Received: 1 March 2022 | Revised: 8 April 2022 | Accepted: 17 April 2022

IET Communications

DOI: 10.1049/cmu2.12405

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Real-time driving style classification based on short-term observations

Xinhu Zheng[1] | **Pengtao Yang**[2] | **Dongliang Duan**[3] | **Xiang Cheng**[2] | **Liuqing Yang**[1]

[1]Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, Minnesota, USA (Email: qingqing@umn.edu)

[2]State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics, Peking University, Beijing, China

[3]Department of Electrical and Computer Engineering, University of Wyoming, Laramie, Wyoming, USA

Correspondence
Xiang Cheng, State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics, Peking University, Beijing, China.
Email: xiangcheng.86@googlemail.com

Funding information
Ministry National Key Research and Development Project, Grant/Award Number: 2020AAA0108101; National Natural Science Foundation of China, Grant/Award Number: 62125101; National Science Foundation, Grant/Award Numbers: CNS-1932413, CNS-1932139

Abstract

Vehicle behaviour prediction provides important information for decision-making in modern intelligent transportation systems. People with different driving styles have considerably different driving behaviours and hence exhibit different behaviour tendency. However, most existing prediction methods do not consider the different tendencies in driving styles and apply the same model to all vehicles. Furthermore, most of the existing driver classification methods rely on offline learning that requires a long observation of driving history and hence are not suitable for real-time driving behaviour analysis. To facilitate personalised models that can potentially improve vehicle behaviour prediction, the authors propose an algorithm that classifies drivers into different driving styles. The algorithm only requires data from a short observation window and it is more applicable for real-time online applications compared with existing methods that require a long term observation. Experiment results demonstrate that the proposed algorithm can achieve consistent classification results and provide intuitive interpretation and statistical characteristics of different driving styles, which can be further used for vehicle behaviour prediction.

1 | INTRODUCTION

In recent years, humankind has made remarkable progress in the transportation system. Specifically, advanced sensors are installed at both the infrastructure and the vehicles to provide improved situational awareness and facilitate machine intelligence during the decision-making while driving, realising the Intelligent Transportation Systems (ITS). Among the various issues, the prediction of the vehicle motion and behaviour is a crucial topic, since it provides critical information to the decision-making of both individual vehicles and the transportation control centre. As a result, there have been many recent studies on vehicle trajectory tracking and behaviour prediction (see, e.g., [1–17]) , and among the studies on behaviour prediction, there are three typical research directions: motion model-based prediction (e.g. [3–5]), manoeuvre-based prediction (e.g. [6–10]) and interaction-aware prediction (e.g. [11–13]).

However, none of them has taken into account the impact of different driving styles in the prediction. Specifically, all drivers are treated identically and the same model is applied to all vehicles to conduct behaviour analysis and prediction. In practice, drivers with different driving styles (e.g. aggressive or conservative) or under different driving conditions (e.g. normal, rush, or even drunken) could lead to considerably different tendency in their driving behaviours. To achieve better behaviour prediction, personalised models for individual drivers would be a better alternative, which can be established if the driving styles or conditions could be obtained in advance.

In the meantime, there are a large amount of works on driving style analysis and recognition (see, e.g., [18–36]). Most works applied data mining techniques to classify different driving styles, typically clustering [18, 24, 26, 28, 30]. Some early works rely on fuzzy logic to conduct the driving style recognition [21, 23]. Different types of probabilistic models are also very popular among researchers for the classification of driving behaviour [25, 27, 29, 31]. With the booming of machine learning, many researchers tend to use learning techniques, for instance, Support Vector Machine [37] to conduct driving style classification [28, 30]. Moreover, domain transformation from time to frequency has also been applied in order to improve the classification accuracy [20] or facilitate the data fusion from different sources [19, 22 36]. The early work [18] that applied data mining techniques made use of principal component analysis (PCA) [38] and hierarchical clustering algorithm (HCA) [39] to deal with the driving data. However, this work [18] is only applicable to very limited pre-selected driving parameters or features. Specifically, the parameters covered here are limited to speed, acceleration and mechanical work of the vehicle. There are no position-related parameters that have been taken into account. Although the authors did mention this work can be significantly improved by taking other factors into consideration, no further updates on the work have been published so far. In a recent work [24], more features were included, and the partitioning around medoids (PAM) [40] and K-means [41] clustering algorithms were adopted to classify drivers into multiple groups with different driving profiles. However, this work requires considerably long observations on historical data of vehicles and drivers. It might not be suitable for online real-time driving behaviour analysis. Specifically, it might be inapplicable for traffic scenarios with complex dynamics. In another recent work, Wang *et al.* [28] proposed a modified semi-supervised SVM in order to mitigate the burden of data labelling while maintaining the performance of traditional supervised SVM approaches. However, they only focus on the longitudinal driving behaviour with two features, namely the vehicle speed and the throttle opening. In addition, during the data pre-processing, to make labelling work easier, K-means clustering method was introduced.

In summary, all above works are based on some pre-defined categories and the rules to determine the different categories could vary significantly for different system setups (e.g. heavy traffic vs. light traffic, or day/night). Or they are data hungry to reach certain accuracy while data may not be available in practice. None of the existing works discussed above has managed to propose an algorithm to achieve consistent classification results with a short-term observation window that would be suitable for real-time applications in any practical system setup. Instead, we attempts to classify the vehicle considering short observation window in order to provide new features for subsequent work, so that it can use a personalised model for prediction to achieve more precise result without worrying about the data amount available. Moreover, by introducing the driving style and the corresponding vehicle trajectories, it can also facilitate the vehicular communication channel modelling and estimation, or contribute to the dynamic resource allocation [42–46], especially in trendy Vehicle-to-everything (V2X) scenarios [45] and the future of beyond 5G and 6G communication systems for vehicular communications [46].

