ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 155

2018 International Conference on Advanced Control, Automation and Artificial Intelligence (ACAAI 2018)

Driver Modeling Based on Vehicular Sensing Data

Zhuowen Wang, Fuqiang Liu, Xinhong Wang and Yuyan Du

School of Electronics and Information Engineering, Tongji University, 4800 Cao'an Highway, Jiading, Shanghai

***Abstract*—In the past few years, the automotive electronics and sensing technologies have developed rapidly. Today, the status of most of the sub-systems in a running vehicle can be accurately monitored. This process produces a huge amount of data. Extracting the potential value of such data, to for instance support developing advanced vehicle diagnosis and active safety applications, has attracted tremendous attentions in both academia and industry. Considering that the sensing data, if sampled with sufficiently high frequency, can accurately represent how a driver maneuvers a vehicle, this paper investigates using the vehicular sensing data to exploit drivers' behaviors in different traffic scenarios. We apply machine learning techniques to construct driving behavior models, and discuss their applications in driver identification.**

Keywords-driving behavior models; vehicular sensing data; machine learning; driver identification

I. INTRODUCTION

In recent years, autonomous vehicles and connected vehicles have attracted the attention both in academic and industrial fields. They all indicate that traffic will become more and more intelligent in the future, and people will travel more conveniently, safely and comfortably. The central concern of autonomous driving or connected vehicle is vehicle's movement mode. By analyzing the normal driver's operation, a safe and effective vehicle handling program can be found. Vehicles are a very sophisticated and complex system whose speed, trajectory and mode of operation are affected in many ways, such as driver's habits, vehicle performance, road conditions, and other environmental factors. In order to explore the behavior patterns of vehicles in traffic, it is very important to construct a driver behavior model.

Building driver model requires a great deal of vehicle driving data. Modern vehicles, in fact, are equipped with a large number of sensors which can be used to record the behaviors of the drivers at all times through high frequency data acquisition mechanism. Now the vehicle data acquisition method has been mainly divided into four kinds: the first method is directly reading the vehicle sensor messages through the vehicle's CAN port [1]. The data sent by the CAN port has the characteristics of high frequency and wide variety. Hoverer, it is necessary to obtain the message protocol of this vehicle model in advance and develop the corresponding data acquisition system, which makes the overall cost expensive. Installing external sensors (gyroscope, inertial sensor, GPS, Video surveillance etc.) is a complement to the above approach, especially in terms of the need of image, the surrounding object and geographic information [2] [3]. Through the mobile phone sensors to obtain data is also a popular way among researchers because the phone is easy to carry and easy to develop [4]. This approach has its own shortcomings at the same time, for instance, fewer data types and relatively low accuracy compared to the vehicle inner sensors. The driving simulator can also generate large amounts of data [5] [6]. It can even simulate dangerous scenes, which is beneficial to the driver's risk assessment, but the price of a simulator is very expensive and the data may differ from the real situation [7].

After obtaining the original data of the vehicle through the above way, the driver's behavior modeling analysis can be carried out. Driver behavior analysis is mainly divided into two categories: one is driving behavior classification, and the other is driver classification. Driving behavior classification is extracting features from speed, acceleration, steering wheel angle and other data, to identify the turn, follow, overtaking and other driving behaviors. These identified actions are mainly used for autopilot-related research, and by modeling the traffic environment and specific behavior, we can teach a vehicle as safe and normal as a human being. Many articles have studied this field. [8] uses dynamic a time warping method to identify left turn, right turn, U-turn and other four kinds of driving behaviors based on GPS data. Similarly, the turning behavior under different roads are identified based on hidden markov model in [9]. [10] [11] have studied the follow behavior of the vehicle. The different stages of the follow action were divided by using the improved C-means algorithm in [10], and unsupervised learning classification results were also described. [11] constructs Gazis–Herman–Rothery model to segments and clusters the data of lateral acceleration and heading angle. The clustering results show that the following behaviors of different types of vehicles are diverse. [12] [13] predict the driver's intention based on the classification of driving behavior. In [12], by gaining pedal, steering, foot movement information, braking intention was predicted using the Bayesian learning method, and the risk of different timing of the brakes was evaluated. Four types of behaviors, i.e., straight/left/right/stop, at the intersection are predicted in [13] through the hidden Markov models (HMM) and hybrid-state system(HSS).

