Chen, Depeng; Chen, Zhijun; Zhang, Yishi; Qu, Xu; Zhang, Mingyang; Wu, Chaozhong

# Driving Style Recognition under Connected Circumstance Using a Supervised Hierarchical Bayesian Model

*Published in:*
Journal of Advanced Transportation

*DOI:*
10.1155/2021/6687378

Published: 02/06/2021

*Document Version*
Publisher's PDF, also known as Version of record

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 6687378, 12 pages
https://doi.org/10.1155/2021/6687378

*Research Article*

# Driving Style Recognition under Connected Circumstance Using a Supervised Hierarchical Bayesian Model

**Depeng Chen**,[1] **Zhijun Chen**,[1] **Yishi Zhang**,[2] **Xu Qu**,[3] **Mingyang Zhang**,[4] **and Chaozhong Wu**[1]

[1]*Intelligent Transportation Systems Research Center, Wuhan University of Technology, Wuhan, China*
[2]*School of Management, Wuhan University of Technology, Wuhan, China*
[3]*School of Transportation, Southeast University, Nanjing, China*
[4]*School of Engineering, Aalto University, Espoo, Finland*

Correspondence should be addressed to Zhijun Chen; chenzj556@whut.edu.cn and Yishi Zhang; yszhang@whut.edu.cn

Received 18 November 2020; Revised 1 May 2021; Accepted 24 May 2021; Published 2 June 2021

Academic Editor: Tomio Miwa



In recent years, the automated driving system has been known to be one of the most popular research topics of artificial intelligence (AI) and intelligent transportation system (ITS). The journey experience on automated vehicles and the intelligent automated driving system could be improved by individualization driving understanding. Although previous studies have proposed methods for driving styles understanding, the individualization driving classification has not been addressed thoroughly. Therefore, in this study, a supervised method is proposed to understand driving behavioral structure and the latent driving styles by incorporating the prior knowledge. Firstly, a novel method is established for driving behavioral encoding and raw driving data mining. Then, the Labeled Latent Dirichlet Allocation (LLDA) is proposed to understand the latent driving styles from individual driving with driving behaviors. Finally, the Safety Pilot Model Deployment (SPMD) data are used to validate the performance of the proposed model. Experimental results show that the proposed model uncovers latent driving styles effectively and shows good agreement to real situations, which provides theoretical guidance on driving behavior recognition for better individual experience on automated driving vehicles.

## 1. Introduction

Automated vehicle technology has been developing rapidly and has been applied in public life in some cities. Automated vehicles, such as automated taxis and automated shared cars, can offer services for people's daily commute. Nowadays, automated taxis have been demonstrated practicable in cities like Guangzhou, Shanghai, and Changsha in China by cooperation between the government and companies like Baidu, Weride, and Robotaxi. It is estimated that automated vehicles will gain more significant growth and attention in the future. In the meantime, the satisfaction of people's ride plays a key role in the success of automated vehicles. Better and more individualized service can be provided if the driving styles of the drivers or the passengers are accurately recognized. However, the traffic context is usually complex and may vary in a short time; heterogeneous drivers may react differently even in the same traffic context, which is called the latent driving style. As illustrated in Figure 1, when approaching a slow-speed car (i.e., the red one), a driver with the moderate driving style (i.e., the blue one in the right-hand side of Figure 1, hereafter called the moderate driver) often tends to follow the front car instead of changing lane and overtaking. In this case, passengers adapting to the aggressive driving style (hereafter called the aggressive passengers) may feel uncomfortable, resulting in low satisfaction; similarly, the moderate passengers may experience discomfort when the aggressive driver decides to change lane to overtake the front car in the same situation. Analyzing and understanding the different driving styles of heterogeneous drivers automatically can thus help reduce the mismatch ratio for the passenger-driver pairs, in such a way as to enhance passenger satisfaction.

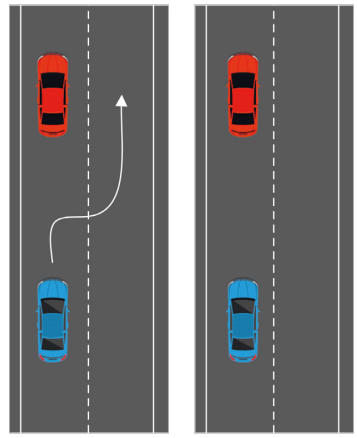


Figure 1: An example of different driving behaviors under same traffic context: (a) aggressive drivers change lane; (b) moderate drivers follow car.

Studies on driving styles have been performed recently; an enormous number of influencing factors result in different definitions of driving styles. Based on different definitions, approaches and sensor data are dissimilar among previous studies which mainly considered that driving styles are related to driver information of themselves and the driving data representing as driving behaviors. For instances of drivers' information, Rios-Torres et al. collect real Hybrid Energy Vehicle (HEV) driving data including average speed, acceleration, and user information like age, gender, and household income to extract driving styles from drivers for further research. Both drivers' internal factors and driving behaviors are taken into consideration for better driving style understanding [1]. Taubman-Ben-Ari et al. discovered relationship of driving styles among driver internal factors including age, gender, job, education, and personality [2]. A self-reporting instrument MDSI is developed to examine the associations between driving styles and factors mentioned. It discovers that self-esteem, need for control, sensation seeking, and extraversion are significantly related to driving styles. Javier et al. adopted heart rate sensors to discover the correlation between driving style and heart rate [3]. The research shows that aggressive behavior is between 2.5% and 3% beats per minute higher than quiet behavior. Our previous study also shows that driving context like pedestrian trajectories also have a significant impact on driving behavior [4]. Focusing on driving behaviors, Deng et al. designed a Driving Behavior Questionnaire to understand and classify drivers, including 28 items about driving behaviors [5]. By constructing speed curve and accel model to modify them, the speed curve and accel curve score are 27% better. However, only driving behaviors and subjective judgements may not correctly represent driving styles hidden in drivers. Marinez et al. indicated that only driver inner factors and driving behaviors may be excessively simplistic to represent driving styles. The driving style of a driver is a necessary form of individual driving, and same driving behaviors under different traffic context may differ extracting driving styles [6]. For example, while a front obstacle is on the road, drivers may drive with high speed or low speed. The former driver tends to be aggressive, while the latter may be categorized into "moderate" or "careful" class. In order to solve the problems of uncovering individualization driving, a model should be proposed to uncover the latent driving style in advance.

Latent Dirichlet Allocation (LDA)-based topic models have been innovatively applied in discovering latent topics from behavioral data rather than natural language processing [7], for instance, mining individual behavior patterns from individual behavioral patterns from (television) TV watching patterns [8], analyzing autonomous mobile robot behavior with a LDA-based method [9], discovering the underlying quantified structure through the LDA model and KFCM (kernel fuzzy C-means)-based algorithm [10], and understanding individualization driving states utilizing the LDA model [11]. However, compared with our previous work that the proposed model is unsupervised, the supervised model can utilize the prior knowledge (i.e., the class labels of the specific task) to learn an effective model, which make more significance for driving style recognition. Besides, focusing on vehicle motion data impairs the importance of traffic scenes in the complicated driving styles recognition, which may result in inaccurate driving style recognition. Overall, we propose a model that uses supervised knowledge and driving context data.

