passenger cars and trucks. Figure 1 shows examples of behaviour of passenger cars on the selected highway section; Figure 2 shows the same for trucks. Just by visual inspection, it can be seen that the signal patterns and limits are different for cars and trucks.

First visual inspection leads us to believe a correlation between traffic state and driver behaviour can be found in these sensor data. For example, Figure 2b could indicate a traffic breakdown.

2.1. Traffic Phase from NDW Data

The traffic information data for determining traffic density is obtained from the Dutch National Data Warehouse for Traffic Information which offers historic open traffic data for the Dutch highway network.

On the road section selected for this study, each loop is approximately 400 m apart, for each car measurement we take the closest loop to be representative of the traffic condition at that time; depending on vehicle speed this means the loops are at most 60 seconds away from each other. Because of the spacing of the loops, not every timestamped car measurement had a corresponding loop measurement; such gaps in the loop data were filled in by extrapolating from the last-known value. If a gap of over 30 seconds occurs at the beginning or the end of the data, they are removed.

For this explorative study, results from traffic simulation using Improved 2D Intelligent Driver Model (2D-IIDM) [11] are used as reference; the density range between 22 and 37 vehicles per kilometer per lane is deemed synchronized flow by [11]; anything below 22 is deemed free flow, and anything above 37 is deemed wide moving jam. The results

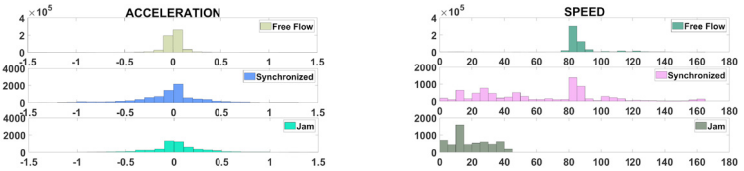

Fig. 3. (a) distribution of acceleration values (b) distribution of speed values

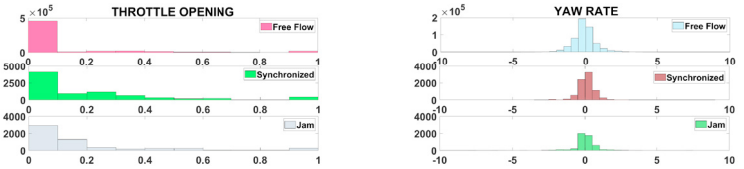

Fig. 4. (a) distribution of throttle values (b) distribution of yaw values

from [11] are for homogeneous traffic; the values for $k1$ and $k2$ will differ for real heterogeneous traffic. This can be observed in a 2D-IIDM-based simulation [13] by Treiber and Kesting [12]. Knowing this, using the values from [11] will introduce an unknown error in determining exact phase change point based on density as some assumptions or parameters from [11] may not apply on data used in this study. At later stage or further research, the actual critical densities for the road sections under study must be calculated from the NDW data, but time constraints prevented inclusion of these in the current study.

With the ranges so defined, the occurrence of traffic phases in each trip is determined to check the distribution of data points across each traffic phase. It is found that the number of data points for free flow are overwhelmingly more than those for synchronized and wide moving jam. This distribution gives a high chance of over-fitting as the model will learn more about free flow and will not be able to accurately detect other phases given new data.

3. Data Analysis and Pre-processing of input signals

Since we cannot assume the sensors to be fully reliable, the first step is to select the data for training the model. To this end, all data files are checked for consistency of data: whether they contain all required signals, and whether there are too many missing values. The throttle signal is min-max normalized as different vehicles have different throttle pedal position ranges.

Figures 3 and 4 show that the number of occurrences of acceleration, speed, throttle and yaw behaviour show different patterns in different phases, which would indicate that machine learning should be able to find these correlations.

Figure 4a shows more variance for throttle in the jam phase as there is continuous use of the throttle pedal to close any gaps as they occur. The value 0 for throttle in free flow and synchronized flow may be due to the use of ACC in most vehicles, but for jam the value 0 is expected as the vehicles would stop and go in congested traffic.

