Accident Analysis and Prevention 139 (2020) 105460

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

A driver behavior assessment and recommendation system for connected vehicles to produce safer driving environments through a "follow the leader" approach

Zihan Hong[a], Ying Chen[a,*], Yang Wu[b]

[a] Department of Civil and Environmental Engineering, Northwestern University, 211 Chambers Hall, 600 Foster Street, Evanston, IL 60208, United States
[b] Department of Electrical Engineering & Computer Science, Northwestern University, Technological Institute, 2145 Sheridan Road, Evanston, IL 60208, United States

ARTICLE INFO

Keywords:
Driver behavior
Vehicle trajectories
Connected vehicles
Gaussian mixture model
Data mining
Recommendation system
Assessment system

ABSTRACT

As part of the emerging world of intelligent transportation, there is considerable interest in developing connected vehicles that are more capable of identifying and guiding individual drivers' behavior than collecting mileage as a moving cart. The two goals of this study are (a) to build a conceptual framework for driver assessment and (b) develop recommendation systems to evaluate individual driving performance and guide driver behaviors, thus improving the network traffic conditions and individuals' perceived safety. A safety score is defined relatively by comparing a driver's individual pattern to a standard "safe driver" pattern. To elaborate, the proposed system adopts advanced data mining techniques to extract, identify, characterize, and display driving behavior patterns. The scoring system provides a basis of assessing individual drivers, who are then recommended to mimic a nearby "safe" driver in a connected environment. To evaluate and implement the proposed conceptual framework, an anonymous trajectory dataset collected from Pittsburgh urban area is applied to build the scoring system, which is then integrated within a virtually simulated environment. The results show that the proposed behavior assessment and recommendation system framework improves the overall performance of a connected traffic system beyond those attained through baseline connectivity principles.

1. Introduction

With the rapid development of information and communication technologies, vehicles equipped with advanced wireless connectivity capabilities are becoming more prevalent in the automotive market. The increase in the number and performance of sensors has allowed for well-designed intra-vehicle communication systems to report any abnormal occurrences to the electrical control units (ECU) in a reliable and timely manner (Ahmed et al., 2007). Accordingly, the vehicle is no longer merely collecting mileage, but also a vast array of information regarding driver behavior and the surrounding environment to improve the efficiency of the overall transportation network. The potential popularity of connected (and autonomous) vehicles encourage researchers and manufacturers to develop conceptual and practical applications to improve drivers' behavior and safety to enhance the overall driving environment, with the focus on detecting and reporting 'abnormal' behaviors.

However, developing solutions for improving driver behavior, let alone defining what such improvements could or should be, remain elusive in research and practice. Moreover, one of the broad challenges is analyzing enormous volumes of data collected from intra- and inter-vehicle communication systems. Rapid advancements in data mining and machine learning algorithms, though, have permitted researchers to more effectively explore these seemingly untamable datasets to discover important underlying patterns and relationships among variables.

The objective of this research is thus to conceptualize a system for driver behavior assessment and recommendation using data analytics and statistical methods, which quantifies drivers' safety levels and provides behavior guidance to reduce risks. To be more specific, it recommends aggressive drivers to follow identified "safer drivers" using connected vehicle tools. Some studies have examined the potential relationship between aggressive behavior and roadway characteristics like the number of lanes, speed limit (Administration, 2016), while other studies modeled the driver's speed profiles to evaluate the aggressiveness of the drivers (Spiegel et al., 2011). Driver behavior modeling and identification has also been explored by researchers using pedal operation data to generate distance-keeping and car-following patterns (Igarashi et al., 2004; Ozawa et al., 2005; Wakita et al., 2005;

* Corresponding author.
E-mail addresses: z.hong@u.northwestern.edu (Z. Hong), y-chen@northwestern.edu (Y. Chen), yangwu2015@u.northwestern.edu (Y. Wu).

https://doi.org/10.1016/j.aap.2020.105460
Received 13 May 2019; Received in revised form 6 November 2019; Accepted 27 January 2020
Available online 02 March 2020
0001-4575/

Miyajima et al., 2006, 2007). However, to the best knowledge of the authors, there is minimal research employing detailed trajectory data to examine aggressive behavior in a connected driving environment. Thus, the main contribution of this paper is to fill this gap in the literature by proposing a framework that relies on rigorous analysis of this data in an effort to improve driver performance and thus the efficiency of the transportation network. Furthermore, a recommendation scheme for driver guidance is put forward based on the results of speed harmonization techniques. It is worth noting that the focus of this paper is not to design a manufacture-level system which is ready to be deployed on vehicles. Therefore, the feasibility of connection, the hardware design of sensors and monitors, and the privacy issue in connected vehicles are beyond the scope of this paper.

The goal of the proposed system is to improve the quality of driving behavior and reduce the risk of accidents. To do so, a machine learning method is first presented to model the algorithmic signature for each driver and identify the underlying driving behavior model. Does a driver usually take a long trip, or a short trip? How about highway trips and back roads? Does he/she accelerate hard from stops or take turns at high speed? The answers to these and other questions form an aggregate and maximally distinct driver profile. Second, by measuring the similarity between and individual driving profile and compared to a "safest driver" behavioral standard, a safety score for each driver is generated. Finally, a simulated connected vehicle environment is used to evaluate and provide driving behavior guidance. In a connected system, the relatively risky drivers are suggested to follow safer ones by connecting to a neighboring safe driver as a leading vehicle. The learning process trains the aggressive and highly deviated drivers by reforming their behavior pattern closer to a normalized pattern and therefore increases the safety of the entire traffic system.

The remainder of this paper is organized as follows. Section 2 synthesizes the existing literature on connected vehicles, driving behaviors studies, and driver risk assessment systems in the United States. Section 3 describes the overall framework for the proposed assessment and recommendation system. Section 4 highlights the methodology to construct an assessment system of driving behaviors, including machine learning models in conjunction with a holistic driving behavior library. Section 5 illustrates the experiment of the recommendation system in a simulated environment using connected vehicle techniques and evaluates the performance of the entire system. Section 6 then provides concluding remarks and discusses challenges for future work and implementation.

2. Literature review

There are three core parts in this study: (1) connected vehicles, which is the foundation of the proposed system that vehicles can be guided in a connected environment to improve the network safety and performance, (2) driver behavior modeling, which supports data analytics to identify the risk drivers, and (3) driver assessment system, which quantifies driver behaviors and then is incorporated in the recommendation system to guide vehicles.

2.1. Connected vehicles

The general perspective of connected vehicles (CV) within intelligent transport systems (ITS) is the seamless integration of infrastructure, vehicles and users. CV are allowed to interact within not only internal environment, i.e., the interactions of vehicle-to-sensor on board (V2S), but also external environment, i.e., vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2R), and vehicle-to-Internet (V2I), vehicle to pedestrian (V2P) and even vehicle to anything (V2X) (Shladover, 2018).

Applications enabled by V2R communications include speed harmonization (Talebpour et al., 2013; Khondaker and Kattan, 2015), intelligent traffic signals (Feng et al., 2015; Tiaprasert et al., 2015), real-time traveler information (Khan et al., 2017), and incident management (Narla, 2013). Due to their well-informed and cooperative characteristics, connected vehicle technologies can be characterized as: (1) Individual vehicle applications, provided by automakers or apps, such as BMW Connected NA (FHA, 2010) and the European eCall (which will be mandatory in all new cars sold within European Union after April 2018 (Anon)); and (2) Beyond individual vehicle applications, such as safety (e.g. collision detection lane change warning, and cooperative merging), as well as smart and "green" transportation (e.g. traffic signal control, intelligent traffic scheduling) (Lu et al., 2014).