*1.1. Literature Review.* In recent years, there have been studies on driving style identifications in scientific research. Various methods are proposed and adapted in driving behavior recognition [11–13, 18]. The problem of driving styles recognition is mainly considered as a classification problem from data associated with driver information.

Many unsupervised machine learning methods for clustering are performed on driving styles recognition. Constantinescu et al. proposed PCA (principal component analysis) and HCA (hierarchical cluster analysis) to characterize behaviors with time-series vehicle motion data [12].

The motion data record each driver's driving behaviors precisely, five categories of driving styles from "nonaggressive" to "very aggressive" after analysis. Van Ly et al. used unsupervised method K-means and supervised learning methods SVM (supported vector machine) to conduct driver classification [13]. Driving event-based vehicle motion data from internal sensors are recognized. With different permutation of Brake, Turning, and Accelerate events, driving styles are classified as the cluster results. Chen et al. used the Gaussian mixture model to analyze driving signals including brake light switch, longitudinal and lateral accelerations, steering wheel angle, and vehicle speed [14]. Visual driving signal distributions of signals as acceleration profile for driving style recognition are presented, which makes benefits for better ADAS (advanced driver assistance system) design. Chu et al. employed the K-means model to cluster driving parameters including minimum time of driving in same lane, acceleration, and time of exceeding limit speed [15]. It reflects three types of driving events, accelerating, driving over speed, and changing lane. Comparing with fuzzy synthetic evaluation, the cluster similarity is over 60%. Zhang et al. proposed the DBSCAN model, an unsupervised clustering method, to classify driving style with vehicle following behavior characteristics represented with the Gipps model [16]. Liu et al. demonstrated a semisupervised Tri-CatBoost method to reduce the label data in driving style recognition [17]. This model effectively reduces label dependency of primitive high-dimensional driving behavioral data and improves classification accuracy.

Some studies point out that driving styles based on driving motion data can be modeled by hidden Markov process model. Deng et al. proposed methods based on the hidden Markov model to process time-series driving braking data [18]. Every brake event is coded with braking force, braking impulse, and time window of brake as hidden Markov state; the proposed model realizes effective discriminant of driving style. Sun et al. proposed a method for driving style recognition via Multidimension Gaussian Hidden Markov Process with the root mean square of vehicle acceleration sample [19]. Similarly, Wang et al. proposed a hierarchical hidden Markov model for driving pattern analysis for driving style extraction [20]. Car following behaviors segmented by accelerations, speeds, and distance to lead vehicle are extracted by HDP-HSMM (hidden Dirichlet process-hidden semi-Markov model). After driving patterns frequency distribution for drivers is shown using HDP-HSMM, driving style is easy to be recognized and labeled. Murphey et al. proposed a driving style classification model with jerk and speed data [21]. With comparison of different window sizes, the driving classification results using different window sizes and same driver may be classified as different driving styles. Wang et al. adapted a semisupervised support vector machine to analyze drivers' driving styles with longitudinal driving behaviors [22]. The proposed model introduces labeled data to help the SVM model build classifiers and then use the unsupervised model to cluster the driving data. To a certain extent, motion data collected from vehicles represent driving behaviors and their driving styles. However, these models neglected the significance of traffic context to driving behaviors.

The other researchers focus on what factors impact driving styles recognition the most, where different factors result in different comprehension of driving behaviors and driving styles. Marinez et al. provided a survey on factors on driving styles, including traffic context and vehicle motion [6]. Although the representation of raw data is various, feature extraction concentrates on speed distribution, acceleration time, deacceleration time, and average speed. Environmental factors (i.e., traffic situation, season, weather, road type, and road condition) are considered to influence driver's judgement which leads to different driving behaviors and different driving styles. Ishibashi et al. constructed a driving style Questionnaire for analysis [23]. The participants are required to answer 18 questions about daily driving. PCA is employed for driving style analysis; however, the questionnaire depends on subject judgement of how was driving experience under different circumstances. Cordero et al. proposed a hierarchical model for driving style recognition [24]. The driving style is composed of three levels of features including emotional state, driver state, and driving style. Applied with chronical approach and Ar2p approach, more complex driving style pattern is recognized better and more precise. Driving styles recognition with vehicle motion data become more high dimensional and complex considering traffic environmental factors. Latent relationship with traffic context is necessary to be discovered for better recognition.

Despite many challenges in driving style classifications, topic discovery models have become popular in feature extraction, such as LDA models [8–11, 25–27]. Zhang et al. developed a time-topic coupled LDA model to analyze the individual behavioral patterns from (television) TV watching patterns [8]; it discovers relationship of watching patterns between both watching behaviors and time periods. Duckworth et al. applied an LDA-based method for autonomous mobile robot behavior analysis [9]. Chen et al. utilized the LDA model for understanding individualization driving states, where three types of driving styles are recognized via driving motion information [11]. Chen et al. applied LDA with SVM to classify scenes with features extracted from pictures [25]. Features hidden under scenes are successfully extracted from various types of scenes. Liu et al. proposed the topic-link LDA model, which can extract hidden features in documents and discover similarity and community closeness between them [26]. Ramage et al. proposed a supervised LDA model to apply prior knowledge with labels, which allows the LDA model to learn from documents with corresponding labels [7]. With prior knowledge, the model itself models corresponded corpora for only some documents in which the labels are observed. Qi et al. provided a modified LDA model to leverage longitudinal driving behavior and discover latent driving style by data mining technique [10], where three types of driving styles are successfully classified by the LDA-based model. Bando et al. presented an LDA model to extract driving topics from discrete scenes segmented by a double articulation analyzer (DAA) from continuous driving behavioral

data [27]. The driving data they adapted only contain vehicle motion data related to the drivers but without traffic context.

*1.2. Objective and Contribution.* As mentioned above, the models shown above applied in behavioral data mining are most traditional generative models within the unsupervised learning framework. Moreover, the driving data they [10, 11, 19, 26] adapted more concentrate on vehicle motion data related to the drivers (i.e., acceleration and brake) to acquire driving topic. However, driving context obviously has significant contribution to driving style (driving topic) recognition. This inspires us to use a supervised topic model that can effectively deal with multiple data sources for driving style recognition. For further exploring the innovative usage of LDA in behavior pattern analysis, LLDA is introduced to adapt expert knowledge on drivers as a supervised driving behavior identification model for better uncovering latent individual driving characteristics. Thus, we utilize the labeled LDA model that connects the motion data and sensor data to comprehensively find out the latent driving styles under the driving behaviors of different drivers. In order to extract driving styles from the raw data, an encoding method is also proposed to mine and understand driving behaviors from the motion data and sensor data. The whole structure of our study is shown in Figure 2.

The main contributions of our work are threefold as follows:

First, the supervised LLDA model is introduced for driving style modeling, which can take into account the prior domain knowledge (i.e., driving styles labeled by experts) as the supervision of the model

Second, an encoding method, the motion and context aggregation model (MCAM) for the motion data and traffic context data is proposed for the driving behavior

Third, the MCAM + LLDA method combining LLDA and MCAM is proposed for effective driving style recognition, which can improve the classification accuracy of the driving styles

This paper is organized as follows. A more extensive description of the word-encoding method and the LDA-based model is presented in Section 2. The experiment and data mining are carried out in Section 3. The results and the analysis of driving styles and the proportions of driving styles for tested drivers are presented in Section 4, and the last section presents our conclusions.