In order to optimize detection accuracy further, feature extraction is used to add more features for the classification learner. This is done using Reconstruction Independent Component Analysis (rica) and Sparse Filtering (sparsefilt). The features were selected by adding features one by one until there was no significant increase in accuracy as compared to the increase in training time with higher dimensionality. Using this procedure, a total of six new features are extracted to avoid increasing dimensionality too far which would significantly increase training time. ‘rica’ creates a linear transformation of input features and ‘sparsefilt’ creates a non-linear transformation of input features to output features. Three sets of training data are prepared. One with initially selected input features and two other prepared using ‘rica’ and ‘sparsefilt’.

We tested for weather influences for our driving data set but found that adding it as a feature had no siginificant effect on the results, and have therefore omitted it from our analysis. We do anticipate however that for larger data sets spanning more time and more drivers, an effect is likely to be found.

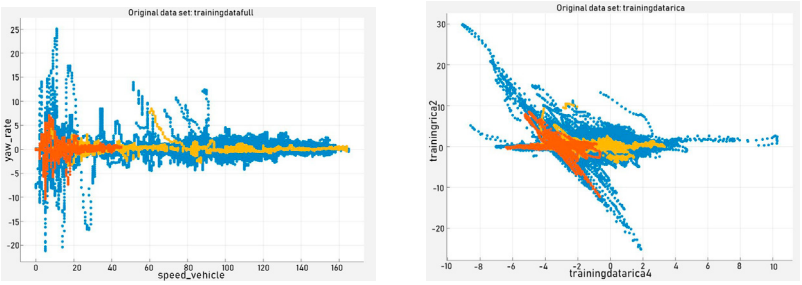

Fig. 5. Scatter plot of two predictors. Blue-Free Flow, Yellow- Synchronized Flow, Orange-Wide Moving Jam. (a) dataset 1 (b) dataset 2

4. Training of Predictive Model

In this study, we use an ensemble learning approach using Bagged trees and RUSBoosted trees. RUSBoosted trees are recommended for imbalanced data set [9]. Accuracy is improved by testing different configurations for the Bagged and RUSBoosted algorithms. The configuration includes parameters such as learning rate, maximum number of learners, and number of splits. Input features are removed or added, and the results are checked for improvement. The parameters of the ensemble algorithms and their effects on training are described below:

- The learning rate was varied between 0.1 to 1. A learning rate of 0.1 would take longer to train but would usually achieve higher accuracy.
- The maximum number of splits control the depth of tree learners. The number of splits was gradually increased as too many branches can lead to over-fitting.
- The number of learners can be increased to achieve higher accuracy at the cost of longer training times.

The parameters are tuned gradually to avoid over-fitting and keep the training time as low as possible. Multiple rounds of training are conducted with different configurations and algorithms. Besides the ensemble algorithms, other algorithms are also tried. All the results from these algorithms are analysed and compared. The algorithms which produced the best results are trained again with multiple configurations. Each configuration is also trained after performing Principal Component Analysis (PCA) on input features to see if that improves results. This is repeated until the best results are achieved from a specific configuration, with minimum training time.

The same process is repeated with the other two data set created using feature extraction. The extracted features are also normalized during feature extraction, so all algorithms which perform better after normalization are also used for training to check if higher accuracy is achieved. Different algorithms have different configuration options which are also changed, and multiple training simulations are run. They are not described here as they produced low accuracy results and ensemble algorithms still performed better. So, ensemble algorithms are trained multiple times with different configurations to achieve best possible results.

5. Results

Three data set are used for training. The first set is the one obtained after combining all selected input features (speed, yaw rate, acceleration and throttle). Speed is the major predictor as it explains most of the differences between all inputs. PCA was used and 99.7% variance could be explained with first principal component, but the results did not improve by using PCA. The other two sets are prepared by using functions rica and sparsefilt.