Therefore, an integrated algorithm which is ready for real-time applications is proposed, which can classify driver styles by their driving behaviours, based on the sensor data related to vehicle motions such as vehicle's position, velocity, and acceleration. Technically, we extract features of the vehicle motion data via data filtering, domain transformation and dimensionality reduction, and then apply the unsupervised learning technique on a historic dataset as training set to cluster the drivers into several different categories with different driving styles. During this training process, the policies for classification can be established to classify any newly observed vehicle. The proposed algorithm is tested on an open-source dataset (https://data.transportation.gov/Automobiles/Next-Generation-Simulation-NGSIM-Vehicle-Trajector/8ect-6jqj). Results show that the proposed algorithm can successfully classify drivers into different categories of different driving styles with reasonable interpretation. Moreover, the classification can be conducted based on data with short observation windows, which makes it feasible for online real-time applications. In addition, the consistency of the proposed algorithm is also validated. With the proposed driving style classification, a new dimension can be added into existing vehicle behaviour prediction algorithms for better prediction and analysis of vehicle motion.

The rest of this paper is organised as follows. In Section 2, an integrated algorithm for vehicle driving style classification is proposed. Section 3 describes the NGSIM dataset and presents the implementation of the proposed algorithm to the dataset. The interpretation of results and performance evaluation are shown in Section 4. Finally, some concluding remarks are given in Section 5.

2 | THE PROPOSED ALGORITHM

In this paper, we aim to classify the drivers into different driving style categories according to the multi-modal and multi-dimensional sensor data of the vehicle motion. The sensor data could be either provided by on-board sensors of vehicles or the road side units (RSUs) in the ITS. Given the heterogeneity nature in the sensor data, it is challenging to develop a physical-based model for driving behaviours. Therefore, we propose a data-driven approach which is composed of two phases for behaviour classification, namely the training and inference phases. In the training phase, some historical data in a particular transportation system (e.g. a segment of a highway) are collected and an unsupervised learning approach [47] is applied to cluster the data into different clusters, which corresponds to the driving style categories. At the same time, the resultant clustering policy will be established and applied to classify any newly observed vehicle in the same system into specific driving style in the inference phase. For both the training and inference phase, in order to accelerate the clustering process, the high-dimensional sensor data is pre-processed and dimensionally reduced such that the features that are closely related to driving behaviour will be

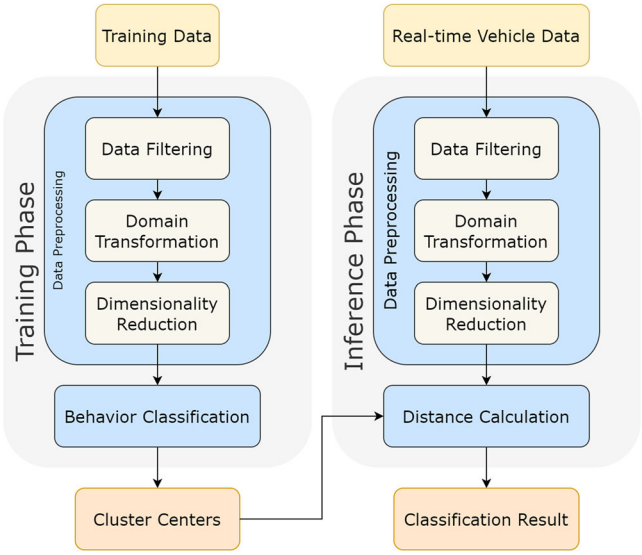


FIGURE 1 Algorithm flow chart

extracted. A flow chart of the proposed algorithm is shown in Figure 1.

2.1 | Data pre-processing

The data pre-processing is divided into two steps: data filtering to select the associated features and domain transformation to generate frequency domain features.

2.1.1 | Data filtering

In general, there might be a plenty of sensor data available that cover many different aspects of vehicle motion. In order to conduct driving behaviour classification, one should filter the enormous sensor data by selecting only features that are either related to the manoeuvre of the vehicle (such as position, speed, and acceleration) or to its temporal and spatial relationship with other vehicles (such as the relative position to the preceding vehicle).

Assuming that n-dimensional related sensor data are available

$$\boldsymbol{x}_i = \left[x^1, x^2, \ldots, x^n\right]^T, \tag{1}$$

Assuming that k-dimensional related sensor data are selected where $k \leq n$, a vehicle's feature at the ith time slot can be represented by an k-dimensional vector $\boldsymbol{x}_i$ as

$$\hat{\boldsymbol{x}}_i = \left[x^1, x^2, \ldots, x^k\right]^T, \tag{2}$$

where each component of the vector corresponds to one feature selected. The driving behaviour is usually characterised by the temporal development of vehicle motion. Assume that there an observation window of N, therefore the sensor data can be organised into a $k \times N$ matrix

$$\left[\hat{\boldsymbol{x}}_1\ \hat{\boldsymbol{x}}_2 \ldots \hat{\boldsymbol{x}}_N\right] = \begin{bmatrix} x_1^1 & \ldots & x_N^1 \\ x_1^2 & \ldots & x_N^2 \\ \vdots & \ddots & \vdots \\ x_1^k & \ldots & x_N^k \end{bmatrix} = \begin{bmatrix} \mathbf{x}_T^1 \\ \mathbf{x}_T^2 \\ \vdots \\ \mathbf{x}_T^k \end{bmatrix}. \tag{3}$$

2.1.2 | Domain transformation

Instead of working on the temporal data directly, the data is transformed into frequency domain before conducting the clustering. There are two reasons for domain transformation [20]. First, the frequency domain can contain more information about the driving behaviour with limited data. For instance, the speed of the vehicle in the time domain describes the instantaneous value of the velocity over a period of time, while in the frequency domain, it can describe the fluctuation of the vehicle speed. Hence, the frequency domain velocity can indicate whether the vehicle is driving smoothly or not directly, while in time domain, the steadiness may not be available given a limited observation window. Second, the driver behaviour is usually insensitive to the absolute time stamp of its actions. For instance, if a vehicle changes the lane three times within 1 min, no matter when those lane-change actions happen, this driver would be considered as an aggressive driver. Transforming the data into frequency domain would further filter the time stamps that are not relevant in deciding the driving behaviour.