## 2. Methodology

The LDA model is a successful topic discovery model to analyze the text in words, which is good at uncovering the latent topics under documents, consisted of words [7]. We use the LDA model to uncover the latent driving styles (topics) from driving behaviors (words) of different drivers (documents). Based on the above settings, the driver could be defined as a mixture of driving styles. Driving styles pervade driving behaviors, which are hidden variables in the proposed model. The driving behaviors are from the raw driving motion data and radar data, which can be observed in our proposed model. The proposed model can automatically organize, understand, and summarize the driving behaviors, which can achieve driving styles distribution estimation. However, the problem of driving behavior constitution from the raw data remains. Here, we introduce a method of words encoding for driving behaviors, i.e., the motion and context aggregation model as follows.

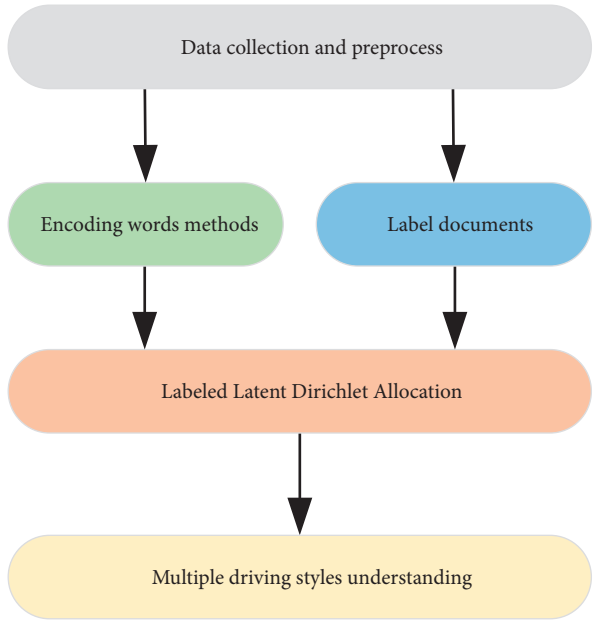


Figure 2: The structure diagram of this study.

*2.1. Motion and Context Aggregation Model.* In this study, driving behaviors are extracted from raw vehicle motion data and radar data through data mining techniques. The method of encoding driving behavior on raw data is shown Figure 2. Although speed and acceleration data are continuous driving data, a slight difference in value is not different in driving behavior recognition. For example, 10 KM/h is not different with 11 KM/h on reflecting the driving style under the same driving context. It is unnecessary to classify them as two driving behavior. So, we merge them to set up a range for a cluster that represent this feature. The driving behaviors consist of five features in the light blue frames. The features can be described as two parts: driving behavior and driving scene. The first part is from the motion data, including speed, acceleration, and turn signals. The other part is from the radar data, which includes front objective and lane offset. For example, acceleration is categorized into two categories, named positive and negative. Both categories of positive and negative acceleration are divided into five types named very low, low, middle, high, and very high, respectively. All types should be concluded in raw data. Hence, for continuous features like acceleration and speed, we set a maximum absolute value as a whole range and then divided into five equal intervals (five types). The combination with categories would become ten possible

features state for one feature. For noncontinuous features like turn signals, we just take all of its states as types to contribute to the behavior word combination. As for obstacles data, in one specific period, obstacles in vision may have more than one obstacle like cars, buses, or others that block the way. We divide the obstacle into two categories and three types, which indicates that each obstacle detected in the radar is processed with one category and one type. In the same period of one driving behavior, only one obstacle is valid in a block that is combined by a category and a type. Six blocks are in the preceding visual area and each one is at a binary state. Zero means no obstacle, and one means the opposite. The light green frames present a simple classification from driving data. Five features can be combined as the corpus of words, and one trip records raw data in sequence as a document.

Based on the above encoding of vehicle motion data and sensor data, the words in the LLDA model can be formed as a tuple constituted of speed factor, acceleration factor, turn signal factor, obstacles factor, and lane offset factor. For example, the driving behavior of a driver is described as "very low speed, very low acceleration, change lane, close front obstacle, on lane" in a specific period. Also, it can be reassembled into a sentence like "the driver change lane with very low speed and very low acceleration while a front obstacle is closed and the vehicle is on a lane." One trip of a driver is considered as a document consisting of these sentences. The word-encoding process is shown in Figure 3, and their encoding of driving behaviors is shown in Table 1. The continuous driving data are formed into discrete driving behavior sequences, which meet the requirement of the LDA model for the natural language process. The frequency distributions of driving behaviors are presented after encoding of driving behavioral data, and driving behaviors represent long-tail feature. We only show the top 100 frequent words in Figure 4.

*2.2. Driving Style Recognition Model.* Given the above information and description, driving behaviors and latent driving styles with individualization driving are well defined. The mathematical notation is shown in Table 2. In the proposed model, the driving styles are categorized into three types named "aggressive driving," "moderate driving," and "careful driving" (the number of driving styles can increase quickly by the topic number K). Each individualization driving has a proportion of driving styles. Encoded driving behaviors determine driving styles. The typical LDA model is an unsupervised model without prior knowledge. Labeled LDA (LLDA) is one of the typical supervised LDA models [6]. It contains the traditional LDA model with a one-to-one correspondence between the latent topics and tags of every documents. The original LLDA model succeeds in discovering restrictive topics in labeled documents with a supervised method. It differs from unmodified LDA, and topics are learned from supervised tags, which enhance the performance from rich information in tags for documents. In this study, the driving styles would be constrained by those topics that correspond to a driver (document) label set. According to the graphic model from Figure 5 and the notation shown in Table 1, each driver $m$ is represented by a list of driving behavior indices $\mathrm{w}_{(\mathrm{m})} = (w_1, \ldots, w_{Nm})$ and a list of binary topic indicators $\Lambda_{(m)} = (l_1, \ldots, l_k)$ where each $w_\mathrm{i} \in \{1, \ldots, V\}$ and each $l_\mathrm{k} \in \{0, 1\}$. For the generative process, different from the traditional LDA, the multinomial distribution corresponds to its labels $\Lambda_{(m)}$. The generative process is shown in Table 3.