Driver classification is also a hot topic in academia, researchers have been hoping to solve the problem of high traffic accident casualty rates, and a driver's poor driving behavior is the root cause of the accident. We are eager to make driver model to describe whether a driver is dangerous or not, skilled or unskilled, stable or unstable, which is usually used as a basis for Usage Based Insurance (UBI). The driver classification model can also provide a corresponding improvement for drivers with poor driving behavior. When the category is large enough, that is, when the type is equal to the number of drivers, the driver classification problem becomes a driver identification problem, it can construct a driver portrait for a separate driver, and can also be used to identify the driver's vehicle which is stolen. The vehicle simulator is used in [14] to evaluate the driver's skill in a specific scene, and the drivers are

divided into three categories: novice, general and expert through extracting DFT features from wheel angle data. References [15] [16], respectively, distinguish driver risk from the perspectives of Bayesian theory and vehicle dynamics theory. The driver's categories in above papers were pre-labeled based on questionnaires and videos. However, in many cases, it is difficult for researchers to obtain records of the driver's violation in recent years. In this case, it is necessary to use the unsupervised learning method to cluster the driver [17] [18]. In addition, references [19] and [20] study driver identification technology. [19] uses the SVM and multivariate logistic regression to perform feature extraction on velocity, heading angle, instantaneous fuel consumption data, and applied it to driver identification, In [20], a driver identification model is constructed from the perspective of the deep learning,

Driver modeling is the basic research of many other applications. Its core idea is to find features that really represent driver behavior and use these driver characteristics to predict and analyze a particular goal. For example, in the driver behavior classification problem, this prediction goal is to distinguish between different driving behaviors, and in the driver classification problem, this goal can be to determine the driver's skill level, risk and so on. Because of the limitations of data collection, some studies have only experimented with vehicle data with fewer data types [16] [17] [20]. In view of the fact that more data sets provide more information, this paper uses a range of commonly used sensor data and several different feature extraction methods are adopted to obtain driver behavior characteristics. Moreover, in many previous studies on driver modeling, data modeling was based on a complete trip, but are all the data in a complete trip valuable? Based on this consideration, several different driving scenes are proposed in this paper, and the performances of different scene combinations are discussed under the application of driver identification.

The rest of this paper is organized as follows: Section II introduces data sources and the method of data preprocessing. Section III discusses feature extraction and specific scenes. Section IV describes the modeling method. Our conclusion and future work are presented in Section V.

II. DATA PREPROCESSING

This paper is based on the SPMD (Safety Pilot Model Deployment) project [22], which provides a comprehensive data collected from nearly 3,000 vehicles equipped with V2V communications devices in real-world scenarios. The data set include basic safety information, vehicle location information, vehicle-driver interaction data, roadside communication unit data and other types of data. The SPMD experiment was carried out in Ann Arbor, Michigan, USA. Each participant vehicle was owned by a volunteer and did not need to perform a special driving operation or a special journey due to the experiment. It is intended to obtain data during routine natural driving. And it can be thought that each car was basically driven by the same driver, different vehicles that represented different drivers, so that data can be used to analyze the differences of drivers' behaviors.

In the data set used in this paper, we only select the dimensions that are relevant to the driver's operating and travel habits, including: GPS time, latitude, longitude, longitudinal velocity, longitudinal acceleration, steering wheel angle. We selected 64 drivers, each driver file contains from 100 to 300 different trips, and the total number of samples is about 80 million.

The raw data of a vehicle can not be used directly, and a series of preprocessing operations are needed before construct a model. The preprocessing of vehicle data mainly includes the following: completing the missing values, smoothing the noise data, deleting the outliers, resolving data inconsistencies, reducing the data dimensions and normalizing data.

III. SELECTION OF DRIVING PATTERNS

A. Driving Scenes

Many previous studies are based on the entire trip to construct variables and obtain the driver's features, but is it reasonable enough for extracting drivers' driving patterns? For example, in the case of traffic jams, the statistics characteristics of the complete trip data will show the driver brakes too frequently and speed is not stable, makes him look like a low-skilled driver. It indicates that the evaluation is unfair and can not represent real characteristics of drivers. Therefore, a whole trip must be divided into several parts to find scenes that really reflect drivers' driving habits [23]. This paper defines 6 scenes, which are:

(1) START: 5 seconds after starting

(2) STOP: 5 seconds before stop

(3) HIGH SPEED: the driving speed is faster than 50 kilometers per hour

(4) TURN: cars turn at corners or intersections

(5) ACCELERATION: continuous acceleration for at least 5 seconds

(6) DECELERATION: continuous deceleration for at least 5 seconds

B. Feature Extraction

The above six scenes are typical and familiar scenes of a journey. By extracting feature values from the scenes, we can distinguish between different drivers. A common way of obtaining features is to extract statistical values of data, which we generalize as follows:

- Mean, maximum and variance of speed.
- Mean, maximum and variance of longitudinal acceleration.
- Mean, maximum and variance of steering angle.
- Positive kinetic energy (PKE) [24] [25] that is defined as the sum of the differences between the squares of the final and initial speeds in successive acceleration manoeuvers, divided by total trip:

$$\mathrm{PKE} = \frac{\sum_I (V_{I+1}^2 - V_I^2)}{D}, v_{i+1} > v_i \quad (1)$$

- Relative positive acceleration (RPA) that is defined as the

product of the instantaneous speed and the instantaneous positive acceleration divided by total trip distance:

$$\mathrm{RPA} = \frac{\sum_I (v_i * a_i)}{D} \quad (2)$$

• Root mean square (RMS) of the power factor (PF) over the different time:

$$\mathrm{RMS(PF)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} PF_I^2} \quad (3)$$

$$\mathrm{PF} = 2 * \mathrm{v} * \mathrm{a} \quad (4)$$

• Probability density value [20] of speed, acceleration and steering angle.

• Road conditions, including highway and urban roads.

• Driving time of a trip.

• Speed, acceleration, and steering angle are continuous numerical variables, and they all have their own range of distribution. Here, take the acceleration scene for example, Figure I and II show the acceleration distributions of the driver 10 and the driver 11 in different acceleration scenes respectively. It can be seen that the same driver has similar probability density distribution at different acceleration scenes, and the distribution characteristics of different drivers are diverse. Since this distribution is not a typical distribution, the parameters of the common distribution (such as the mean value and variance value of the Gauss distribution) cannot be directly used as features. At this point, we can divide the distribution area into several equal bins and compute the probability density of each bin, which is used as the characteristic value of the driver. Moreover, the geographical position and driving time, which can also affect the driver's behavior, are taken in consideration.

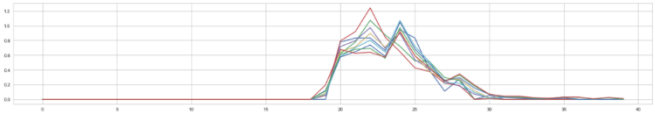

FIGURE I. DISTRIBUTION OF ACCELERATION DATA IN ACCELERATION SCENE: (DRIVER 10)

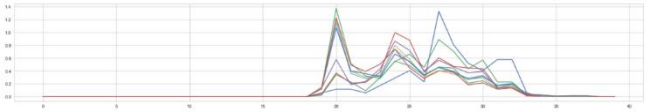

FIGURE II. DISTRIBUTION OF ACCELERATION DATA IN ACCELERATION SCENE (DRIVER 11)

IV. DRIVER MODELING

Due to the lack of driver's background information (such as the number of violations, the number of accidents etc.), it's temporarily unable to give each driver a label of risk degree. So this article will discuss performance under the application of driver identification using above modeling method. The performance of a model will be influenced by the input variables, the model algorithm and the model parameters. In the following, we will discuss scene selection, and model selection to get the optimal model. The model parameters had been adjusted when using the model.

A. *Evaluation Method*

Driver identification is a typical classification problem and the acquired features need to be put into the model of machine learning to predict the categories they belong to. The prediction accuracy of a model is obtained through dividing all correctly identified samples by the total number of samples.

The five-fold cross validation will be used to determine the model's real prediction accuracy, and the given modeling sample will be taken out of 4/5 of the samples to train the model, and the remaining samples will be predicted with the newly established model. After repeated several times, the average accuracy is recorded as the evaluation index of the model, which is close to the actual test accuracy. When the average accuracy is high, we can say that the model performs well.

B. *Scene Selection*

In section III, we propose six kinds of driving scenes used to characterize the driver's driving behavior. Each scene can extract features that are described in section III (B). And a vector, contains all the features of a scene, can be treated as a sample point. One of the dimensions of this vector is the sign of the scene. The accuracy of the model can be obtained by putting the sample points into the model.