Performing the Fourier transform of each of the k-dimensional time domain features, one can obtain the frequency domain features as follows:

$$\begin{bmatrix} \mathbf{x}_T^1 \\ \mathbf{x}_T^2 \\ \ldots \\ \mathbf{x}_T^k \end{bmatrix} \overset{\mathcal{F}(\cdot)}{=\!=} \begin{bmatrix} \mathbf{x}_F^1 \\ \mathbf{x}_F^2 \\ \ldots \\ \mathbf{x}_F^k \end{bmatrix} = \boldsymbol{X}_{k\times N}, \tag{4}$$

in which $\mathcal{F}(\cdot)$ represents the row-wise Fourier transform.

2.2 | Dimension reduction

Instead of using the raw frequency domain data $\boldsymbol{X}_{k\times N}$ we obtained in Section 2.1.2, we conduct dimension reduction of the data before the clustering. There are several reasons for doing this. First of all, during the dimensional reduction, the correlation among different components of the data can be studied and the major structure in the data would be revealed. This greatly facilitates the clustering process. Second,

the noise would usually be reduced during the dimension reduction. Third, the computational complexity for the training process would be greatly reduced with the dimensionally reduced data.

Principal component analysis (PCA) is one of the most widely used data dimensionality reduction algorithms. The main idea of PCA is to sequentially find a set of mutually orthogonal coordinate axes from the original space to map a higher k-dimensional features to a lower r-dimensional vector. The selection of the new coordinate axes is closely related to the data itself, wherein, the first new coordinate axis selection is the direction with the largest variance in the original data, and the second new coordinate axis selection is the plane orthogonal to the first coordinate axis. The value of the resultant dimensionality r is selected according to the variance ratio of each dimension. Let the variance ratio of the dimension with the ith largest ratio be v_i, then

$$r = \arg\min_{r} \sum_{i=1}^{r} v_i \geq V_{th}, \tag{5}$$

in which V_{th} is the variance ratio threshold and can be treated as a hyper-parameter in the proposed algorithm. When V_{th} is larger, r is larger, hence more dimensions will be retained after the dimension reduction, and vice versa.

2.3 | Driving behaviour classification by clustering

After data pre-processing in Section 2.1 and feature dimensionality reduction in Section 2.2, the K-means algorithm which is an unsupervised learning algorithm, is then applied to cluster the frequency domain features.

In order to determine the value of K, Calinski–Harabasz score (CH score)[48] is used to evaluate the results of the clustering. In clustering, the internal-cluster variance is defined as

$$S_W(K) = \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in C_i} \|\boldsymbol{x} - \boldsymbol{m}_i\|^2, \tag{6}$$

in which $\boldsymbol{C}_i$ is the set of all points in the ith cluster, $\boldsymbol{m}_i$ is the centre point of the date in the ith cluster and inter-cluster variance is defined as

$$S_B(K) = \sum_{i=1}^{K} n_i \|\boldsymbol{m} - \boldsymbol{m}_i\|^2, \tag{7}$$

in which $\boldsymbol{m}$ is the center point of all data, $\boldsymbol{m}_i$ is the centre point of the date in the ith cluster and n_i is the number of data in the ith cluster. Then the CH Score is defined as

$$CH(K) = \frac{S_B(k)/(k-1)}{S_W(k)/(n-k)}. \tag{8}$$

It can be seen that the larger the CH Score is, the more close the data are to each other within the same category, and the more dispersed the data between different categories. In other words, the clustering performance is better.

2.4 | Inference phase

During the training phase of this unsupervised learning approach described in Sections 2.1, 2.2, 2.3, drivers in a particular system are classified into different categories with different driving styles. When a new driver in the same system is observed, one can apply the rules obtained during the clustering to classify this driver.

Specifically, with the available training data, suppose that M different categories are generated. For any new vehicle, its distance to the M cluster centres in the feature domain can be calculated by the data in a short time, and one can calculate the probability that this vehicle belongs to each category by the normalised distance function. With the same data processing method and the dimensionality reduction procedure described above, the feature of the observed vehicle is obtained, and the distance between the vehicle feature and the ith centre in the feature space can be denoted as d_i. According to the central limit theorem, it is assumed that the data follows Gaussian distribution. Then the probability density that the vehicle belongs to the ith class is

$$f_i(d_i) = \mathcal{A}\, e^{-\frac{{d_i}^2}{B}}, \tag{9}$$

where $\mathcal{A}$ and B are constant. Then the normalisation is applied to find the probability for the ith class

$$P_i(d_i) = \frac{f_i(d_i)}{\sum_{j=1}^{M} f_j(d_j)} = \frac{e^{-{d_i}^2}}{\sum_{j=1}^{M} e^{-{d_j}^2}}, \tag{10}$$

that is,

$$P_i(d_i | d_1, d_2, \ldots, d_M) = \text{softmax}({d_i}^2 | {d_1}^2, {d_2}^2, \ldots, {d_M}^2). \tag{11}$$

3 | EXPERIMENTS

We use some real-world data to verify the proposed algorithm with detailed information given in the following subsection.