Recent applications have demonstrated data fusion capabilities from multiple vehicles, external data sources, and a built-in data library which collects and stores historical driving data from drivers (Li et al., 2014). That is, researchers have begun exploring more sophisticated mechanisms for extracting and learning from large-scale data to adjust and suggest improvements to vehicle operations (Siegel et al., 2017). Furthermore, cloud computation services and storage allow applications to be more scalable and accessible with faster computational performance for more users (Siegel, 2013; Guerrero-Ibanez et al., 2015; Wilhelm et al., 2015).

2.2. Driver behavior modeling

Pedal operation data is the most common source for identifying individual patterns in driving behavior models. The existing literature is comprised of numerous investigations of distance-keeping and car-following patterns. For example, Igarashi et al. (2004) adopted Gaussian Mixture Models (GMM) to investigate the uniqueness of individual driving behaviors in acceleration and deceleration. Their objective is to achieve safer driving behavior in emergency scenarios. Ozawa et al. (2005) also adopted GMM driver model, but with more information, including velocity, car-following distance, and gas or brake pedal operations to investigate the driving behavior for accident prevention. Wakita et al. (2005) proposed a method to identify driving behavior using car following observations. They also applied a driving simulator to generate and measure driving behavior signals (i.e. features) to describe the vehaviors. They compared parametric models and non-parametric models, concluding that the latter acheives better performance. Wakita et al. (2006) conducted spectral analysis of driving behavior signals, including pedal operation for acceleration and deceleartion, an applied a GMM to achieve better accuracy in driver identification both in similation and practice. Furthermore, they analyzed car-following behaviors (Miyajima et al., 2007), where the GMM models were applied to describe the relationship between following distance and velocity along with pedal operational patterns (i.e. acceleration and deceleration). They did a field test with 276 drivers and achieved an accuracy rate of 76.8 %.

2.3. Driver risk assessment system

Driver risk assessment systems have been explored and designed within the automobile insurance field to formulate "reasonable" premiums for drivers. The data applied fed into these systems typically include individual credit score, employment history, driving record, insurance history, and accident history (Daniel, 2012). Additionally, some companies provide each vehicle a portable device with a unique key such that individual activities of the driver can be collected from the device, sent to the data center, and assessed to update premiums.

Accordingly, in this comprehensive system one component is used in collecting information from the driver, another component serves as an on-board diagnostic system, while a third functions to provide driver assistance technology. The assessment scores could be provided based on the information from these different components (Parameshwaran, 2016).

An alternative approach to driver risk assessment systems is to report and score onboard events in real-time (Cook and Gilles, 2013). The

system is designed to evolve via repeated calibration from the stream of event reports and resulting data analysis (Cook and Gilles, 2016). Moreover, it could be continuously improved by optimizing the data transmission and detecting driving events more effectively. The on-board score is compared with historical values which are stored in the system (Cook et al., 2014). The driver score can also be normalized according to key environmental factors (Cook and Etcheson, 2014).

Existing driver risk assessment systems from vehicle manufacturers as well as from insurance companies are designed to focus more on individual performance. However, from the perspective of traffic network and safety, it is worth noting that the neighboring vehicles and environmental conditions should be integrated in the system, made possible byconnected vehicle techniques. It has also been shown that GMM-type models can perform well in driver identification and signature, which enables the modeling and data analaytics to be independent from the traditionally preinstalled sensor and device for each driver. Thus, this paper proposes a novel driver identification, assessment and guidance system to take advantage of the superior predictive prowess of machine learning algorithms, and modern connected vehicle techniques, to achieve a more comprehensive assessment of driver performance with the goal of improving traffic network safety.

3. Framework

The proposed integrated assessment and recommendation system (Fig. 1) includes three components: measurement, analysis, and feedback. The details of the data adopted in this study is explained further in Section 4.1.

In the analysis component, pattern recognition algorithms are applied to raw data to extract meaningful information by identifying nine behavioral features, such as lane changes and turns with or without acceleration (Toledo et al., 2008), harsh brakes, harsh cornering and excessive acceleration. The features are selected and extracted to describe the safety score of a driver trip as follows:

- Feature 1: trip distance.
- Feature 2: trip duration.
- Feature 3: average speed.
- Feature 4: maximum speed.
- Feature 5: standard deviation of speed.
- Feature 6: number of hard brake (acceleration < -7 mph/s).
- Feature 7: number of hard speed-up (acceleration > 7 mph/s).
- Feature 8: number of high-speed turning at local road (speed direction turn > 40°, speed value in [20 mph, 50 mph]).
- Feature 9: number of lane changing or curvilinear movements incidents (a movement: moving direction turn in [5°, 22.5°], speed > 30 mph) per unit time (Salvucci et al., 2007; Deng, 2013).

A Gaussian Mixture Model-Universal Background Model (GMM-UBM) system is then adopted: to expound, the Universal Model is built to detect the general environmental (or background) factors from the information of all drivers in the dataset, and simultaneously individual driving model (i.e. Gaussian Mixture Model) for each driver is constructed to characterize individual behavior profiles. These models provide two possible ways of assessing a driver with different reference points: if the historical incidental record or related insurance pricing is available, the safest driver with best historical record can be defined as the reference; otherwise, the driver with the most normalized behavior among the profiles' distribution is taken as the standard. In this study, the second approach is applied for the following three reasons: (1) in