The labels are generated using a Bernoulli coin toss, for each topic $k$, with a labeling prior probability $\Phi_\mathrm{k}$. The driver-specific label projection matrix $L_{(m)}$ is of size $D_\mathrm{m}$ times $K$ for each driver $m$. The matrix is defined as follows. For each row $i \in \{1, \ldots, M_\mathrm{d}\}$ and column $j \in \{1, \ldots, K\}$, matrix $\mathrm{L}_{(m)}$ is assigned by

$$L_{ij}^{(m)} = \begin{cases} 1, & \text{if } \lambda_\mathrm{i}^{(\mathrm{m})} = j, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The Dirichlet topic prior $\alpha = (\alpha_1, \ldots, \alpha_\mathrm{K})$ $T$ is projected by $L_{(m)}$ to a lower-dimensional vector $\alpha_{(d)}$ as follows:

$$\alpha^{(m)} = L^{(m)} \times \alpha = \left( \alpha_{\lambda_1^{(m)}}, \ldots, \alpha_{\lambda_{D_m}^{(m)}} \right). \tag{2}$$

This means the latent topics are restricted to their labels. Given the labels $\Lambda(m)$, the labeling prior $\Phi$ is separated from the rest of the model. Gibbs sampling is also applied for model training. The sampling probability for the driving style z in a driver $m$ in the labeled LDA model is given in

$$P(z_i = j | z_i) \propto \frac{n_{i,j}^{w_i} + \beta_{w_i}}{n_{i,j}^{(\cdot)} + \beta^T 1} - \frac{n_{i,j}^{m} + \alpha_j}{n_{i,j}^{(m)} + \alpha^T 1}. \tag{3}$$

Equation (3) will be applied in the Gibbs sampling in our study, and finally the proportion of driving styles from the raw driving data and given labels is obtained, where $n_{i,j}^{w_i}$ means the count of driving behavior $w_i$ in driving style $j$ but not including the currently assigned driving style $z_w$. The results of the Gibbs sampling on labeled LDA will be shown in the next section.

## 3. Experiment and Data Mining

*3.1. Dataset.* The SPMD (Safety Pilot Model Deployment) data we used in the proposed model are collected from the equipment implemented on vehicles and roadside devices [28]. This model is conducted by UMTRI (the University of Michigan Transportation Research Institute) in Ann Arbor, Michigan. These data were collected during the Safety Pilot Model Deployment (SPMD). The datasets that these entities will provide include basic safety messages (BSMs), vehicle trajectories, and various driver-vehicle interaction data, as well as contextual data that describe the circumstances under which the model deployment data were collected. Large portion of the data contained in this environment is obtained from on board vehicle devices and roadside units.

The experimental environment we used is running in python 3.6. For this data file with large volume, these files are processed by python package "pandas" to divide into small volume files. The dataset includes BrakeEvents, BSMEvents, DataFrontTargets, HV_Primary, and HV_Radar. The data

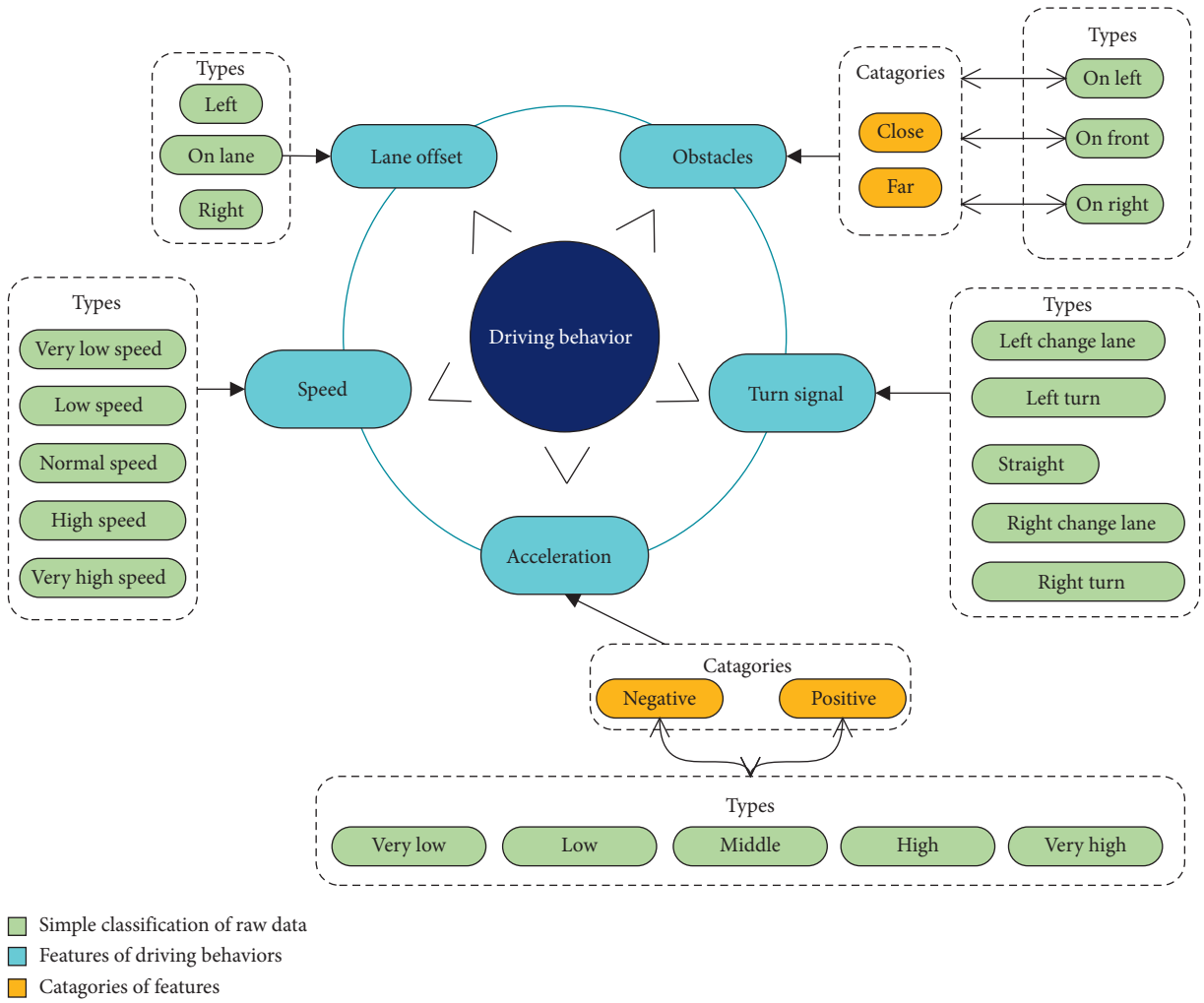


Figure 3: Words encoding model for five features of driving behaviors.

Table 1: The driving behavior of the sample driving data.

| Lane offset | Obstacle | Acceleration | Speed | Signal |
|---|---|---|---|---|
| On lane | Close on front and close on left | Negative low | Very high speed | Straight |
| On lane | Close on front and close on left | Negative low | Very high speed | Straight |
| On lane | Close on front and close on left | Negative low | Very high speed | Straight |
| On lane | Close on front and close on left | Negative low | Very high speed | Straight |
| On lane | Close on front and close on left | Negative low | Very high speed | Straight |
| On lane | Close on left | Negative very high | High speed | Turn left |
| On lane | Close on left | Negative very high | High speed | Turn left |
| On lane | Close on left | Negative very high | High speed | Turn left |
| On lane | Close on left | Negative very high | High speed | Turn left |
| On lane | Close on left | Negative very high | High speed | Turn left |

used in this paper are HV_Primary and HV_Radar file. The HV_Primary.csv file contains all detailed operation data during the test, including instantaneous driving motion data such as device ID, Trip number, speed, acceleration, the turn signal, and GPS position. The HV_Radar file provides the data collected from a radar unit which are compatible with device ID and Trips. All obstacles detected with speed and range in $X$ and Y coordinates are also included. Sample raw datasets are shown in Table 4.