Firstly, only the individual scenes are used to identify two drivers. The random forest model was used as the classification model. Random forest model, a classic machine learning model, has advantages of less overfitting and easier processing of high-dimensional data. As shown in Figure. III the horizontal axis represents the number of the characters, and the vertical axis represents the model prediction accuracy. It can be seen that the acceleration, deceleration and high-speed scene have the highest identification rate, the turn scene is in the middle position, and the accuracies of the starting and stopping scenes are the worst. However, unlike the expected result, the performance of the individual scenes is not as good as the performance of the entire trip. The reason may be that although the complete trip has some redundant data, it still contains most of the driving scene information. So if the performance of a single scene model is not good enough, can the integration of multiple scenes improve the overall recognition rate? That means putting samples of multiple scenes into the classification model.

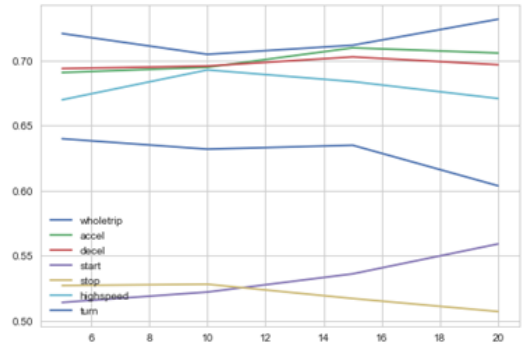

FIGURE III. PREDICTION WITH SINGLE SCENE

Figure IV tried combinations of three kinds of the scenes to construct models (only partial combinations are shown here).

Compared with single scene, the highest identification rate of combinations has been significantly improved, and the highest combination of the accuracy rate can reach 0.76. It can be seen that the performances of most combinations exceed the performance of full trip. In order to find the best combination of scenes, all the situations have been exhausted. Finally, we find that the combination of start, acceleration and deceleration, can achieve the best performance.

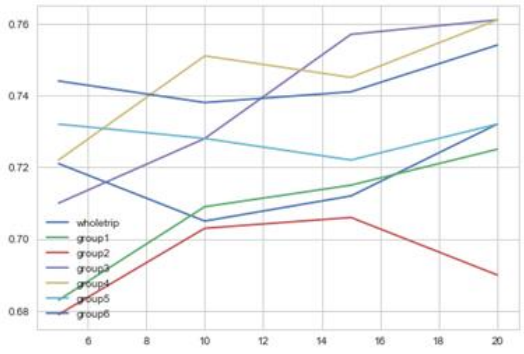

FIGURE IV. PREDICTION WITH COMBINATIONS OF THREE SCENES

C. Model Selection

The recognition rate of machine learning model is related to the matching degree between internal algorithm and data set. It is desirable to find an optimal model to process vehicle data. In addition to the random forest model, three models are used in this section, which are k-nearest neighbors (KNN), artificial neural network (ANN) and support vector machine (SVM), respectively. The KNN algorithm is a model based on the proximity point. In the category decision, the algorithm is only related to the adjacent samples. Therefore, the KNN method is more suitable than the other methods for the crossover or overlapping sample sets. The SVM is a very commonly used supervised learning algorithm that divides different types of data sets from high-dimensional spaces by calculating category distances and setting kernel functions. The ANN algorithm simulates the pattern of human biological neurotransmission signals by constantly injecting new sample points to stimulate neurons and update the weights, it can handle very complex nonlinear problems when setting multilayer neurons.

Figure. V shows the predicted probability of different models under different number of features. It can be seen that with the growth of numbers, the recognition probability has been greatly improved, which is due to the distribution in a more detailed division provides more information. It is worth noting that this value is not as bigger as better, when the number of types closes to the number of sample points will cause the curse of high-dimensional, the model over-fitting will result in a decline in the probability of prediction. Among the four models, random forest and ANN have better performance than others, especially when the number of features is large, the performance of ANN is superior to random forest, and ANN's highest recognition rate reaches 0.83. SVC also does well when the number is small. However, its performance is not significantly improved with the increase of the number of features. The overall recognition rate of KNN is poor. The above classification are two classification problems. When more driver are added to the model, the recognition rate will have a certain degree of decline, But the results are still within the acceptable range, as shown in Figure VI.