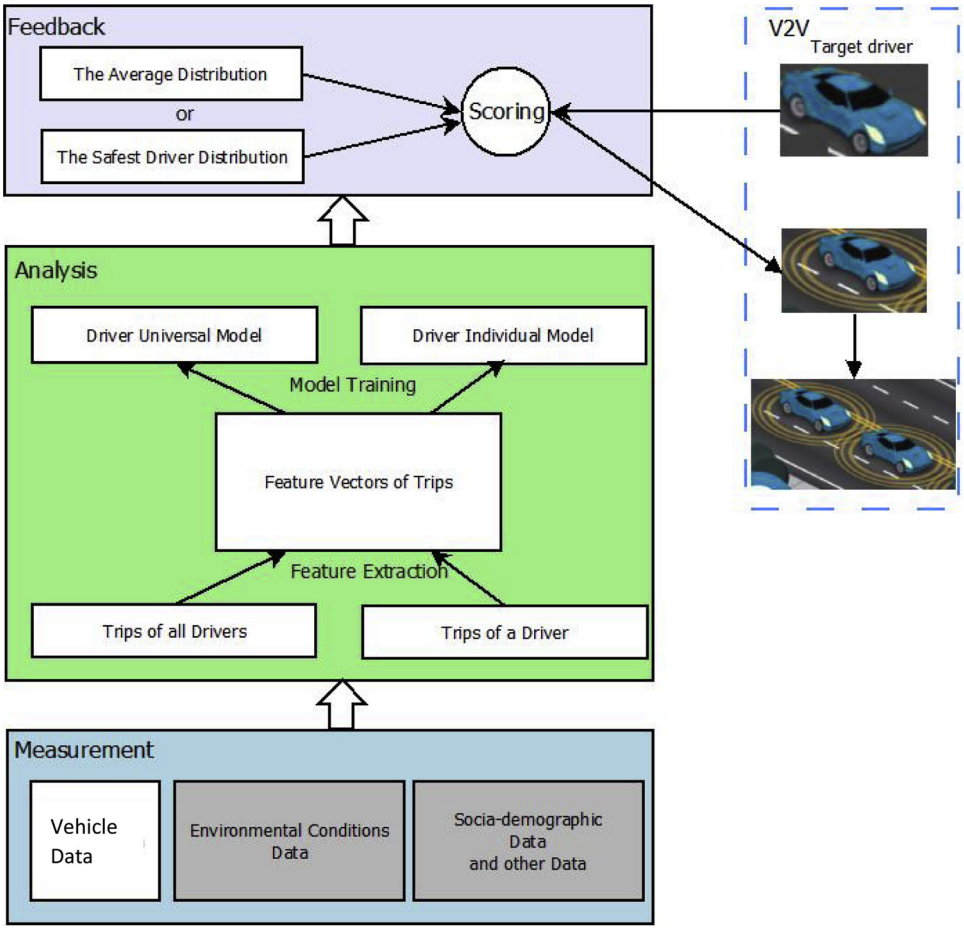

Fig. 1. Overall framework of driver assessment and recommendation system.

most cases, the former one is not applicable due to the lack of data and privacy concerns; (2) the safest driver may not be a good leading driver in every scenario — as a driver with best historical record might be someone always driving at low speed to avoid accidents in a local area, but he/she may induce more lane changing as a result; and (3) the historical record cannot show any real-time features governed by road pavement and geometry. Following this definition, any deviated behaviors can be regarded as outliers in the profile distribution.

In the feedback component, when a target driver's trip information is ready, a maximum likelihood detection is adopted to match the target driver with its individual model, which is followed by the calculation of a similarity metric between the target driver's distribution and the "standard" distribution, given as a percentage. The scoring system is assumed to have been installed and enabled within the target vehicle. Once the measurement and analysis components are performed, the lower-scored driver will receive warnings when engaging in aggressive and unsafe driving behavior followed by recommended guidance from a better (safer) leading driver, which is obtained from the currently connected nearby drivers or a built-in driving behavior library. The built-in library is assumed to be available in the connected and/or autonomous vehicles environment, where the driving patterns from similar historical driving circumstances are stored as reference. Circumstances are identified through connected vehicle techniques through communication among vehicles, infrastructure and internet to obtain a complete profile of environment and traffic network conditions.

4. Methodology

This section describes the method to construct a driver risk assessment system using machine learning and data mining techniques with an anonymous dataset. First, the anonymous trajectory dataset collected from an urban area are described; this dataset used in testing the driver scoring system and building the behavior library is obtained from a Kaggle project (https://www.kaggle.com/c/axa-driver-telematics-analysis), which is provided by the AXA Insurance Company. Second, the machine learning models are presented to identify and construct drivers' behavior profiles. Third, the driving behavior library is constructed to assess relative driver performance.

4.1. Data description

Current technologies allow the vehicle to "explore its surroundings" and gather data on the roadway, which could be used to provide technical support in the backend. Specifically, engine data and accelerometer data are combined with GPS tracking data to enhance the overall picture of vehicle operations. Newer generations of vehicles have more sensors which constitute a sophisticated data collection system, which includes transmissions, brakes, safety systems, and electrical systems. Collectively, these provide rich data for researchers to study the driver's behavior and help reduce the possibility of accidents.

The dataset used in this paper includes nearly 50,000 trips from 2736 anonymous drivers, and each driver has 200 trips. For each trip, the position of the car in meters is recorded every second. To protect the privacy of the drivers' positions, the original trips have been randomly relocated with origin centered at 0.0, 0.0. The data of a sample trip is showed in Table 1; Fig. 3 shows some sample trip plots of a driver in this dataset. Note that the unit for x and y coordination is meters.

The majority of 200 trips belong to the driver, but a small and random number of fake trips were also added. These fake trips were collected from other drivers who were not included in this dataset to avoid similarity with the given drivers. It is reasonable to have "fake" trips for each vehicle, since the same vehicle can be driven by any family member in a household or even by friend. In most cases of insurance companies, a vehicle is tagged with an identifier, and in this study, this identifier is treated as an identified driver. For this reason, the fake trips are not to be excluded to build the behavior library. Rather, the fake trips can help evaluate the accuracy of the proposed algorithmic signature method — the details of fake trips for each driver, i.e. trip id and the number of trips, are unknown to users. However, Kaggle website provided an evaluation and cross validation interface, which allows the participants to upload the results, and accuracy of identification is sent back when results are uploaded for evaluation.

4.2. Machine learning models

4.2.1. Driver signature

The driver signature is extracted from a driving behavior library using an individual driving behavior model. Identifying the general driving behavior of a specific target driver helps minimize the random noises generated within the trip data and emphasizes habitual behaviors for a more generalized assessment.

The maximum likelihood method and GMM-UBM system are adopted in finding the best-match of this target driver among all possible drivers. The GMM-UBM system (Reynolds et al., 2000) is commonly utilized in the Speaker Verification field. It uses a Gaussian Mixture model, a type of weighted probabilistic model, to profile a person. This model is selected due to the parallels between driver identification and speaker verification. For example, if face verification is needed, the face of a person does not change too much in different headshots; however, a speaker's voice varies a lot and is dependent on both individual pattern (mood, tone, etc.) and the context. Similarly, a driver can exhibit a variety of behaviors depending on road conditions, personal mood and weather conditions. Given the proposed analogy between driver verification and speaker verification, this method is adopted for the current study. The general framework of GMM-UBM system is presented in Fig. 2.

In the GMM-UBM system, UBM is trained using all driver trajectories to represent the universal driver behavior features (regarded as general or background factors). Given data to train a UBM, the iterative expectation-maximization (EM) algorithm (Dempster et al., 1977) is used to obtain the final model.

On the other hand, each GMM is trained to represent an individual driver's behavioral features. Given a feature vector x of a target driver, the maximum likelihood is calculated and selected in Eq. (1) to find the best-match GMM for the target driver behavior from the behavior library. This GMM will represent the target driver in the scoring procedure.

$$argmax_i(L(x))$$

$$L(x) = \log\left(\frac{UBM(x)}{GMM_i(x)}\right),\ i = 1, \cdots n \tag{1}$$

4.2.2. Drive risk assessment

Given a matched GMM Y of the target driver, and a GMM S of a "standard" driver (the average of all drivers or the safest driver), the "standard" driver is scored 100. The target driver is then assessed using a similarity measurement between the distribution of Y and S, which is calculated by determining the overlap of the distributions corresponding to the two GMMs in Eq. (2).

The entire process can be stated as a basic hypothesis test between

H0. Y is regarded as a safe driver similar with S, and

H1. Y is not regarded as a safe driver similar with S.

The optimal test to reject or accept the null hypothesis is a likelihood ratio test given by

$$\lambda = \frac{p(Y|H_0)}{p(Y|H_0)} \begin{cases} if\ \lambda \geq \theta & accept \quad H_0 \\ if\ \lambda < \theta & reject \quad H_0 \end{cases} \tag{2}$$

Table 1
One Trip Data of Driver (id = 11).