3.1 | Dataset description

The data used in this paper are provided by Federal Highway Administration's NGSIM project [49]. There are two typical datasets available: U.S. Highway 101 and Interstate 80 in California, and both datasets provide the vehicle trajectory data extracted by video cameras. We worked on the U.S. Highway 101 dataset, where vehicle trajectory data are

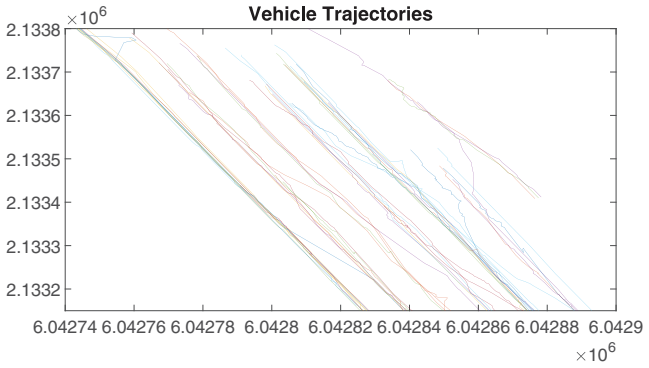


FIGURE 2 50 randomly selected vehicle trajectories in the dataset

TABLE 1 Raw features used in behaviour classification

Name	Description
Local_X	Lateral coordinate of the front centre of the vehicle in feet with respect to the left-most edge of the section in the direction of travel.
Local_Y	Longitudinal coordinate of the front centre of the vehicle in feet with respect to the left-most edge of the section in the direction of travel.
Global_X	X Coordinate of the front centre of the vehicle in feet based on CA State Plane III in NAD83.
Global_Y	Y Coordinate of the front centre of the vehicle in feet based on CA State Plane III in NAD83.
V_vel	Instantaneous velocity of vehicle in feet/second.
V_acc	Instantaneous acceleration of vehicle in feet/second square.
Space_headway	Space Headway in feet. Spacing provides the distance between the front centre of a vehicle to the front center of the preceding vehicle.
Time_headway	Time Headway in seconds. Time Headway provides the time to travel from the front center of a vehicle (at the speed of the vehicle) to the front center of the preceding vehicle.

collected from a section of U.S. Highway 101, Los Angeles, CA. The length of this section is 640 m, and it consists of five lanes, one auxiliary lane and two ramps. The dataset records vehicle motion information for 45 min from 7:50 AM to 8:35 AM on 15 June 2005. Each piece of data is recorded at 0.1 s intervals, that is the sampling frequency is 10 Hz (https://data.transportation.gov/Automobiles/Next-Generation-Simulation-NGSIM-Vehicle-Trajector/8ect-6jqj). An illustration of 50 randomly selected vehicle trajectories in the dataset are shown in Figure 2 for description of the selected highway section.

The data in the dataset contains 25 features such as *Vehicle_ID, Frame_ID, Local_X, Local_Y, Global_X*, and *Global_Y*, which describe the absolute position, relative position, relationship with the preceding vehicle, speed, and acceleration. *Vehicle_ID* can distinguish which vehicle the data belongs to. In total, there are 3233 vehicles and 118,505,266 data frames in the dataset, and the number of data frames of each vehicle ranges from 217 to 9834 depending on how long the vehicle stays within the area of interest.

TABLE 2 Ratio of variance after PCA

Domain	Dimension order				
Time domain	88.22%	11.77%	0.0011%	0.00045%	0.00020%
Frequency domain	99.77%	0.088%	0.050%	0.040%	0.028%

3.2 | Algorithm implementation

In this section, an example to show how the algorithm implementation on the NGSIM dataset is presented.

3.2.1 | Data preprocessing

The data in NGSIM dataset contains 25 dimensions. Some dimensions are used to identifying the owner of the data, such as *Vehicle_ID*, and *Frame_ID*, and the others describe some information related to vehicle motion. Among many dimensions, there are eight dimensions of the dataset been selected as the features of interests, as shown in Table 1 to conduct analysis on driver behaviours.

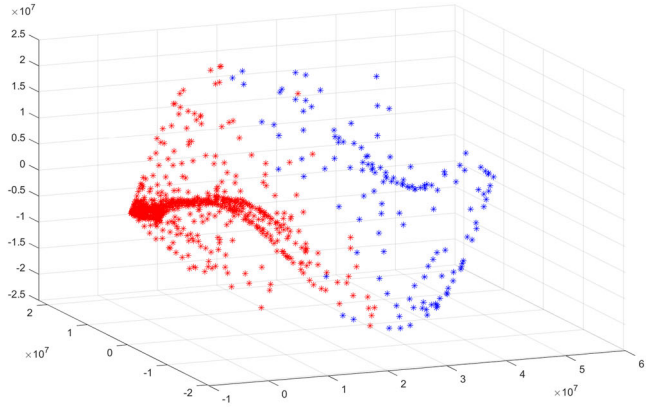
As mentioned previously, a vehicle's feature at the ith frame can be represented by an 8-dimensional vector $\boldsymbol{x}_i$

$$\boldsymbol{x}_i = \left[x_{Local\_X},\ x_{Local\_Y},\ x_{Global\_X},\ x_{Global\_Y},\ x_{V\_vel},\ x_{V\_acc},\ x_{Space\_headway},\ x_{Time\_headway}\right]^T. \tag{12}$$

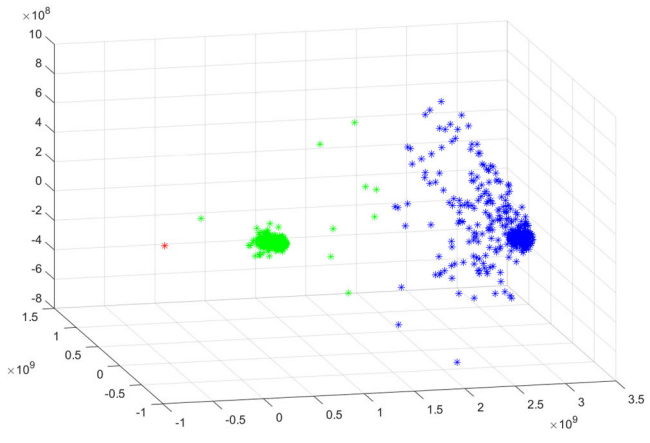
Taking an observation window with N frames, then a time-domain feature matrix for each vehicle is obtained as $\left[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \ldots\ \boldsymbol{x}_N\right]$. Then, as described in 2.1.2, frequency domain feature matrix is obtained after taking the Fourier transform of each row

$$\begin{bmatrix} \mathbf{x}^T_{Local\_X} \\ \mathbf{x}^T_{Local\_Y} \\ \mathbf{x}^T_{Global\_X} \\ \mathbf{x}^T_{Global\_Y} \\ \mathbf{x}^T_{V\_vel} \\ \mathbf{x}^T_{V\_acc} \\ \mathbf{x}^T_{Space\_headway} \\ \mathbf{x}^T_{Time\_headway} \end{bmatrix} \overset{\mathcal{F}(\cdot)}{=\!=} \begin{bmatrix} \mathbf{x}^F_{Local\_X} \\ \mathbf{x}^F_{Local\_Y} \\ \mathbf{x}^F_{Global\_X} \\ \mathbf{x}^F_{Global\_Y} \\ \mathbf{x}^F_{V\_vel} \\ \mathbf{x}^F_{V\_acc} \\ \mathbf{x}^F_{Space\_headway} \\ \mathbf{x}^F_{Time\_headway} \end{bmatrix} = \boldsymbol{X}_{8\times N}. \tag{13}$$