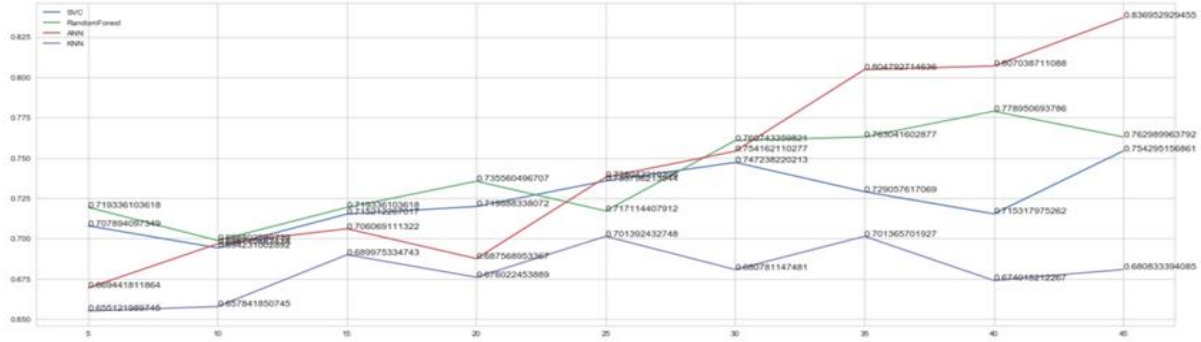

FIGURE V. THE IDENTIFICATION RATE OF DIFFERENT MODELS UNDER DIFFERENT NUMBER OF FEATURES

V. CONCLUSION

In this paper, we discuss the driver modeling process, including data acquisition, preprocessing, feature extraction, and driver modeling. The acceleration, speed, steering angle, GPS time, location information are used to construct the driver model, six kinds of typical driving scenes are proposed in this paper, including the start, stop, high speed, turn, acceleration, deceleration. All of the scenes were modeled using the random forest algorithm, the results showed that the combination of start, acceleration, deceleration scenes could be able to carry out the best identification of drivers. Otherwise, four kinds of machine learning models (random forest, k-nearest neighbors, artificial neural network and support vector machine) are compared. It is found that both artificial neural network models and random forest models are superior in performance, in which the maximum recognition probability of artificial neural network model is up to 0.83. After adding categories, that is, the number of drivers, the model's recognition probability has declined, but is still within the acceptable range.

Many future works can be done. In fact, in most cases, we are not concerned about the accurate recognition of each individual

driver, but focus on the drivers' categories. Is a driver aggressive or non-aggressive, and what is the range of accidents he will have in the next three years? Is he an ordinary office worker or a taxi driver? Because of the separation of the drivers' background information and the driving data, there is still a lack of a good way to catch a driver with an accurate classification label only through driving data. The driver modeling method of this paper can be used as the basis of future driver classification.

ACKNOWLEDGMENT

This work was funded in part by the Fundamental Research Funds for the Central Universities (0400219331).

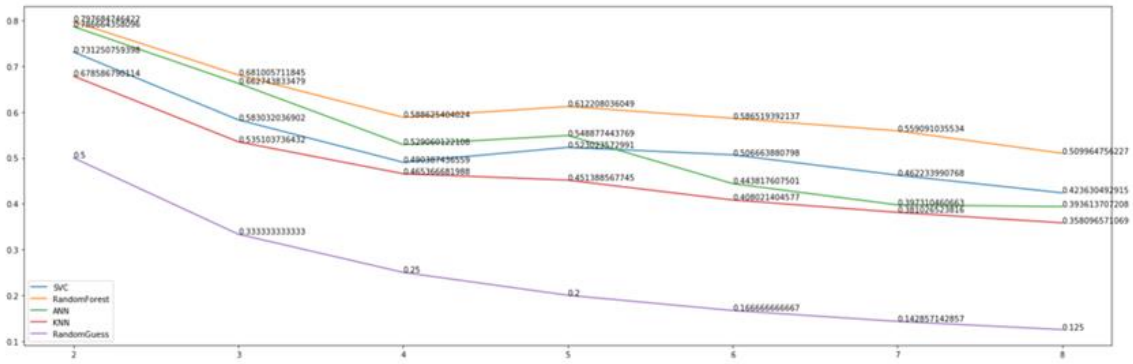

FIGURE VI. THE IDENTIFICATION RATE OF DIFFERENT MODELS WITH MULTIPLE DRIVERS (40 FEATURES)

REFERENCES

[1] Jang W, Jong D, Lee D. "Methodology to improve driving habits by optimizing the in-vehicle data extracted from OBDII using genetic algorithm." *International Conference on Big Data and Smart Computing*. IEEE, 2016:313-316.

[2] Toledo T, Musicant O, Lotan T. "In-vehicle data recorders for monitoring and feedback on drivers' behavior." *Transportation Research Part C: emerging Technologies*, 2008, 16(3): 320-331.