Sequence No.	x	y	Sequence No.	x	y	Sequence No.	x	y
1	0	0	20	145	191.6			
2	5.5	9.1	21	154.5	201.2	239	597.4	1465.7
3	12	17.5	22	164	212	240	597.4	1465.7
4	18.4	27.1	23	173.9	223.2	241	597.4	1465.7
5	25.1	37.2	24	183.3	235.5	242	597.3	1465.6
6	32.8	46.5	25	193.4	246.9	243	597.3	1465.6
7	40.3	57.2	26	202.9	259.2	244	597.3	1465.6
8	48.1	68.1	27	212.5	271.5	245	597.3	1465.6
9	56.8	78.3	28	221.4	285	246	597.3	1465.6
10	64.6	89.3	29	231.5	297.8	247	597.3	1465.6
11	72.3	100.1	30	240.9	311.6	248	597.3	1465.6
12	80.7	110.1	31	250.6	325.6	249	597.3	1465.6
13	88.4	120.9	32	259.7	340.6	250	597.3	1465.6
14	96.1	131.8	33	269.7	354.9	251	597.3	1465.6
15	104.4	141.6	34	279	370	252	597.3	1465.6
16	111.9	152.3	35	287.4	385.9	253	597.3	1465.6
17	120.3	162.3	36	295.6	401.5	254	597.3	1465.6
18	128.4	172	37	303.2	416.8	255	597.4	1465.7
19	137.2	180.9	38	309.8	432.6	256	597.4	1465.7

$$p(YH_0) = \int_{-inf}^{inf} \min(GMM_{target}(x),\ GMM_S(x))dx \tag{3}$$

where $p(Y|H_0)$ is the cumulative distribution function (Eq. (3)) to evaluate the similarity of the driving behavior model between GMM_{target} of the target driver and GMM_s of the standard driver.

4.3. Building a driving behavior library

4.3.1. Feature generation

The behavior of drivers will differ when dealing with position changes along the road, such as accelerating or decelerating, changing lanes and passing, to keep a safe and comfortable distance from the neighboring vehicles (Ohta, 1993; Miyajima et al., 2007). In the existing driver style analysis, the main goal is to extract the reasonable and relevant features to detect the aggressive drivers. It is also proved in the literature that aggressive driving behavior is critical to road safety as it is highly correlated with accidents and other traffic safety hazards (Association, 2009).

In this study, each vehicle is moving in the xy-plane and its location is represented by a pair of x, y coordinates in meters at one second time resolution. The distance, travel speed, acceleration, and driving direction at each time interval can be retrieved as D_t in mile, $v_t$in mph, $a_t$in mph/s, and $d_t$in degrees from the location-based dataset as follows, where the unit conversion from meter to mile and second to hour are conducted as needed.

$$D_t = \sqrt{(y_t - y_{t-1})^2 + (x_t - x_{t-1})^2} \tag{4}$$

$$v_t = D_t/t \tag{5}$$

$$a_t = (v_t - v_{t-1})/t \tag{6}$$

$$d_t = (y_t - y_{t-1})/(x_t - x_{t-1}) \tag{7}$$

As listed in Section 3, each feature of selected nine features has been normalized to [-1, 1], and thus the "standard" driver is the one with the GMM model most aligned with the normal distribution with mean vector of nine 0's and a covariance matrix corresponding to the nine elements, as the average of all drivers is defined as a standard model in this study.

4.3.2. Sample score distribution

To evaluate the driver behaviors of the library dataset, the proposed GMM-UBM approach is adopted. The dataset is first split to train the UBM model. 2236 drivers are taken out for UBM model while the remaining 500 drivers are used for GMM model training and cross validation. That is, 90 % of the individual trips from 500 drivers (i.e. 180 trajectory files) are used to train the individual driver model (GMM) and the remaining 10 % of the trips are treated as cross validation samples to evaluate the accuracy of the algorithm. It is worthy noting that the 180 trajectory files are randomly selected from the dataset, which may also include the fake trips. However, with iterative training of different random ones, the true driver behavior can be detected and extracted using the model.

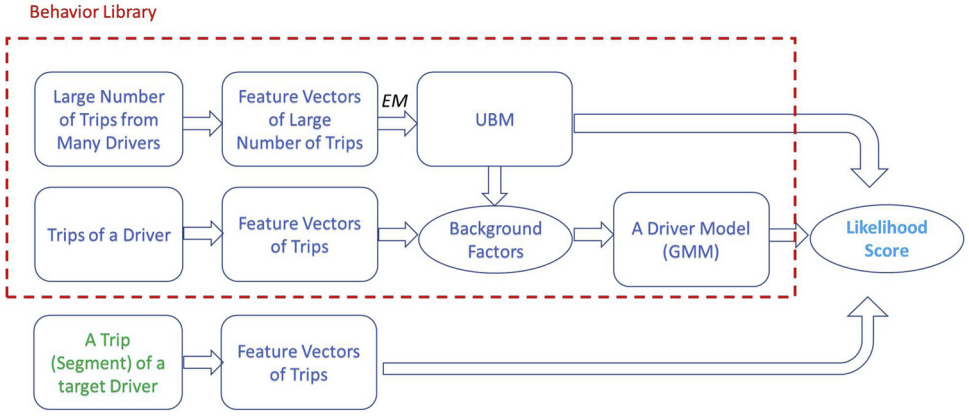

Fig. 2. Framework of GMM-UBM approach.

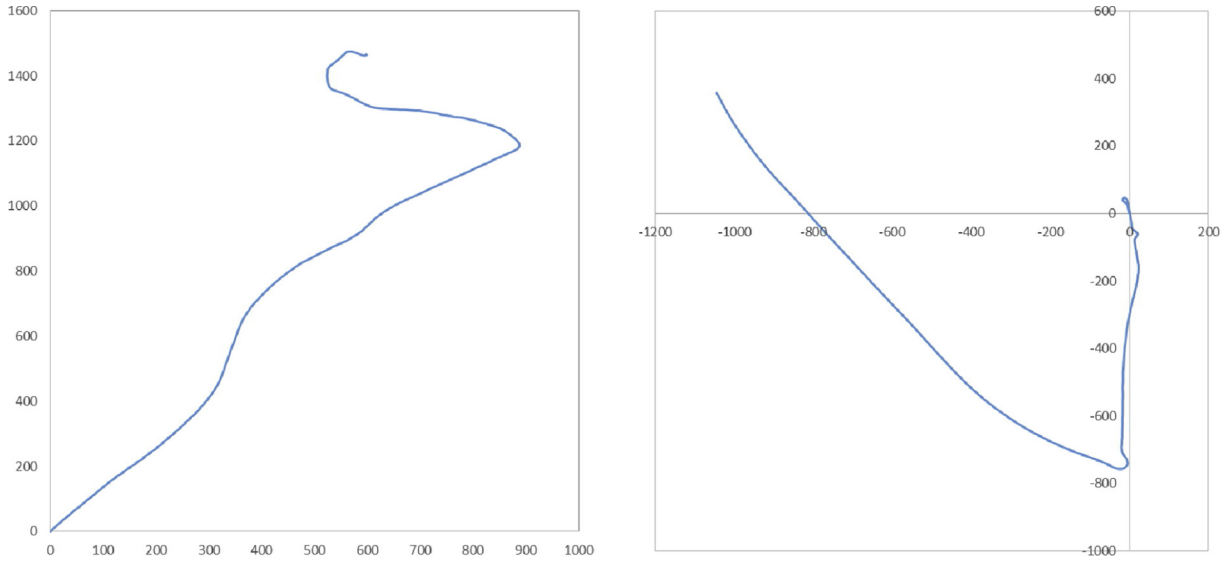

Fig. 3. Two trips of driver with id = 11.