Many training simulations were run using all three data set. The ensemble algorithms (Bagged and RUSBoosted trees) were used. Other algorithms were also tried but their results were not significant (very low accuracy), therefore they are not described here. As expected the imbalance in the data did affect the results. Figures 5a and 5b show scatter plots of two predictors with different colours indicating different traffic phases of first and second data set respectively. It is clearly visible that the data points for free flow (blue coloured) are more in number.

The ensemble algorithms were tuned using different parameter values to achieve the best possible results given the data set available. The most optimum results of data set are formulated in Table 2 (precision and recall rate for

synchronized flow). The results of third data set (sparsefilt) have been omitted here as they displayed very poor results. Precision or positive predictive rate indicates the accuracy with new data or validation data. Recall rate or true positive rate indicates accuracy of training data.

Table 2. Accuracy achieved with precision and recall rate of synchronized flow phase.

Algorithm Used	Configuration		Results	
First data set				
Bagged Trees	No. of Learners	200	Accuracy	99.3%
	Max. No. of Splits	50000	Time	8053.6 secs
			Precision	92%
			Recall Rate	65%
RUSBoosted Trees	No. of Learners	250	Accuracy	96.7%
	Max. No. of Splits	1500	Time	982.15 secs
	Learning Rate	1	Precision	33%
			Recall Rate	82%
Second data set				
Bagged Trees	No. of Learners	200	Accuracy	99.2%
	Max. No. of Splits	100000	Time	15304 secs
			Precision	62%
			Recall Rate	91%
RUSBoosted Trees	No. of Learners	400	Accuracy	98.7%
	Max. No. of Splits	2000	Time	2491.4 secs
	Learning Rate	1	Precision	64%
			Recall Rate	75%

For the first data set using bagged trees, it is observed that the recall rate is low for synchronized flow, but the precision is high. The jam phase shows high recall and precision, this is due to it being easily distinguishable by the major predictor speed. The model is unable to correctly learn the synchronized phase due to too few data points for this phase against the large number of free flow points, and over-fitting occurs. As a result, some of the data points from synchronized class are also classified as free-flow. Also, as we can see in the scatter plot, free flow and synchronized flow coincide at almost all points so other predictors need to be used for differentiating, and since there are so much more data points for free-flow the model by default assigns free-flow to most data points when it is confused. The large number of data points for free flow is also the reason overall accuracy is not the real performance indicator because even if the other two classes (synchronized and jam) have low number of true positives, the overall accuracy would still be high if free flow is correctly predicted. The precision of the model is pretty good; out of all data points recognized as synchronized, 92% are correctly classified. A traffic manager would likely prefer high precision, so that the chances of unnecessarily diverting traffic due to false detection of synchronized flow are low.

Using RUSBoosted trees algorithm with first data set did improve the recall rate of synchronized flow phase but it also decreased the precision; that is, a greater number of synchronized flow points were correctly predicted during training. This is the result of random under sampling of the free flow class, balancing the data set. But this also results in less training of free flow data points and in turn the model predicts more free flow points as synchronized. Therefore, the precision of the model decreases, and the model performs poorly with new data due to over-fitting.

For the second data set (rica), using bagged trees did not improve the results (Table 2) over the original data set and the training time also increased due to higher dimensionality with six features. Using RUSBoosted trees however, did improve the precision but decreased the recall rate as compared to data set 1. This indicates that the model performed a little better but still lacks some information needed which could be corrected with more sample data.

Figure 6 shows the confusion matrices for results of first data set using bagged trees for reference. Figure 6(a) shows the main confusion matrix indicating the number of correctly predicted observations. Figure 6 (b) and Figure 6 (c) show the True Positive Rate (Recall Rate) matrix and Positive Predictive Rate (Precision) matrix respectively.