Each record in this dataset has a trip number for driver recognition. The dataset is sorted by trip number then listed as time sequence records, which has time resolution of 100 milliseconds. Disqualified or invalid records in dataset have been removed for denoising. As explained in MCAM, the continuous features like speed or acceleration, we set the maximum speed value as 30 meter per second which is 108 kilometer per hour, and the record with speed value which is over the maximum value would be removed.

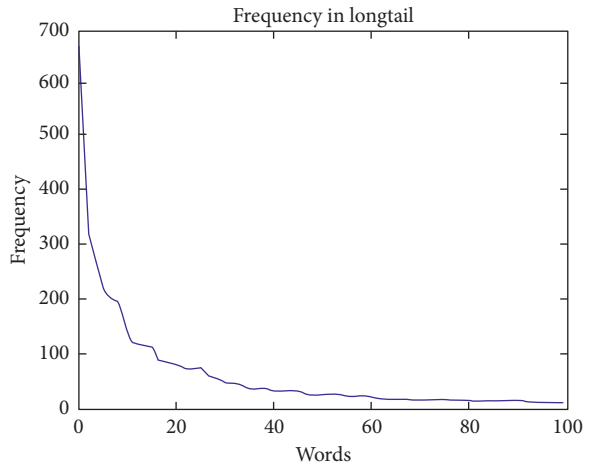


Figure 4: The frequency distributions over driving behavior words on parted driving data.

Table 2: Mathematical notation for the labeled LDA model.

| Notation | Description |
|---|---|
| $\Theta$ | Multinomial distribution over driving styles |
| $\Psi$ | Multinomial distribution over driving behaviors |
| A | Dirichlet prior parameters for $\Theta = (\theta_1, ..., \theta_M)$ |
| B | Dirichlet prior parameters for $\Psi = (\psi_1, ..., \psi_K)$ |
| $M$ | Number of drivers |
| $N_m$ | Number of driving behaviors in m-th drivers |
| $K$ | Number of driving styles |
| $z_w$ | The topic of the driving behavior $w$ |
| $w_i$ | The ith driving behavior |
| $\Lambda^{(m)}$ | The label vector of driver m |
| $\Phi_k$ | Label prior for topic k |

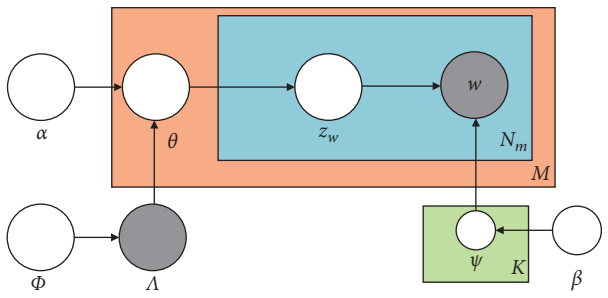


Figure 5: The graphical model of LLDA

*3.2. Data Processing.* The variety and multidimension of sample data fulfil our requirement for the driving behavior description, and we take 300 trips from the data as 300 documents in our model. For the supervised model, over a million records of driving behaviors in a document are applied in train set and test set. We encode the raw data as the format illustrated in Table 4 with our motion and context aggregation model into five features. Then, the driving behavior consists of these five features.

We used 10-fold cross validation (random seed = 1, ..., 10) to promote the accuracy result; for example, the test dataset 1 represents ten drivers randomly selected in the whole dataset. In our study, the types and quantities of vehicle motion data not only enhance the performance of data mining result of driving behaviors but also raise the difficulty and complexity of the analysis process. The continuous data are discretely encoded into words as shown from Figure 2. In LDA, the driver-driving style distribution is restricted to prior parameter alpha (if there is prior knowledge about the distribution, it can be embedded by setting alpha elaborately). The topic-words distribution is restricted to the prior parameter beta (similar to alpha). Usually, each entry of alpha (and beta) is heuristically set to be a same value because we do not have any prior information about such two distributions. Following the suggestion of the study of Thomas and Mark [29], the Dirichlet prior parameters value alpha is set as 16.67 which is 50 divided by 3 (the number of topic) and beta is set as 0.01. The beta which is set as 0.01 and relatively small means it can be expected to result in a fine-grained decomposition of the driving behaviors into driving styles.

Table 3: Generative process for the labeled LDA model.

| No. | Steps |
|---|---|
| 1 | For each topic $k \in \{1, ..., K\}$: |
| 2 | Generate $\Psi_k = (\psi_{k,1}, ..., \psi_{k,V})^T \sim \text{Dir}(\cdot\|\beta)$ |
| 3 | For each driver $d$: |
| 4 | For each topic $k \in \{1, ..., K\}$ |
| 5 | Generate $\Lambda(m)$ $k \in \{0, 1\} \sim \text{Bernoulli}(\cdot\|\Phi_k)$ |
| 6 | Generate $\alpha^{(m)} = L^{(m)} \times \alpha$ |
| 7 | Generate $\theta^{(m)} = (\theta_{l_1}, ..., \theta_{l_{D_m}})^T \sim \text{Dir}(\|\alpha^{(m)})$ |
| 8 | For each $i$ in $\{1, ..., N_m\}$: |
| 9 | Generate $z_i \in \{\lambda_1^{(m)}, ..., \lambda_{D_m}^{(m)}\} \sim \text{Mult}(\cdot\|\theta^{(m)})$ |
| 10 | Generate $w_i \in \{1, ..., V\} \sim \text{Mult}(\cdot\|\psi_{z_i})$ |

For adapting prior knowledge in our proposed model, expert knowledge to label driving data is necessary. The speed and acceleration curves from a driving sample of prior information about drivers which helps us to label driving styles from drivers are shown in Figure 6. Three driving styles are assumed for drivers in our study, and we labeled the small amount of collected data with most possible two labels by expert knowledge. For each driver in our validation dataset, the speed curve and acceleration curve are illustrated. Driver's average speed, average acceleration, max speed, and max acceleration are quantified and discretized into three levels for driving style labeling. For instance, the average speed curve of driver 2 in dataset 1 is around 15 m/s and average acceleration is around 2 m/s$^2$. The max speed is not over 35 m/s. All in all, this driver is labeled as "moderate" and "careful" according to the speed and acceleration curve feature.