Fig. 4(a)–(c) show the GMM models of three randomly selected individual drivers. The dashed lines for each individual driver show the distribution of each feature, and the solid red line is the Gaussian mixture model of all features for each driver. Fig. 4(d) shows the comparisons between the standard driver as the solid black line and selected individual driver in a dashed line. It is observed that Driver 1 with the red dashed line shares the most portion with the standard driver and has the highest safety score of 95.03 %. Driver 3, with the blue dashed line, shares the least portion and has a score of 86.11 %. Fig. 4(e) shows the scoring distribution of the tested 500 drivers. It is observed that most drivers are considered safe, with scores over 90, and only around 10 % of drivers scored below 85.

5. Experiment

The scoring system provides a basis of assessing individual drivers. To utilize the system in a practical manner and to improve the overall traffic network, the proposed system can suggest the relatively risky drivers to follow another nearby “safe” driver in a connected environment. The intent of this action is to limit the potential negative impact on network safety and influence of “bad” drivers on system performance. Although the warnings and suggestions may be useful in some cases, it cannot be ignored that some less-experienced drivers may not have the skills to interpret the warnings and translate them into correcting behaviors. A more intelligent mechanism enabled with connected vehicle techniques helps lower-scored drivers to follow higher-scored ones and reduce risks to safety. The nearby good drivers as well as good historical behavior patterns stored in the built-in library will ensure a satisfactory match for improving these drivers’ performances.

To implement this concept and evaluate its applicability to connected vehicles, the proposed system is integrated within a micro-simulation environment. In the baseline scenario, all the vehicles are initiated as unconnected vehicles. Two other scenarios with connectivity are then simulated and compared to this baseline scenario.

Case 1. All vehicles are designated as regular vehicles without any connectivity or scoring system.

Case 2. The scoring system is disabled, but connected vehicle technology is enabled so that 10 % vehicles are connected with a random generation algorithm proposed in (Talebpour et al., 2013).

Case 3. The driver scoring system is enabled; the top 10 % aggressive vehicles (10 % drivers with lowest scores) are targeted with an automatic search triggered to find the safest driver among the neighbouring vehicles and/or from the built-in library, and a recommendation option is provided to the targeted vehicles. Once the connection between them is established, the aggressive vehicle would follow the driving pattern of the safer one under a predefined compliance rate (including the acceleration/deceleration and lane changing scheme), and the status of the following vehicles is modelled as connected vehicles. The compliance rate allows drivers to reject the recommended connection.

It may be argued that the drivers might not be willing to follow the guidance system due to privacy concerns. Previous literature has investigated compliance rate and the willingness to follow a driver guidance/warning system. For instance, a survey on high-speed differential warning (HSDW) application was presented in (Li et al., 2017). It shows that 77 % of drivers have high potential compliance rates, and 55 % of them would react to a warning when HSDW is triggered for both lane-changing and acceleration/deceleration. Accordingly, it is assumed that for the connected vehicles, the compliance rates in cases 2 and 3 is 60 %, a number between 55 % and 77 % from literature. One may argue that drivers operating under a driver scoring system in case 3 may care more on safety issues and have higher potential compliance rates than the drivers without it in case 2. In this study, however, it is more important to focus on the effect of the driver scoring system, instead of evaluating sensitivity to compliance rates, confirming feasibility of connected vehicles, or checking privacy issues.

Each case runs 10 times with a different random seed each time in order to minimize the effect of stochastic variability. The experimental cases are conducted in the offline mode such that the driving score of case 3 is calculated based on the simulated trajectories of case 1, which means that each driver in case 3 has a “habitual behavior” match that can be extracted from the results of case 1.

The proposed scoring system is applied to the individual simulated trip data obtained from a 3.5-mile highway segment of Interstate 290 in Chicago (Fig. 5). The microsimulation tool utilizes different behavior rules for regular and connected vehicles. (Talebpour et al., 2016). It is designed to be able to capture the interactions between them and the collective effects on traffic flow dynamics. A total of 2500 vehicles are generated and simulated for 30-minute morning peak with a simulation interval of 0.1 s.

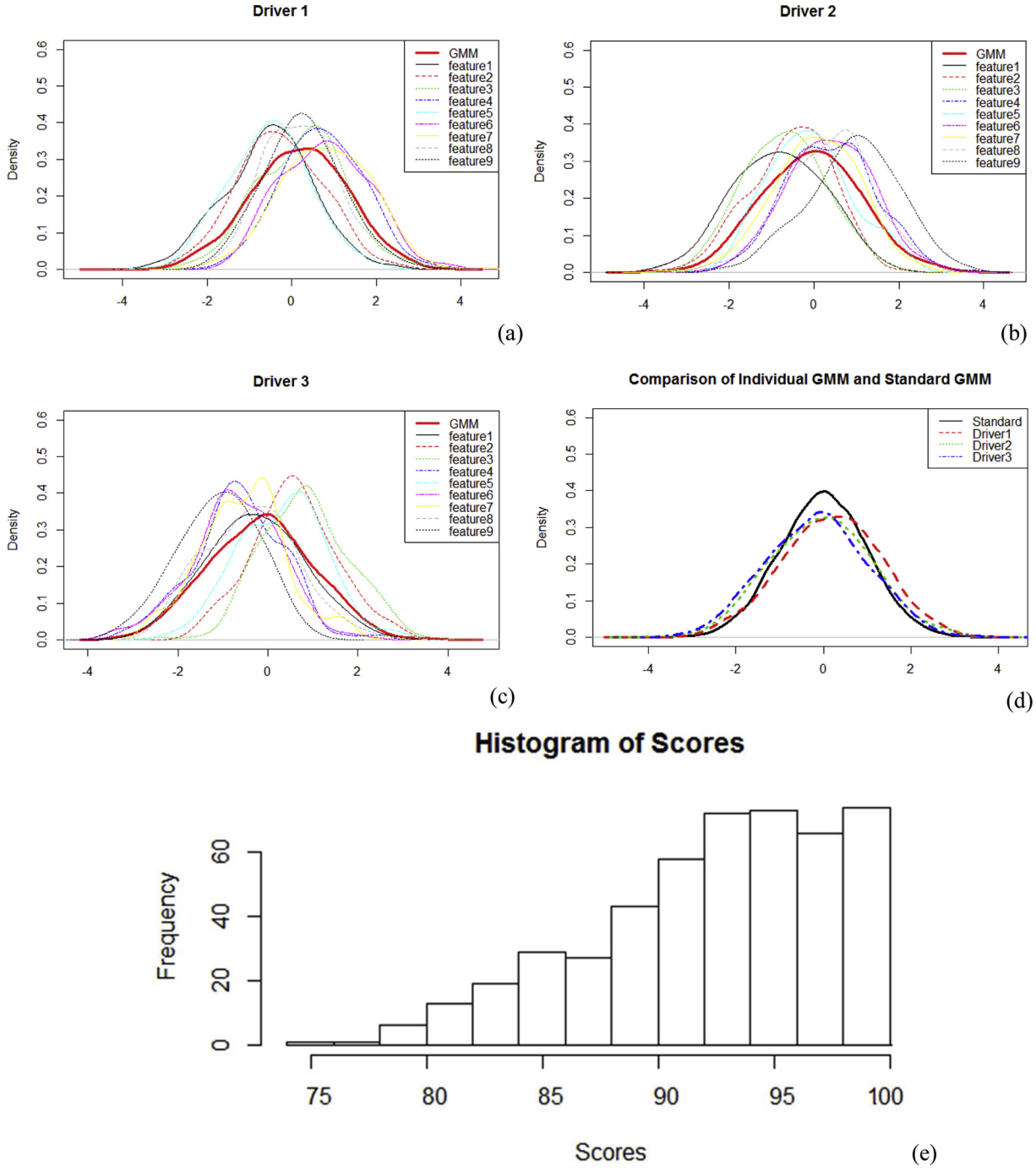

Fig. 4. (a)-(d) Examples of scoring results; (e) Histogram of 500 drivers' scores.