For real-time applications, and considering the selected highway section, the observation window is set to 200 frames ($N = 200$),



(a) Time-domain.



(b) Single principal component.

FIGURE 3 Clustering using different features

that is the data with a continuous 20 s record. This indicates that the proposed algorithm can determine the driving style of a vehicle with a very short observation window. Note that, not all of the 3233 vehicles in the original dataset contains 200 frames of data, and 2356 vehicles were selected after removing bad data.

3.2.2 | Dimension reduction

PCA is applied to this dataset to reduce and select the dimensionality of the frequency domain feature. The variance ratio of the first five dimensions after their dimensionality reduction is shown in Table 2.

It can be seen that the variance ratio of the first dimension in frequency domain exceeds 99% and the summation of first two dimension in time domain exceeds 99.99%. The summation of first three dimensions of time domain and frequency domain both exceeds 99.99%. As shown in Figure 3(a), the time domain clustering result does not show clear boundary between clusters and the points within each cluster are dispersed. The single frequency-domain principal component clustering result shown in Figure 3(b), the number of clusters is still 3 as the proposed scheme. However, the points within each cluster are not quite concentrated as compared with the result shown in

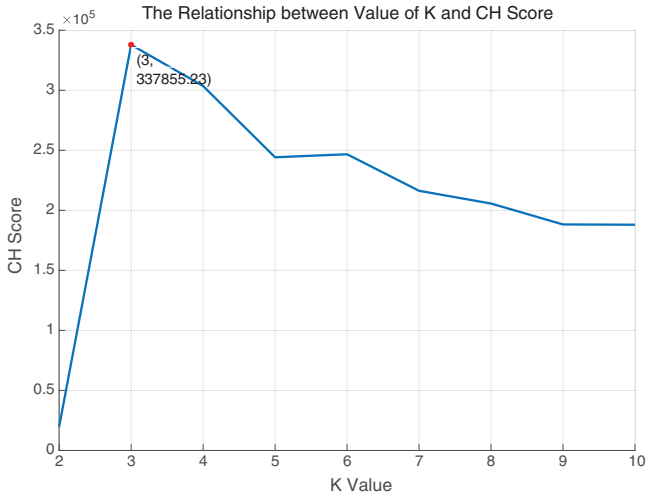


FIGURE 4 The relationship between value of K and CH score

Figure 5 which uses three principal components for flustering. Therefore, the first three principal components are selected as features for the following clustering procedure.

3.2.3 | Driving behaviour classification

The final step of the algorithm is to cluster the extracted features. As described in Section 2.3, the CH-Score index is used to judge the clustering effect, so some comparative experiments were performed to obtain the best K value. It can be seen from Figure 4 that the clustering performance is best when $K = 3$. Hence, the value of K is set to 3 for the selected dataset.

4 | RESULT AND ANALYSIS

In this section, statistical results of clustering are presented, and then the characteristics of the drive styles in different categories are analysed. Finally, we introduce how the labels obtained by clustering can be used to classify new observed vehicles.

4.1 | Clustering results

After the dimensionality of the frequency domain feature is reduced, the clustering results with highest CH-score are shown in Figure 5, and the statistical characteristics of three cluster are shown in Table 3. In the following, we will refer to the categories of red, green and blue as category 0, category 1 and category 2, respectively.

4.2 | Characteristic in different categories

By analysing the statistical characteristics of each type of driving style (as shown in Table 3), the characteristics of each category's vehicles driving style can be interpreted as follows:

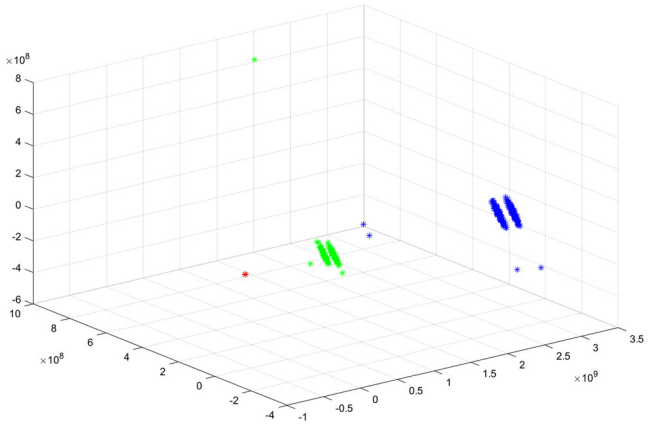


FIGURE 5 Clustering result visualisation: The number of category 0 (red) is 1767, the number of category 1 (green) is 294 and the number of category 2 (blue) is 295