## 4. Results and Discussion

Based on the data and the proposed model, the latent driving styles distribution is uncovered from the driving behavioral data. The average accuracy of the driving styles identification is shown in Table 5. We compare the following clustering algorithms:

Table 4: The driving behavior features of the sample driving data.

| Radar_X | Radar_Y | Acceleration | Speed | Left signal | Right signal | Left change | Right change | Lane width | Right mark |
|---|---|---|---|---|---|---|---|---|---|
| −5.41 | 82.78 | −0.30 | 16.66 | 0 | 0 | 0 | 0 | 3561.08 | 2077.0 |
| −5.98 | 78.24 | −0.30 | 16.66 | 0 | 0 | 0 | 0 | 3561.08 | 2077.0 |
| −5.95 | 76.22 | −0.30 | 16.66 | 0 | 0 | 0 | 0 | 3561.08 | 2077.0 |
| −5.95 | 74.17 | −0.30 | 16.66 | 0 | 0 | 0 | 0 | 3561.08 | 2077.0 |
| −6.17 | 72.16 | −0.30 | 16.66 | 0 | 0 | 0 | 0 | 3561.08 | 2077.0 |
| 1.984 | 8.704 | −2.78 | 9.61 | 1 | 0 | 0 | 0 | 2799.08 | 1361.0 |
| 1.89 | 9.216 | −2.78 | 9.34 | 1 | 0 | 0 | 0 | 2905.76 | 1412.0 |
| 1.95 | 9.152 | −2.78 | 9.34 | 1 | 0 | 0 | 0 | 3032.76 | 1463.0 |
| 1.92 | 8.32 | −2.31 | 9.22 | 1 | 0 | 0 | 0 | 3169.92 | 1513.0 |
| 1.89 | 8.288 | −2.16 | 8.72 | 1 | 0 | 0 | 0 | 3205.48 | 1549.0 |

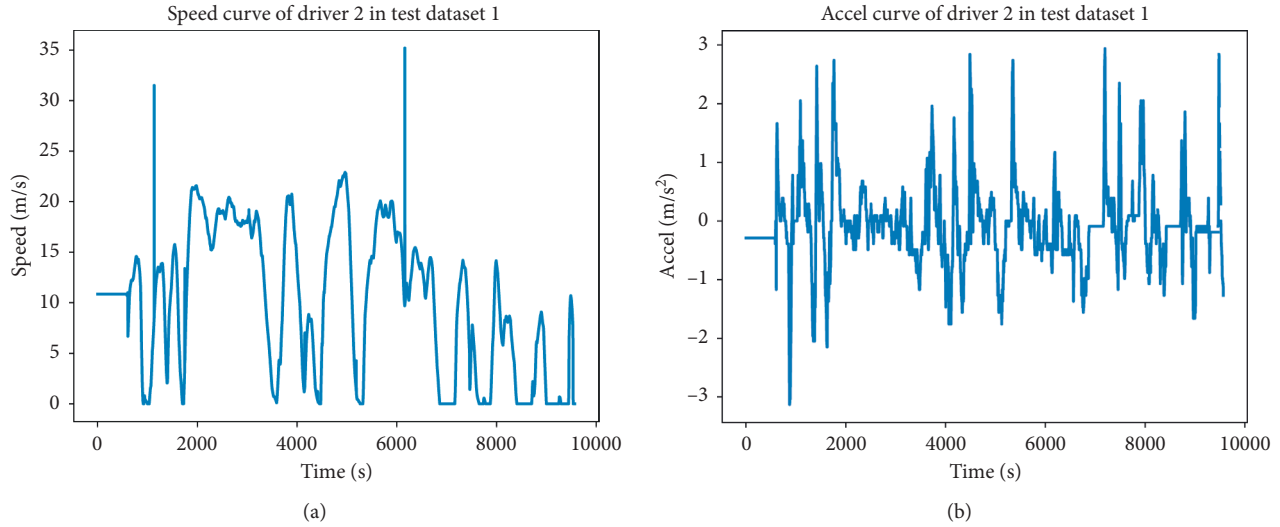


Figure 6: The driving sample of vehicle motion from Driver 2 in test dataset 1: (a) speed curve of three sample driving; (b) acceleration curve of three sample driving.

Table 5: Results of driving styles identification average accuracy in test datasets.

| Model dataset | SVM | NBC | KNN | Proposed method |
|---|---|---|---|---|
| Test dataset 1 | 0.121 | 0.346 | 0.47 | 0.50 |
| Test dataset 2 | 0.191 | 0.354 | 0.46 | 0.60 |
| Test dataset 3 | 0.198 | 0.396 | 0.46 | 0.50 |
| Test dataset 4 | 0.119 | 0.320 | 0.46 | 0.70 |
| Test dataset 5 | 0.201 | 0.369 | 0.45 | 0.60 |
| Test dataset 6 | 0.188 | 0.314 | 0.45 | 0.70 |
| Test dataset 7 | 0.208 | 0.391 | 0.40 | 0.65 |
| Test dataset 8 | 0.208 | 0.377 | 0.47 | 0.70 |
| Test dataset 9 | 0.191 | 0.338 | 0.47 | 0.50 |
| Test dataset 10 | 0.204 | 0.353 | 0.44 | 0.60 |
| Averages | 0.183 | 0.356 | 0.463 | 0.605 |

(i) SVM [29]—support vector machine—is a supervised learning model used for classification analysis. The data would be separated by a constructed hyperplane.

(ii) NBC [30]—Naive Bayes classifier—is a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independent assumptions between the features.

(iii) KNN [31]—k-nearest neighbour—is a method used by measuring distances between features for classification and clustering.

The average accuracy (60.5%) of the proposed model is higher than that of compared methods. In order to evaluate the driver individualization classification, the mean performance of macro F1 is also shown in Tables 5 and 6. The macro-F1 value is defined as the following harmonic average of recall and precision:

$$F_1 = \frac{1}{N}\sum_{k=1}^{|C|}\delta C_k \frac{2pr}{p+r}, \tag{4}$$

where $N$ denotes the sample size and $\delta C_k$ denotes the number of the samples with class label $C_k$. The higher

Table 6: Results of driving styles identification average macro-F1 in test datasets.

| Model | SVM | NBC | KNN | Proposed method |
|---|---|---|---|---|
| Test dataset 1 | 0.180 | 0.348 | 0.39 | 0.504 |
| Test dataset 2 | 0.266 | 0.354 | 0.349 | 0.636 |
| Test dataset 3 | 0.274 | 0.385 | 0.34 | 0.529 |
| Test dataset 4 | 0.176 | 0.312 | 0.329 | 0.712 |
| Test dataset 5 | 0.277 | 0.380 | 0.35 | 0.667 |
| Test dataset 6 | 0.262 | 0.309 | 0.283 | 0.598 |
| Test dataset 7 | 0.285 | 0.331 | 0.405 | 0.642 |
| Test dataset 8 | 0.285 | 0.388 | 0.375 | 0.694 |
| Test dataset 9 | 0.266 | 0.325 | 0.327 | 0.471 |
| Test dataset 10 | 0.281 | 0.357 | 0.306 | 0.639 |
| Averages | 0.222 | 0.349 | 0.345 | 0.612 |

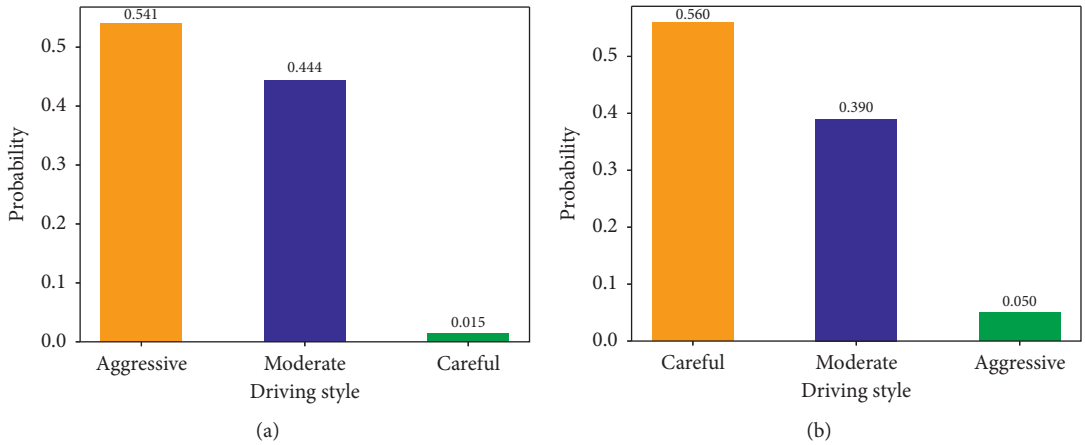


Figure 7: The example of the general distribution of driving styles for two drivers: (a) the distribution of driver 3 in test dataset 2; (b) the distribution of driver 8 in test dataset 2.

weighted macro F1 can demonstrate the better performance shown in a method.