5.1. Microsimulation system

Car following (acceleration) and lane-changing modeling are core elements of microsimulation traffic models. The acceleration framework recognizes the differences in the longitudinal behavior of regular, connected, and autonomous vehicles, and is thus intended to capture distinct operational decision-making processes. Lane changing is a cause of perturbations in multilane traffic, and is especially sensitive to human error, particularly at high speed in high density environments. It aims to capture the tactical driving decision-making process (Mahmassani, 2016). This paper adopts the microsimulation model, within a duration-based framework at the tactical level and a utility-based framework at the operational level (Hamdar et al., 2008; Talebpour et al., 2013).

5.1.1. Modeling unconnected vehicles

A state-of-the-art car-following model is adopted in this study to model the acceleration behavior of regular vehicles (Talebpour et al., 2011). Utilizing prospect theory by Kahneman and Tversky (1979), a typical value function form (as showed in Eq. (8)) is introduced along with a typical weighting function to evaluate the acceleration choice by Hamdar et al. (2008).

$$U_{PT}(a_n) = \frac{[w_m + (1 - w_m)(\tanh(a_n) + 1]}{2} * \left(\frac{\left(\frac{a_n}{a_0}\right)}{1 + \left(\frac{a_n}{a_0}\right)^2} \right)^{\gamma}, \gamma > 0 \quad (8)$$

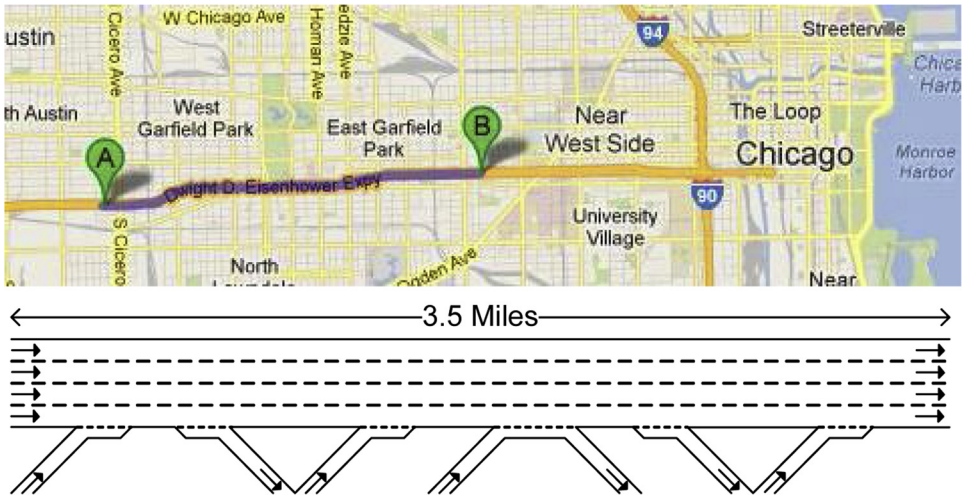

Fig. 5. Geometric characteristics of the selected segment in Chicago, IL.

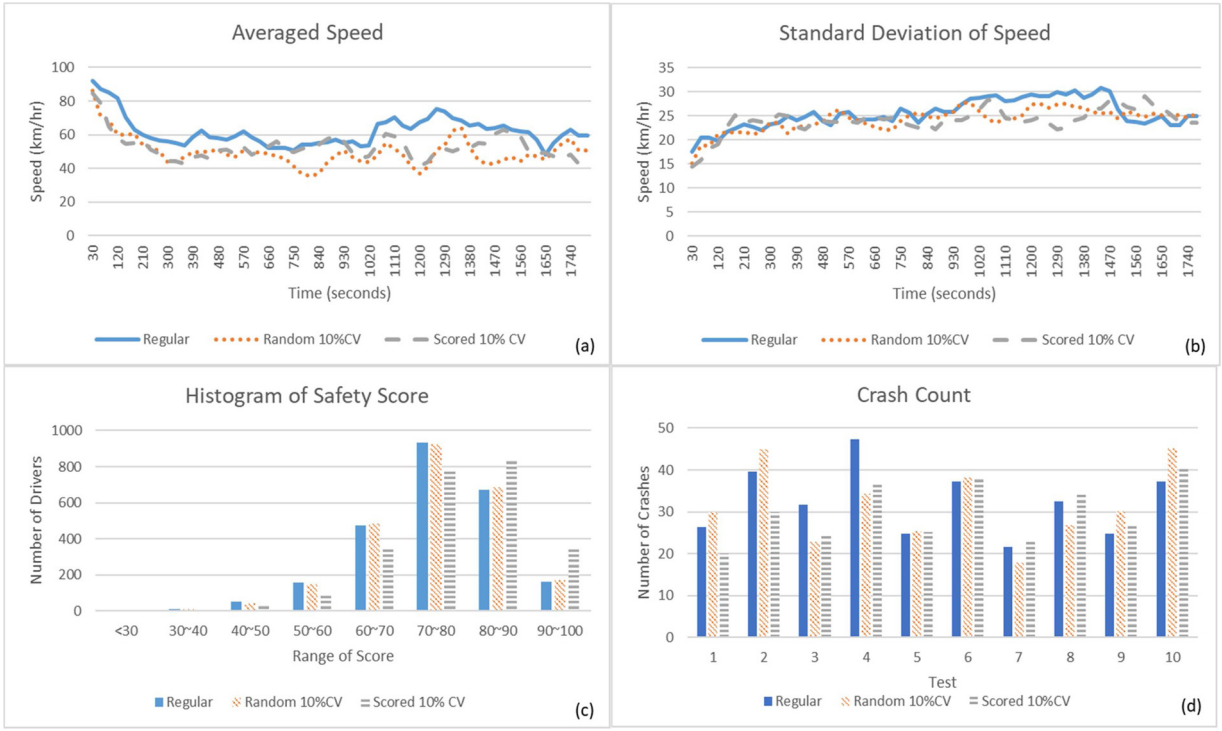

Fig. 6. Traffic system performance measures under three experimental cases.

where U_{PT} denotes the prospect theory acceleration value function, α_0 is the normalization parameter, γ and w_m are parameters to be estimated α_n, is the acceleration chosen by the driver at the unit of m/s^2.

In the majority of cases, the benefit to drivers is valued as U_{PT} by α_n taking as their acceleration rate unless they are involved into a rear-end collision. The disutility U from a crash is then modeled with the crash seriousness term $k(v, \Delta v)$, for the velocity of v and associated change in the velocity Δv (Talebpour et al., 2011).

$$U(a_n) = (1 - p_{n,i})U_{PT}(a_n) - p_{n,i}w_c k(v, \Delta v) \tag{9}$$

where $p_{n,i}$ is the probability of being involved in a rear-end collision and w_c is a crash-weighting parameter, which is lower for aggressive drivers $U_{PT}(a_n)$ is derived from Eq. (8).