TABLE 3 Statistical characteristics of clustering

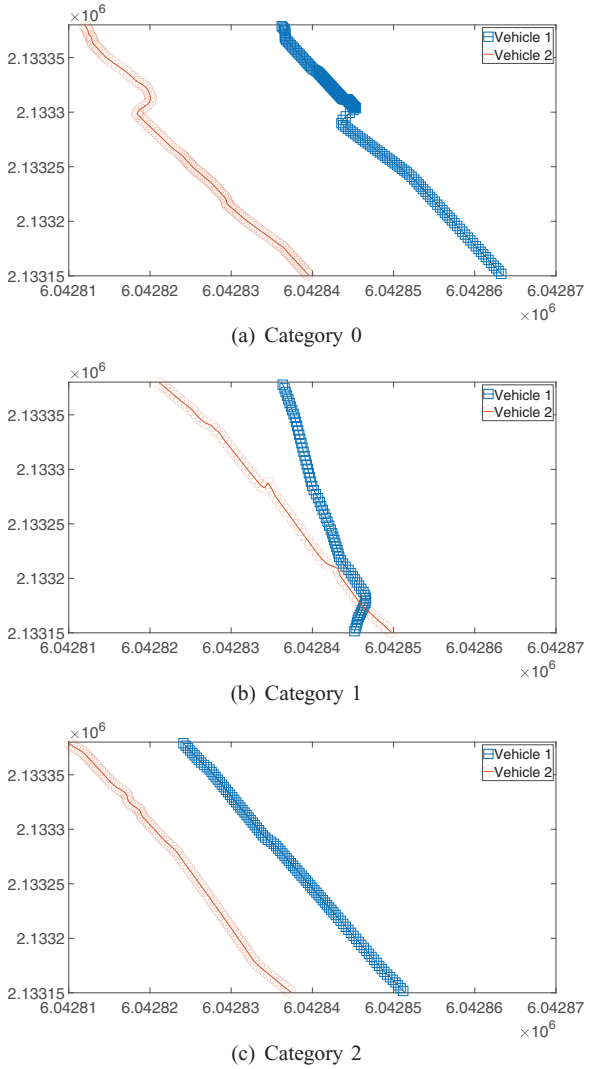
Category	Attribute	Mean	Variance
0	*Local_X*	35.64 ft	5.29 ft^2
	V_vel	27.36 ft/s	47.12 ft^2/s^2
	V_acc	-0.11 ft/s^2	29.77 ft^2/s^4
	Space_headway	67.46 ft	728.23 ft^2
	Time_headway	144.83 s	1,103,061 s^2
1	*Local_X*	36.03 ft	6.22 ft^2
	V_vel	29.40 ft/s	48.56 ft^2/s^2
	V_acc	-0.11 ft/s^2	27.56 ft^2/s^4
	Space_headway	72.70 ft	947.14 ft^2
	Time_headway	58.96 s	440,478 s^2
2	*Local_X*	31.26 ft	4.25 ft^2
	V_vel	31.85 ft/s	48.90 ft^2/s^2
	V_acc	-0.07 ft/s^2	30.23 ft^2/s^4
	Space_headway	75.93 ft	2319.85 ft^2
	Time_headway	127.01 s	819,074 s^2

4.2.1 | Category 0 'conservative drivers'

The average speed (27.36) is the slowest of the three categories. It can be seen from the exemplary trajectory that the speed is slower when making turns. At the same time, the time taken to reach the position of preceding car is significantly longer than the other two categories, which means that the drivers in this category are more conservative to prevent rear-end collision.

4.2.2 | Category 1 'aggressive drivers'

The average speed (29.04) is faster, second only to category 3. The time taken to reach the position of the preceding car is significantly less than the other two categories, indicating that they tend to follow closer to the preceding cars. In addition, the average variance of the distance from the left side of the road is the largest, indicating that the drivers tend to take lane changes much more frequently.



(a) Category 0

(b) Category 1

(c) Category 2

FIGURE 6 Typical trajectory of each category

4.2.3 | Category 2 'experienced drivers'

The average speed (31.85) is the fastest, but the driving is relatively stable, because the absolute value of the average acceleration (-0.07, the other two categories are -0.11) is the smallest, indicating that the brakes are the least used. The average variance of the distance from the left side of the road is the smallest, indicating that the driver is good at selecting routes and change lanes less often. And the average distance from preceding car is the largest, but the time taken to reach the position of preceding car is moderate, indicating that the driver can better control the distance from the preceding car and the speed while following the preceding car.

The example trajectories as shown in Figure 6 can be analysed for intuitive interpretations, where all trajectories are cho-

sen from the same fragment of X-axis from 6,042,810 to 6,042,870 and Y-axis from 2,133,150 to 2,133,380. It can be further confirmed that the trajectories match the driving styles from conservative to aggressive and experienced. Since each data point represents the position of the vehicle at each time slot, denser points indicate slower speed. In Figure 6(a), the driver slows down significantly before performing the lane change. In Figure 6(b), one of the vehicles, marked by blue square, does not keep the line, and another vehicle, marked as brown circle, does not slow down when performing the lane change. The experienced drivers, shown in Figure 6(c), keep within a relatively straight lane and maintain a steady speed, that is the data point density does not change significantly.

It should be noted that the three categories obtained here are only applicable to this particular dataset. For other transportation systems such as a different road section, there might be drivers with different number of driving styles which all have quite different characteristics in their behaviour. However, for any system, similar procedure could be followed to analyse some historical observations on the vehicles in the system and obtain the driving styles among them.

4.3 | Performance for classification

Now for the NGSIM dataset, three cluster centres have been found. Then the probability for the ith class is

$$P_i(d_i) = \frac{f_i(d_i)}{\sum_{j=0}^{2} f_j(d_j)} = \frac{e^{-{d_i}^2}}{\sum_{j=0}^{2} e^{-{d_j}^2}}, \tag{14}$$

that is

$$P_d(r_i|r_0, r_1, r_2) = \text{softmax}({d_i}^2|{d_0}^2, {d_1}^2, {d_2}^2). \tag{15}$$