To further value the performance of the proposed LDA-based model, the perplexity (PPL) is also reported. The perplexity intuitively interprets the expected size of words with uniform word distribution that the model needs to generate a specific sentence. The lower value of perplexity indicates that a natural language processing model requires fewer possibilities to choose words with a corpus of documents, which offers a lower misrepresentation and better understanding of the words of the documents by the learned latent topics. The log perplexity value is defined as the follows:

$$\log P\left(\widetilde{\vec{w}}_{\widetilde{m}}|M\right) = \sum_{t=1}^{V} n_{\widetilde{m}}^{(t)} \log\left(\sum_{k=1}^{k} \varphi_{k,t} \cdot \vartheta_{\widetilde{m},k}\right), \quad (5)$$

where $n_{\widetilde{m}}^{(t)}$ denotes the number of times of driving behavior $t$ has been observed in driver $\widetilde{m}$ and $\vartheta_{\widetilde{m},k}$ is derived by querying model after the Gibbs sampling, which denotes the driving style $k$ distribution of the driver $\widetilde{m}$. The average log perplexity is 88.83 using the proposed model. 1843 tokens of corpus from each sample driving data are extracted. This also confirms the superiority of our proposed model.

Results are the proportions of driving styles on each driver. The latent driving styles of two drivers learned from the introduced model are illustrated in Figure 7. The driving style is a mixture combination of the three kinds of driving styles. For example, in Figure 7(a), this driver's driving styles hold proportions that include 54.1% of aggressive driving, 44.4% of moderate driving, and 1.5% of careful driving.

The deduction of driving style to this driver is done carefully, but aggressive driving style is also a considerable part, which matches the label of this driver as "careful moderate." The results show that the proposed model can effectively uncover the driving styles and advance the performance of driving behavior understanding. The average proportions of driving states for individualization driving and their driving style corresponding labels are presented in Table 7 in the experiment, and it shows the proportions of driving style for 30 documents in dataset 2 which is used in cross validation.

Table 7: Results of the proportions of driving styles for individualization driving in dataset 2.

| Driver ID | Aggressive | Moderate | Careful | Labels |
|---|---|---|---|---|
| 1 | 0.009 | 0.173 | **0.817** | Moderate, careful |
| 2 | 0.985 | 0.007 | 0.007 | Aggressive, careful |
| 3 | 0.047 | 0.048 | **0.903** | Aggressive, careful |
| 4 | 0.968 | 0.014 | 0.016 | Moderate, careful |
| 5 | 0.062 | **0.929** | 0.009 | Moderate, aggressive |
| 6 | 0.003 | 0.013 | **0.984** | Moderate, careful |
| 7 | **0.992** | 0.004 | 0.002 | Aggressive, careful |
| 8 | 0.11 | **0.871** | 0.01 | Moderate, careful |
| 9 | 0.981 | 0.011 | 0.008 | Moderate, careful |
| 10 | 0.026 | 0.031 | **0.943** | Moderate, careful |
| 11 | 0.028 | **0.962** | 0.01 | Moderate, careful |
| 12 | 0.965 | 0.02 | 0.013 | Moderate, careful |
| 13 | 0.29 | **0.628** | 0.081 | Moderate, careful |
| 14 | 0.992 | 0.005 | 0.001 | Moderate, careful |
| 15 | 0.001 | 0.002 | **0.997** | Moderate, careful |
| 16 | 0.002 | 0.003 | **0.995** | Moderate, careful |
| 17 | 0.956 | 0.036 | 0.008 | Moderate, careful |
| 18 | 0.008 | **0.988** | 0.004 | Moderate, careful |
| 19 | 0.002 | 0.002 | **0.995** | Moderate, careful |
| 20 | 0.004 | 0.004 | **0.992** | Aggressive, careful |
| 21 | **0.991** | 0.004 | 0.001 | Aggressive, careful |
| 22 | 0.001 | 0.001 | **0.998** | Aggressive, careful |
| 23 | **0.983** | 0.013 | 0.004 | Aggressive, careful |
| 24 | **0.532** | 0.465 | 0.003 | Aggressive, careful |
| 25 | 0.046 | 0.105 | 0.849 | Moderate, aggressive |
| 26 | 0.004 | **0.992** | 0.004 | Moderate, aggressive |
| 27 | 0.98 | 0.016 | 0.002 | Moderate, careful |
| 28 | 0.001 | 0.002 | **0.997** | Moderate, careful |
| 29 | 0.005 | 0.003 | 0.992 | Moderate, aggressive |
| 30 | **0.988** | 0.009 | 0.003 | Moderate, aggressive |

## 5. Conclusions

Understanding driving behaviors provides an option of application in automated individual driving, which is necessary to promote the experience for the automated vehicle journey. The research focuses on proposing a modified supervised LDA model on latent driving styles for driving individualization. Firstly, we encode the continuous vehicle motion data and traffic context data into discrete driving behavior series. Then, the driving styles of drivers are described as a mixture driving style distribution for individualization driving of drivers, which are based on the supervised labeled LDA model and given prior knowledge. Finally, a case study is carried out, and experimental results show that the proposed model uncovers latent driving styles effectively, showing good agreement to real situations. In addition, the results provide information about the proportion of driving style of each driver and the distribution of driving behaviors for each driving style, which can be implemented in automated driving for individualization driving.

Compared with our previous study, we proposed a novel encoding model for introducing driving context data combining with driving motion data to adapt the bag-of-words model. In addition, instead of unsupervised topic generalization, supervised hyperparameters are introduced to restrict the driver-driving style distribution and driving style-driving behavior distribution. We plan to further optimize our words encoding methods that utilize these continuous data more effectively and improve the performance of the proposed model. Using more driving context data for samples in labeled LDA may lower the perplexity of interpreting driving behaviors and provide a better latent driving styles understanding. The co-occurrence relation between driving behaviors and traffic context is revealed naturally from labeled LDA with prior knowledge.

## Data Availability

The data used are the SPMD (Safety Pilot Model Deployment) data [28]. This model is conducted by UMTRI (the University of Michigan Transportation Research Institute) in Ann Arbor, Michigan. For further dataset access, please visit https://www.its-rde.net/.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported in part by the National Key R&D Program of China under grant 2018YFB1600600; in part by the National Natural Science Foundation of China under grants 61703319, 71702066, 51775396, and U1764262; in part

by the Major Project of Technological Innovation of Hubei Province under grant 2017CFA008; and in part by the Fundamental Research Funds for the Central Universities under grant WUT:2021VI042.