The stochastic nature of the acceleration choice is modelled with the logistic functional form as follows (Talebpour et al., 2011):

$$f(a_n) = \begin{cases} \dfrac{e^{\beta_{PT}U(a_n)}}{\int_{a_{min}}^{a_{max}} e^{\beta_{PT}U(a')}da'} & a_{min} < a_n < a_{max} \\ 0 & Otherwise \end{cases} \tag{10}$$

where β_{PT} is a free parameter that reflects the sensitivity of choice to the utility $U(a_n)$.

Note that the parameters of this model are calibrated against Next Generation Simulation (NGSIM) data (Federal Highway Administration, 2006).

5.1.2. Modeling connected vehicles

The acceleration behavior of connected vehicles is modelled based on the Intelligent Driver Model (IDM) (Kesting et al., 2010) as IDM has been widely applied in the literature to simulate CV car following behavior. IDM specifies a following vehicle's acceleration as a continuous function of the vehicle's current speed (v_n), distance s_n to the leading

vehicle, and the difference between the leading and the following vehicles' velocities (Δv_n). Perceptive parameters such as desired acceleration, desired gap size, and comfortable deceleration are considered in this model (Treiber et al., 2000; Kesting et al., 2010):

$$a_{IDM}^{n}(s_n, v_n, \Delta v_n) = \bar{a_n}\left[1 - \left(\frac{v_n}{v_0^n}\right)^{\delta_n} - \left(\frac{s^*(v_n,\Delta v_n)}{S_n}\right)^2\right] \quad (11)$$

$$s^*(v_n,\Delta v_n) = s_0^n + T_n v_n + \frac{v_n \Delta v_n}{2\sqrt{\bar{a_n}\bar{b_n}}} \quad (12)$$

where δ_n (Free acceleration exponent), T_n (Desired time gap), $\bar{a_n}$ (Maximum acceleration), $\bar{b_n}$ (Desired deceleration), s_0^n (Jam distance), and v_0^n (Desired speed) are parameters to be calibrated. s^* is the desired (safe) gap. Note that the braking term in the IDM is designed to preclude crashes in the simulation.

5.2. Simulation results

Fig. 6 shows the experimental results. In Fig. 6(a) and (b), where the vertical axis of each figure is the speed and standard deviation of speed, and the horizontal axis displays the simulation time in seconds, it is confirmed that the connected vehicle (CV) scenarios, operating under either random selection or scoring system recommendation, exhibit the more stable speed distribution with less standard deviation compared to baseline. Also, the average speed of both CV cases is generally lower than the baseline case with regular vehicles, but the case with a scoring system manages to maintain speed stability beginning from the 300th second, much earlier than the randomly selected CV case.

Meanwhile, in Fig. 6(c), in which the horizontal axis is the range of scores and the vertical axis shows the number of vehicles in each range, it is confirmed that the safety score of all drivers have been generally improved by shifting to the higher score. It thus indicates that the driving behavior of all drivers is more normalized, including acceleration, deceleration, and lane changing behavior. The other two cases do not show much difference in the driving behavior distribution. The system with more normalized driving behaviors and fewer high-deviation drivers produces a safer driving environment as a result of smoother traffic conditions.

Fig. 6(d) shows the number of crashes for the individual test in the 10 runs. The number of crashes differs among the 10 runs, varying from 18 to 47. In case 1, the number of crashes fluctuates from 22 to 47, and it goes from 18 to 45 in case 2, and 20–40 in case 3. It indicates that with the driver scoring system, the maximum number of crashes is dropped from 45 to 40 and the total number. which is 323, 316, and 299 in each case, drops more than 10 %. It also verifies the system safety is improved compared with regular case and random 10 % CV case.

One may argue that the safety related measures (e.g. jerk, time to collision, gaps, etc.) to assess the performance should be included. In Fig. 6(c), according to the definition and feature extraction for the safety score, jerks are taken into consideration in this study. In Fig. 6(d), the total number of crashes are displayed, while showing that the average time to collision is reduced among connected vehicles enabled with the scoring system.

6. Conclusion and future work

This study developed a driver assessment and recommendation system to evaluate individual motorists' driving performance and improve the traffic conditions and safety from the driver's perspective. Machine learning methods, namely the GMM-UBM model and the maximum likelihood method, are adopted to capture driver signature; a similarity metric is then used to calculate the relative safety score of drivers' behaviors. To evaluate the proposed algorithm, the driver telematics data obtained from over 2000 anonymous drivers in an urban area were used to build the driving behavior library. The evaluation results confirm the capability of the GMM-UBM system to provide accurate detection, driving behavior characterization, and safety score calculation.

To verify the improvement in traffic conditions and safety with the proposed system, a series of microsimulation tests using an urban highway segment in Chicago were conducted. Once the scoring system is enabled, it provides behavior guidance by recommending the less safe driver to follow a safer driver using connected vehicle techniques. By comparing the simulation results with other cases, the tests show the significant improvement of the system stability as a result of the scoring system.

The simulation experiments are presented as an initial exploration of the potential of real-time follower-leader matching on the basis of driving scores to improve the quality and fluidity of traffic flow. Considerable effort and additional exploration are desired to perfect the concept and ensure its applicability in more practical environment under uncertainty. Nonetheless, the potential shown in this paper is sufficiently promising to warrant such additional development. Furthermore, the work here provides an example of how new sources of "big data" from vehicle trajectories can be leveraged towards enhanced safety and flow conditions.

The main contributions of this paper can therefore be summarized as follows: (1) it fills a gap in the literature by analyzing driver data to examine aggressive behavior in a connected driving environment, (2) it proposes a practical framework that employs rigorous data analytics to identify and mitigate aggressive driving behaviors to improve the overall driving environment through connected vehicle technologies, and (3) it demonstrates the potential safety improvement mechanisms resulting from the deployment of driving scores to guide driving recommendations via simulation.

Finally, several extensions of this work are proposed. First, different levels of connectivity for longer tests with more vehicles within other networks, especially in rural areas, would be worthwhile. Second, the data library could be improved by introducing more trajectory data in addition to other types of data describing the driving situations, particularly weather, road conditions, and the driving culture (i.e. social norms) in the area/city/country where the driving data is collected. Third, the proposed system is extendable to an on-line case which can be updated in real-time. Fourth, as stated in Section 5.2, the importance of compliance rate could be further explored with a set of more systematically designed experiments. With additional training data and more robust simulations, the attractiveness of this system for deploying a wider range of traffic management interventions and individual driver guidance is indeed possible.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Northwestern University Transportation Center (NUTC), and we are grateful to Professor Hani S. Mahmassani for his insight and expertise that greatly assisted this research. Dr. Alireza Talebpour and Dr. Samer Hamdar contributed to the development of the microsimulation tool used in the numerical experiments conducted for this paper. The authors remain solely responsible for the contents of the paper.

References

Administration, N.H.T.S., 2016. 2015 motor vehicle crashes: overview. Traffic Safety Facts Research Note 2016. pp. 1–9.