In the evaluation experiment, the clustering results obtained from all NGSIM data were used for cross-validation and verification of the training and testing results. In order to validate the model and avoid underfitting or overfitting, k-fold cross-validation is applied here [50]. Set $1/k$ portion of the vehicles from the dataset as the testing set, and the rest as training set, where k is an integer and the total number of vehicles is 3233. Note that 2356 vehicles were selected after removing vehicles that contain less than 200 frames of data. The training set is applied to obtain the cluster centres, and the test set is then classified into the clusters. Finally, the classification result of the test set is compared with the clustering results obtained from the whole dataset to obtain the classification accuracy. In order to further justify the consistency of the proposed algorithm, we evaluate the performance of the classifier with the different setting of the random state and max iterations. The default settings are *random_state* = *None* and *max_iter* = 300, and experiments are also conducted with settings of *random_state* from 0 to 10 and *max_iter* increasing to 500 and 1000. The training and testing results are consistent and shown in Table 4. The accuracy is 79% at $k = 3$ and is increased to 87% at $k = 4$. However, the macro average of precision at $k = 4$ is decreased to 0.5 from 0.79 at $k = 3$ because the distinction between category 2 and category 1 is blurred due to the testing and training split settings. The best performance is obtained at $k = 5$, and the proposed algorithm achieves a 92% accuracy, and both precision and recall are higher than previous settings. Any $k > 5$ will result in overfitting which means that the metrics of classification results all equal to 1. This shows that the proposed algorithm can successfully learn and extract important features about driving behaviour, making it possible to classify vehicle driving style based on the vehicle motion information with a short observation window.

TABLE 4 Training and testing results

k-fold	Accuracy	Attribute	Macro avg.	Weighted avg.
3	0.79	*precision*	0.79	0.92
		recall	0.81	0.79
		f1-score	0.74	0.82
4	0.87	*precision*	0.50	0.81
		recall	0.67	0.87
		f1-score	0.56	0.83
5	0.92	*precision*	0.87	0.95
		recall	0.91	0.92
		f1-score	0.87	0.93

5 | CONCLUSIONS AND FUTURE WORK

An algorithm for driving style classification is proposed. The proposed method can successfully classify driving styles with limited data from a short observation window and achieve consistent classification results. The training phase consists of a series of data mining techniques, including data filtering, domain transformation, dimension reduction and clustering. The first three techniques are conducted in order to enable the clustering to work with limited data. In the inference phase, it can classify the driving style based on the sensor data including vehicle motion and position data which collected in a short observation window. Moreover, distinct vehicle trajectories of different driving styles can easily be obtained. It should be noted that the classifier obtained would be only applicable to the particular system where the training data are collected. For any new system, training must be re-conducted to learn the driving styles among the vehicles in that system. In the future, we plan to update the classifier to more generalised dataset and further utilise the categories of driving styles as new feature to build personalised models for better vehicle behaviour and trajectory prediction.

ACKNOWLEDGMENTS

This work was in part supported by the Ministry National Key Research and Development Project under Grant

2020AAA0108101, the National Natural Science Foundation of China under Grants 62125101, and the National Science Foundation under Grants CNS-1932413 and CNS-1932139.

CONFLICT OF INTEREST

The authors do not have a conflict of interest to disclose.

ORCID

Xinhu Zheng https://orcid.org/0000-0002-9898-5543
Dongliang Duan https://orcid.org/0000-0003-1015-2481
Xiang Cheng https://orcid.org/0000-0002-5943-0326