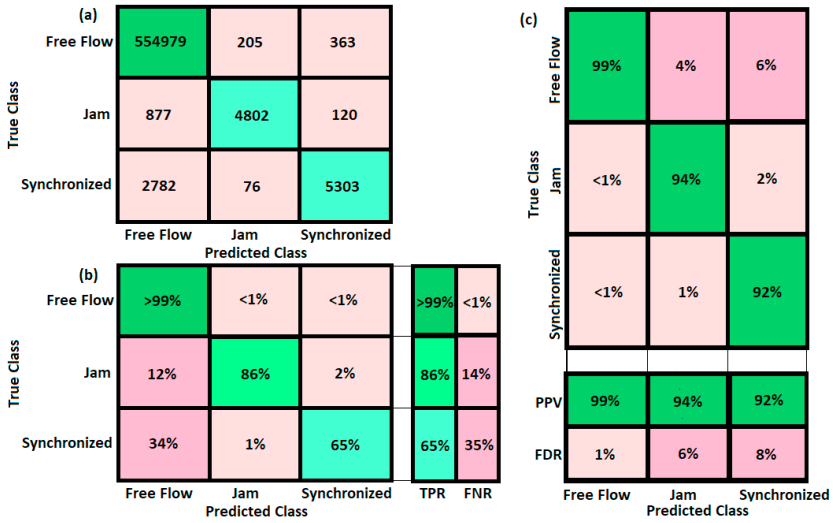

Fig. 6. (a) Confusion Matrix (b) True Positive Rates Matrix (c) Positive Predictive Rates Matrix (Using bagged trees for data set 1) TPR-True Positive Rate, FNR-False Negative Rate, PPV-Positive Predictive Value, FDR-False Discovery Rate

5.1. Training results

Even though the speed input largely dominates the other inputs, the other input features still relevantly influence the classification, as removing those decreases accuracy. The imbalance in the data is very evident in the results, but the model is still able to classify with reasonable accuracy (high precision). More data, specifically from congestion phases, is needed to further improve classification results. Feature extraction did not help as there is no significant improvement in classification performance. This could mean that the maximum possible performance is achieved given the original data set. High recall and precision could not be achieved simultaneously. This is caused by the insufficient quantity of data. The high precision with bagged trees shows that the model performs reasonably well with new data. Although high recall (sensitivity) is achieved using RUSBoosted trees, it fails to perform well with new data, as the precision decreases. So, there is trade-off here between recall and precision. For this problem, high precision seems more desirable as the objective is to detect synchronized phase as a predictor (precursor) to the jam phase. With high recall many free flow data points are also recognized as synchronized, which would lead to wrong prediction of the jam phase, that is, if a model such as this is implemented, and traffic is redirected based on this model then there is a high chance that the model detects free flow as synchronized and traffic is unnecessarily diverted.

6. Limitations

The data collected came mostly from trucks (19) compared to passenger vehicles (5). 87.5% of the collected data is of trucks. This means the model as it is currently trained will be more effective in detecting traffic states through truck behaviour data. The model could be improved by retraining it with more data from passenger cars, as this would better reflect the typical traffic mixture on highways. The traffic states for training of the model have been approximated by using results of a 2D-IIDM-based traffic simulation. Using the actual road configuration would further improve accuracy and usefulness of the model.

7. Conclusions

The training results show that traffic phases can be classified using driving behaviour. Driver behaviour changes as the traffic phase change and these changes can be correlated to these traffic phases using machine learning. The results of this explorative study are encouraging that the correlation can be established between driver behaviour and traffic

phases, but a larger data set and more relevant data (synchronized and jam phase) is needed. The data used in this study is biased, the effects of which is evident in results. Free flow detection has high accuracy because much more data is available to describe it. Although the recall of synchronized flow of the model is low, it shows high precision, which is good for our purposes as the goal is to detect precisely the synchronized flow, and the trained model can detect it with high accuracy. That is, what the model classifies as synchronized flow is mostly correct (92% in results). High recall is also achieved using a RUSBoosted trees algorithm, but in this case the model detects many free-flow points as synchronized, which is not desired. This study thus illustrates a way to correlate driver behaviour with traffic phases and presents a technique to use machine learning and detect traffic phases through driver behaviour. It is concluded that with more relevant data, this technique could be very useful and more reliable.