Ahmed, M., Saraydar, C.U., Elbatt, T., Yin, J., Talty, T., Ames, M., 2007. Intra-vehicular wireless networks. In: Proceedings of the Globecom Workshops. IEEE. pp. 1–9.
Anon, Https://developers.Google.Com/maps/documentation/distance-matrix/intro.
Association, A.A, 2009. Aggressive Driving: Research Update. American Automobile Association Foundation for Traffic Safety.
Cook, B., Ellegaard, P., Gilles, L., 2014. Driver risk assessment system and method employing selectively automatic event scoring. Google Patents.
Cook, B., Etcheson, J., 2014. Driver risk assessment system and method employing automated driver log. Google Patents.
Cook, B., Gilles, L., 2013. Driver risk assessment system and method having calibrating automatic event scoring. Google Patents.
Cook, B., Gilles, L., 2016. Driver risk assessment system and method having calibrating automatic event scoring. Google Patents.
Daniel, I.S., 2012. System and method for determining an objective driver score. Google Patents.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. J. R. Stat. Soc. Ser. B 1–38.
Deng, W., 2013. A Study on Lane-change Recognition Using Support Vector Machine. Phd dissertation at university of south florida. .
Federal Highway Administration, 2006. Next Generation Simulation: Us-101 Highway Dataset.
Feng, Y., Head, K.L., Khoshmagham, S., Zamanipour, M., 2015. A real-time adaptive signal control in a connected vehicle environment. Transp. Res. Part C Emerg. Technol. 55, 460–473.
Fha, D., 2010. Travel Time Reliability: Making it there on Time, All the Time. Us department of transportation, federal highway administration.
Guerrero-Ibanez, J.A., Zeadally, S., Contreras-Castillo, J., 2015. Integration challenges of intelligent transportation systems with connected vehicle, cloud computing, and internet of things technologies. IEEE Wirel. Commun. 22 (6), 122–128.
Hamdar, S., Treiber, M., Mahmassani, H., Kesting, A., 2008. Modeling driver behavior as sequential risk-taking task. Transp. Res. Rec. 2088, 208–217.
Igarashi, K., Miyajima, C., Itou, K., Takeda, K., Itakura, F., Abut, H., 2004. Biometric identification using driving behavioral signals. In: Proceedings of the Multimedia and Expo. IEEE International Conference on ICME'04. 2004. pp. 65–68.
Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47 (2), 263–291.
Kesting, A., Treiber, M., Helbing, D., 2010. Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity. Philos. Trans. Math. Phys. Eng. Sci. 368 (1928), 4585–4605.
Khan, S.M., Dey, K.C., Chowdhury, M., 2017. Real-time traffic state estimation with connected vehicles. IEEE Trans. Intell. Transp. Syst. 18 (7), 1687–1699.
Khondaker, B., Kattan, L., 2015. Variable speed limit: a microscopic analysis in a connected vehicle environment. Transp. Res. Part C Emerg. Technol. 58, 146–159.
Li, L., Werber, K., Calvillo, C.F., Dinh, K.D., Guarde, A., König, A., 2014. Multi-sensor soft-computing system for driver drowsiness detection. Soft Computing in Industrial Applications. Springer, pp. 129–140.
Li, W., Wu, G., Boriboonsomsin, K., Barth, M.J., Rajab, S., Bai, S., Zhang, Y., 2017. Development and evaluation of high-speed differential warning application using vehicle-to-vehicle communication. Transp. Res. Rec. 2621, 81–91.
Lu, N., Cheng, N., Zhang, N., Shen, X., Mark, J.W., 2014. Connected vehicles: solutions and challenges. IEEE Internet Things J. 1 (4), 289–299.
Mahmassani, H.S., 2016. 50th anniversary invited article—autonomous vehicles and connected vehicle systems: flow and operations considerations. Transp. Sci. 50 (4), 1140–1162.
Miyajima, C., Nishiwaki, Y., Ozawa, K., Wakita, T., Itou, K., Takeda, K., 2006. Cepstral analysis of driving behavioral signals for driver identification. Proceedings of the Acoustics, Speech and Signal Processing, 2006. IEEE International Conference on ICASSP 2006 Proceedings pp. V-V.
Miyajima, C., Nishiwaki, Y., Ozawa, K., Wakita, T., Itou, K., Takeda, K., Itakura, F., 2007. Driver modeling based on driving behavior and its evaluation in driver identification. Proc. IEEE 95 (2), 427–437.
Narla, S.R., 2013. The evolution of connected vehicle technology: from smart drivers to smart cars to… self-driving cars. Ite J. 83 (7), 22–26.
Ohta, H., 1993. Individual differences in driving distance headway. Vision in vehicles 4, 91–100.
Ozawa, K., Wakita, T., Miyajima, C., Itou, K., Takeda, K., 2005. Modeling of Individualities in Driving Through Spectral Analysis of Behavioral Signals.
Parameshwaran, R., 2016. Driver assessment and recommendation system in a vehicle. Google Patents.
Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted gaussian mixture models. Digit. Signal Process. 10 (1), 19–41.
Salvucci, D.D., Mandalia, H.M., Kuge, N., Yamamura, T., 2007. Lane-change detection using a computational driver model. Hum. Factors J. Hum. Factors Ergon. Soc. 49 (3), 532–542.
Shladover, S.E., 2018. Connected and automated vehicle systems: introduction and overview. J. Intell. Transp. Syst. Technol. Plan. Oper. 22 (3), 190–200.
Siegel, J.E., 2013. Cloudthink and the avacar: embedded design to create virtual vehicles for cloud-based informatics, telematics, and infotainment. Massachusetts Inst. Technol.
Siegel, J.E., Erb, D.C., Sarma, S.E., 2017. A survey of the connected vehicle landscape—architectures, enabling technologies, applications, and development areas. IEEE Trans. Intell. Transp. Syst. 19 (8), 2391–2406.
Spiegel, S., Gaebler, J., Lommatzsch, A., De Luca, E., Albayrak, S., 2011. Year. Pattern recognition and classification for multivariate time series. Proceedings of the Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data 34–42.
Talebpour, A., Mahmassani, H., Hamdar, S., 2011. Multiregime sequential risk-taking model of car-following behavior. Transp. Res. Rec. 2260, 60–66.
Talebpour, A., Mahmassani, H., Hamdar, S., 2013. Speed harmonization: evaluation of effectiveness under congested conditions. Transp. Res. Rec. 2391, 69–79.
Talebpour, A., Mahmassani, H.S., Bustamante, F.E., 2016. Modeling driver behavior in a connected environment: integrated microscopic simulation of traffic and mobile wireless telecommunication systems. Transp. Res. Rec. 2560 (1), 75–86.
Tiaprasert, K., Zhang, Y., Wang, X.B., Zeng, X., 2015. Queue length estimation using connected vehicle technology for adaptive signal control. IEEE Trans. Intell. Transp. Syst. 16 (4), 2129–2140.
Toledo, T., Musicant, O., Lotan, T., 2008. In-vehicle data recorders for monitoring and feedback on drivers' behavior. Transp. Res. Part C Emerg. Technol. 16 (3), 320–331.
Treiber, M., Hennecke, A., Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. Phys. Rev. E 62 (2), 1805.
Wakita, T., Ozawa, K., Miyajima, C., Takeda, K., 2005. Year. Parametric versus nonparametric models of driving behavior signals for driver identification. Proceedings of the AVBPA 739–747.
Wakita, T., Ozawa, K., Miyajima, C., Igarashi, K., Itou, K., Takeda, K., Itakura, F., 2006. Driver identification using driving behavior signals. IEICE Trans. Inf. Syst. 89 (3), 1188–1194.
Wilhelm, E., Siegel, J., Mayer, S., Sadamori, L., Dsouza, S., Chau, C.-K., Sarma, S., 2015. Cloudthink: a scalable secure platform for mirroring transportation systems in the cloud. Transport 30 (3), 320–329.