[3] Noble A M, Dingus T A, Doerzaph Z R. "Influence of in-vehicle adaptive stop display on driving behavior and safety." *IEEE transactions on intelligent transportation systems*, 2016, 17(10): 2767-2776.

[4] Paefgen J, Kehr F, Zhai Y, et al. "Driving behavior analysis with smartphones: insights from a controlled field study." *Proceedings of the 11th International Conference on mobile and ubiquitous multimedia*. ACM, 2012: 36.

[5] Gregoriades A, Florides C, Lesta V P, et al. "Driver behaviour analysis through simulation." *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE*, 2013: 3681-3686.

[6] Freeman P, Rodriguez J, Wagner J, et al. *"Validation of a fixed-base automotive simulator for run-off-road safety and recovery training."* Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2015, 229(5): 574-589.

[7] Hong J H, Margines B, Dey A K. "A smartphone-based sensing platform to model aggressive driving behaviors." *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014: 4047-4056.

[8] Freeman P, Rodriguez J, Wagner J, et al. "Validation of a fixed-base automotive simulator for run-off-road safety and recovery training." *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 2015, 229(5): 574-589.

[9] Johnson D A, Trivedi M M. "Driving style recognition using a smartphone as a sensor platform." *International IEEE Conference on Intelligent Transportation Systems*. IEEE, 2011:1609-1615.

[10] Mitrovic D. "Reliable method for driving events recognition." *IEEE transactions on intelligent transportation systems*, 2005, 6(2): 198-205.

[11] Ma X, Andreasson I. "Behavior measurement, analysis, and regime classification in car following." *IEEE Transactions on Intelligent Transportation Systems*, 2007, 8(1): 144-156.

[12] Higgs B, Abbas M. "Segmentation and clustering of car-following behavior: Recognition of driving patterns." *IEEE Transactions on Intelligent Transportation Systems*, 2015, 16(1): 81-90.

[13] Hong J H, Margines B, Dey A K. "A smartphone-based sensing platform to model aggressive driving behaviors." *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014: 4047-4056.

[14] Carmona J, García F, Martín D, et al. "Data fusion for driver behaviour analysis." *Sensors*, 2015, 15(10): 25968-25991.

[15] Zhang Y, Lin W C, Chin Y K S. "A pattern-recognition approach for driving skill characterization." *IEEE Transactions on Intelligent Transportation Systems*, 2010, 11(4): 905-916.

[16] Wang W, Xi J, Li X. "Statistical pattern recognition for driving styles based on bayesian probability and kernel density estimation". *International Journal of Automotive Technology*, 2016.

[17] Filev D, Lu J, Prakah-Asante K, et al. "Real-time driving behavior identification based on driver-in-the-loop vehicle dynamics and control." IEEE International Conference on Systems, Man and Cybernetics. IEEE Xplore, 2009:2020-2025.

[18] Malta L, Miyajima C, Takeda K. "A study of driver behavior under potential threats in vehicle traffic." *IEEE Transactions on Intelligent Transportation Systems*, 2009, 10(2): 201-210.

[19] Toledo T, Musicant O, Lotan T. "In-vehicle data recorders for monitoring and feedback on drivers' behavior." *Transportation Research Part C: Emerging Technologies*, 2008, 16(3): 320-331.

[20] Quek Z F, Ng E. "Driver identification by driving style." Technical report in CS 229 Project, Stanford university, 2013.

[21] Dong W, Li J, Yao R, et al. "Characterizing driving styles with deep learning." arXiv preprint arXiv:1607.03611, 2016.

[22] Jeon S I, Jo S T, Lee J M. "MultiMode Driving Control of a Parallel Hybrid Electric Vehicle Using Driving Pattern Recognition". *Journal of Dynamic Systems Measurement & Control*, 2002, 124(1):489-494.

[23] U.S. Department of Transportation, Federal Highway Administration. Safety Pilot Model Deployment One Day Sample Data Handbook [EB/OL].https://www.its-rde.net/index.php, October.2016.

[24] Younes Z, Boudet L, Suard F, et al. "Analysis of the main factors influencing the energy consumption of electric vehicles." *Electric Machines & Drives Conference (IEMDC), 2013 IEEE International*. IEEE, 2013: 247-253.

[25] Jeon S I, Jo S T, Lee J M. "MultiMode Driving Control of a Parallel Hybrid Electric Vehicle Using Driving Pattern Recognition." *Journal of dynamic Systems Measurement & Control*, 2002, 124(1):489-494.