## References

[1] J. Rios-Torres, J. Liu, and A. Khattak, "Fuel consumption for various driving styles in conventional and hybrid electric vehicles: integrating driving cycle predictions with fuel consumption optimization," *International Journal of Sustainable Transportation*, vol. 13, no. 2, pp. 123–137, 2019.

[2] O. Taubman-Ben-Ari, M. Mikulincer, and O. Gillath, "The multidimensional driving style inventory-scale construct and validation," *Accident Analysis & Prevention*, vol. 36, no. 3, pp. 323–332, 2004.

[3] J. E. Meseguer, C. T. Calafate, and J. C. Cano, "On the correlation between heart rate and driving style in real driving scenarios," *Mobile Networks and Applications*, vol. 23, no. 1, pp. 128–135, 2018.

[4] Z. Chen, H. Cai, Y. Zhang et al., "A novel sparse representation model for pedestrian abnormal trajectory understanding," *Expert Systems with Applications*, vol. 138, Article ID 112753, 2019.

[5] Z. Deng, D. Chu, C. Wu, Y. He, and J. Cui, "Curve safe speed model considering driving style based on driver behaviour questionnaire," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 65, pp. 536–547, 2019.

[6] C. M. Martinez, M. Heucke, F. Y. Wang, B. Gao, and D. Cao, "Driving style recognition for intelligent vehicle control and advanced driver assistance: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 666–676, 2018.

[7] D. Ramage, D. Hall, R. Nallapati, C. D. Manning, and L. D. A. "Labeled, "A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, vol. 1, no. 1, pp. 248–256, Association for Computational Linguistics, Singapore, 2009.

[8] Y. Zhang, W. Chen, H. Zha, and X. Gu, "A time-topic coupled LDA model for IPTV user behaviors," *IEEE Transactions on Broadcasting*, vol. 61, no. 1, pp. 56–65, 2015.

[9] P. Duckworth, M. Alomari, J. Charles, D. C. Hogg, and A. G. Cohn, "Latent dirichlet allocation for unsupervised activity analysis on an autonomous mobile robot," in *Proceedings of the Thirty-First AAAI Conference On Artificial Intelligence*, pp. 3819–3826, San Francisco, CA, USA, 2017.

[10] G. Qi, Y. Du, J. Wu, and M. Xu, "Leveraging longitudinal driving behaviour data with data mining techniques for driving style analysis," *IET Intelligent Transport Systems*, vol. 9, no. 8, pp. 792–801, 2015.

[11] Z. Chen, Y. Zhang, C. Wu, and B. Ran, "Understanding individualization driving states via latent Dirichlet allocation model," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 2, pp. 41–53, 2019.

[12] Z. Constantinescu, C. Marinoiu, and M. Vladoiu, "Driving style analysis using data mining techniques," *International Journal of Computers Communications & Control*, vol. 5, no. 5, p. 654, 2010.

[13] M. Van Ly, S. Martin, and M. M. Trivedi, "Driver classification and driving style recognition using inertial sensors," in *Proceedings of the 2013 IEEE Intelligent Vehicles Symposium*, pp. 1040–1045, IEEE, Gold Coast, Australia, 2013.

[14] C. Lv, X. Hu, A. Sangiovanni-Vincentelli, Y. Li, C. M. Martinez, and D. Cao, "Driving-style-based codesign optimization of an automated electric vehicle: a cyber-physical system approach," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 4, pp. 2965–2975, 2018.

[15] D. Chu, Z. Deng, Y. He, C. Wu, C. Sun, and Z. Lu, "Curve speed model for driver assistance based on driving style classification," *IET Intelligent Transport Systems*, vol. 11, no. 8, pp. 501–510, 2017.

[16] X. Zhang, Y. Huang, K. Guo, and W. Li, "Driving style classification for vehicle-following with unlabeled naturalistic driving data," in *Proceedings of the 2019 IEEE Vehicle Power And Propulsion Conference*, pp. 1–5, IEEE, Hanoi, Vietnam, 2019.

[17] W. Liu, K. Deng, X. Zhang et al., "A semi-supervised tri-CatBoost method for driving style recognition," *Symmetry*, vol. 12, no. 3, p. 336, 2020.

[18] C. Deng, C. Wu, N. Lyu, and Z. Huang, "Driving style recognition method using braking characteristics based on hidden Markov model," *PLoS One*, vol. 12, no. 8, 2018.

[19] B. Sun, W. Deng, J. Wu, Y. Li, B. Zhu, and L. Wu, "Research on the classification and identification of drivers driving style," in *Proceedings of the 2017 10th International Symposium on Computational Intelligence and Design*, vol. 1, pp. 28–32, Hangzhou, China, 2017.

[20] W. Wang, J. Xi, and D. Zhao, "Driving style analysis using primitive driving patterns with Bayesian nonparametric approaches," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 2986–2998, 2018.

[21] Y. L. Murphey, R. Milton, and L. Kiliaris, "Driver's style classification using jerk analysis," in *Proceedings of the IEEE Workshop On Computational Intelligence In Vehicles And Vehicular Systems*, pp. 23–28, IEEE, Nashville, TN, USA, 2009.

[22] W. Wang, J. Xi, A. Chong, and L. Li, "Driving style classification using a semisupervised support vector machine," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 5, pp. 650–660, 2017.

[23] M. Ishibashi, M. Okuwa, S. I. Doi, and M. Akamatsu, "Indices for characterizing driving style and their relevance to car following behavior," in *Proceedings of the Annual Conference on the Society-of-Instrument-and-Control-Engineers*, pp. 1132–1137, IEEE, Iizuka, Japan, 2007.

[24] J. Cordero, J. Aguilar, K. Aguilar, D. Chávez, and E. Puerto, "Recognition of the driving style in vehicle drivers," *Sensors*, vol. 20, no. 9, p. 2597, 2020.

[25] S. Chen and Y. Tian, "Evaluating Effectiveness of Latent Dirichlet Allocation Model for Scene Classification," in *Proceedings of the 2011 20th Annual Wireless And Optical Communications Conference (WOCC)*, pp. 1–6, IEEE, Newark, NJ, USA, 2011.

[26] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-link LDA: joint models of topic and author community," in *Proceedings of the 26th Annual International Conference on Machine Learning*, p. 665—672, Montreal, Canada, 2009.

[27] T. Bando, K. Takenaka, S. Nagasaka, and T. Taniguchi, "Unsupervised drive topic finding from driving behavioral data," in *Proceedings of the 2013 IEEE Intelligent Vehicles Symposium (IV)*, pp. 177–182, IEEE, Gold Coast, Australia, 2013.

[28] Safety Pilot Model Deployment and the Members of the Test Conductor Team, *Safety Pilot Model Deployment – Sample Data*, U.S. Department of Transportations USDOT, Washington, DC, 2014.

[29] J. Platt, "Sequential minimal optimization: a fast algorithm for training support vector machines," pp. 1–21, 1998, Technical Report MSR-TR-98-14.
[30] M. E. Maron, "Automatic indexing: an experimental inquiry," *Journal of the ACM*, vol. 8, no. 3, pp. 404–417, 1961.
[31] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.