8. Future directions for research

The learning model can be improved using more data specifically with more observations from non-free-flow phases. Another suggestion would be to include steering angle or steering rate input, which would more directly take direct driver input into account and might produce better results. Also, using headway as an input variable could potentially improve results, as short headway combined with variable throttle is highly likely to occur in the synchronized flow phase. Using deep learning with a larger set of data is also recommended. A planned next phase will aim to bolster these early findings with a larger data set and implement the trained model for real-time predictions on running cars. The goal is to have a more balanced data set between trucks and passenger cars to gain more insight in the behaviour differences and how they correlate to the traffic phases.

References

[1] Condurat, M., Nicuţă, A.M., Andrei, R., 2017. Environmental Impact of Road Transport Traffic. A Case Study for County of Iaşi Road Network. Procedia Engineering 181, 123–130.
[2] Feng, F., Bao, S., Sayer, J.R., Flannagan, C., Manser, M., Wunderlich, R., 2017. Can vehicle longitudinal jerk be used to identify aggressive drivers? An examination using naturalistic driving data. Accident Analysis & Prevention 104, 125–136.
[3] Ferreira, Júnior, J., Carvalho, E., Ferreira, B.V., de Souza, C., Suhara, Y., Pentland, A., Pessin, G., 10-Apr-2017. Driver behavior profiling: An investigation with different smartphone sensors and machine learning. PLOS ONE 12, e0174959.
[4] Fugiglando, U., Massaro, E., Santi, P., Milardo, S., Abida, K., Stahlmann, R., Netter, F., Ratti, C., 2017. Driving Behavior Analysis through CAN Bus Data in an Uncontrolled Environment. arXiv:1710.04133 [physics] arXiv:1710.04133.
[5] Ito, T., Kaneyasu, R., 2017. Predicting traffic congestion using driver behavior, in: Procedia Computer Science, pp. 1288–1297.
[6] Kerner, B.S., 2009. Definitions of The Three Traffic Phases, in: Introduction to Modern Traffic Flow Theory and Control: The Long Road to Three-Phase Traffic Theory. Springer-Verlag, Berlin Heidelberg, pp. 9–40.
[7] Li, G., Li, S.E., Cheng, B., Green, P., 2017. Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities. Transportation Research Part C: Emerging Technologies 74, 113–125.
[8] Ma, C., Dai, X., Zhu, J., Liu, N., Sun, H., Liu, M., 2017. DrivingSense: Dangerous Driving Behavior Identification Based on Smartphone Autocalibration. Mobile Information Systems 2017, 15.
[9] Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A., 2008. RUSBoost: Improving classification performance when training data is skewed, in: 2008 19th International Conference on Pattern Recognition, IEEE, Tampa, FL, USA. pp. 1–4.
[10] Teja, M.S.S., 2014. Driver Behavior Detection System with Inter-Vehicle Communication. International Journal of Engineering Research 3, 5.
[11] Tian, J., Jiang, R., Li, G., Treiber, M., Zhu, C., Jia, B., 2016. Improved 2D Intelligent Driver Model simulating synchronized flow and evolution concavity in traffic flow. arXiv:1603.00264 [nlin, physics:physics] arXiv:1603.00264.
[12] Treiber, M., Kesting, A., 2013. Traffic Flow Dynamics: Data, Models and Simulation. Springer-Verlag, Berlin Heidelberg.
[13] Treiber, M., Kesting, A., 2018. Microsimulation of Traffic Flow. http://www.traffic-simulation.de/.
[14] Wang, W., Xi, J., Chen, H., 2014. Modeling and recognizing driver behavior based on driving data: A survey. Mathematical Problems in Engineering 2014.