REFERENCES

1. Zhang, Y., Song, B., Du, X., Guizani, M.: Vehicle tracking using surveillance with multimodal data fusion. IEEE Trans. Intell. Transp. Syst. 19, 1–9 (2019)
2. Scheel, A., Reuter, S., Dietmayer, K.: Vehicle tracking using extended object methods: An approach for fusing radar and laser. in 2017 IEEE International Conference on Robotics and Automation (ICRA), (2017)
3. Huang, J., Tan, H.S.: Vehicle future trajectory prediction with a dgps/ins-based positioning system. in American Control Conference, 2006, (2006)
4. Sorstedt, J., Svensson, L., Sandblom, F., Hammarstrand, L.: A new vehicle motion model for improved predictions and situation assessment. IEEE Trans. Intell. Transp. Syst. 12(4), 1209–1219 (2011)
5. Polychronopoulos, A., Tsogas, M., Amditis, A.J., Andreone, L.: Sensor fusion for predicting vehicles' path for collision avoidance systems. IEEE Trans. Intell. Transp. Syst. 8(3), 549–562 (2007)
6. Houenou, A., Bonnifait, P., Cherfaoui, V., Yao, W.: Vehicle trajectory prediction based on motion model and maneuver recognition. IEEE/RSJ International Conference on Intelligent Robots & Systems (2013)
7. Xie, G., Gao, H., Qian, L., Huang, B., Li, K., Wang, J.: Vehicle trajectory prediction by integrating physics- and maneuver-based approaches using interactive multiple models. IEEE Trans. Ind. Electron. 65(7), 5999–6008 (2018)
8. Schreier, M., Willert, V., Adamy, J.: An integrated approach to maneuver-based trajectory prediction and criticality assessment in arbitrary road environments. IEEE Trans. Intell. Transp. Syst. 17, 1–16 (2016)
9. Deo, N., Rangesh, A., Trivedi, M.M.: How would surrounding vehicles move? A unified framework for maneuver classification and motion prediction. IEEE Trans. Intell. Veh. 1–1 (2018)
10. Deo, N., Trivedi, M.M.: Convolutional social pooling for vehicle trajectory prediction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1468–1476 (2018)
11. Schlechtriemen, J., Wirthmueller, F., Wedel, A., Breuel, G., Kuhnert, K.D.: When will it change the lane? A probabilistic regression approach for rarely occurring events. 2015 IEEE Intelligent Vehicles Symposium (IV), (2015)
12. Kafer, E., Hermes, C., Wohler, C., Ritter, H., Kummert, F.: Recognition of situation classes at road intersections. 2010 IEEE International Conference on Robotics and Automation (ICRA), (2010)
13. Lawitzky, A., Althoff, D., Passenberg, C.F., Tanzmeister, G., Buss, M.: Interactive scene prediction for automotive applications. IEEE Intelligent Vehicles Symposium (IV), (2013)
14. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. 34(7), 1409–1422 (2011)
15. Asha, C.S., Narasimhadhan, A.: Adaptive learning rate for visual tracking using correlation filters. Procedia Comput. Sci. 89, 614–622 (2016)
16. Yao, B., Zhang, Q.L.: Optimal control for large-scale descriptor systems with symmetric circulant structure. J. Northeastern Univ. (2003)
17. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. 37(3), 583–596 (2015)
18. Constantinescu, Z., Marinoiu, C., Vladoiu, M.: Driving style analysis using data mining techniques. Int. J. Comput. Commun. Contr. 5(5), 654–663 (2010)
19. Johnson, D.A., Trivedi, M.M.: Driving style recognition using a smartphone as a sensor platform. 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 1609–1615. IEEE, Piscataway, NJ (2011)
20. Daza, I.G., Hernández, N., Bergasa, L.M., Parra, I., Yebes, J.J., Gavilán, M., Quintero, R., Llorca, D.F., Sotelo, M.: Drowsiness monitoring based on driver and driving data fusion. 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 1199–1204. IEEE, Piscataway, NJ (2011)
21. Aljaafreh, A., Alshabatat, N., Al-Din, M.S.N.: Driving style recognition using fuzzy logic. 2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES 2012), pp. 460–463. IEEE, Piscataway, NJ (2012)
22. Van Ly, M., Martin, S., Trivedi, M.M.: Driver classification and driving style recognition using inertial sensors. 2013 IEEE Intelligent Vehicles Symposium (IV), pp. 1040–1045. IEEE, Piscataway, NJ (2013)
23. Dörr, D., Grabengiesser, D., Gauterin, F.: Online driving style recognition using fuzzy logic. 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 1021–1026. IEEE, Piscataway, NJ (2014)
24. Figueredo, G.P., Agrawal, U., Mase, J.M., Mesgarpour, M., Wagner, C., Soria, D., Garibaldi, J.M., Siebers, P.-O., John, R.I.: Identifying heavy goods vehicle driving styles in the united kingdom. IEEE Trans. Intell. Transp. Syst. 20(9), 3324–3336 (2018)
25. Sundbom, M., Falcone, P., Sjöberg, J.: Online driver behavior classification using probabilistic arx models. 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), pp. 1107–1112. IEEE, Piscataway, NJ (2013)
26. Vaitkus, V., Lengvenis, P., Žylius, G.: Driving style classification using long-term accelerometer information. 2014 19th International Conference on Methods and Models in Automation and Robotics (MMAR), pp. 641–644. IEEE, Piscataway, NJ (2014)
27. Filev, D., Lu, J., Tseng, F., Prakah-Asante, K.: Real-time driver characterization during car following using stochastic evolving models. 2011 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1031–1036. IEEE, Piscataway, NJ (2011)
28. Wang, W., Xi, J., Chong, A., Li, L.: Driving style classification using a semisupervised support vector machine. IEEE Trans. Human Mach. Syst. 47(5), 650–660 (2017)
29. Li, G., Li, S.E., Cheng, B., Green, P.: Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities. Trans. Res. Part C: Emerg. Technol. 74, 113–125 (2017)
30. Chandrasiri, N.P., Nawa, K., Ishii, A.: Driving skill classification in curve driving scenes using machine learning. J. Mod. Transp. 24(3), 196–206 (2016)
31. Wang, W., Xi, J., Zhao, D.: Driving style analysis using primitive driving patterns with Bayesian nonparametric approaches. IEEE Trans. Intell. Transp. Syst. 20(8), 2986–2998 (2018)
32. Kohno, T.: Automobile driver fingerprinting. Proc. Privacy Enhancing Technol. 2016(1), 34–50 (2015)
33. Martínez, M., Echanobe, J., del Campo, I.: Driver identification and impostor detection based on driving behavior signals. 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), pp. 372–378. IEEE, Piscataway, NJ (2016)
34. Jafarnejad, S., Castignani, G., Engel, T.: Towards a real-time driver identification mechanism based on driving sensing data. 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp. 1–7. IEEE, Piscataway, NJ (2017)
35. Marchegiani, L., Posner, I.: Long-term driving behaviour modelling for driver identification. 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 913–919. IEEE, Piscataway, NJ (2018)
36. Martinez, C.M., Heucke, M., Wang, F.-Y., Gao, B., Cao, D.: Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. IEEE Trans. Intell. Transp. Syst. 19(3), 666–676 (2017)
37. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. 20(3), 273–297 (1995)

38. Wold, S.: Principal component analysis. Chemo. Intell. Lab. Syst. 2(1), 37–52 (1987)
39. Johnson, S.C.: Hierarchical clustering schemes. Psychometrika 32(3), 241–254 (1967)
40. Kaufmann, L.: Clustering by means of medoids. Proc. Statistical Data Analysis Based on the L1 Norm Conference, Neuchatel, pp. 405–416 (1987)
41. Hartigan, J.A.: Algorithm as 136: A k-means clustering algorithm. J. R. Stat. Soc. 28(1), 100–108 (1979)
42. Zajic, A.G., Stuber, G.L.: Space-time correlated mobile-to-mobile channels: Modelling and simulation. IEEE Trans. Veh. Technol. 57(2), 715–726 (2008)
43. Matolak, D.W.: Channel modeling for vehicle-to-vehicle communications. IEEE Commun. Mag. 46(5), 76–83 (2008)
44. Ding, J.-W., Wang, C.-F., Meng, F.-H., Wu, T.-Y.: Real-time vehicle route guidance using vehicle-to-vehicle communication. IET Commun. 4(7), 870–883 (2010)
45. Chen, S., Hu, J., Shi, Y., Peng, Y., Fang, J., Zhao, R., Zhao, L.: Vehicle-to-everything (V2X) services supported by IET-based systems and 5G. IEEE Commun. Stand. Mag. 1(2), 70–76 (2017)
46. Cheng, X., Huang, Z., Chen, S.: Vehicular communication channel measurement, modelling, and application for beyond 5G and 6G. IET Commun. 14(19), 3303–3311 (2020)
47. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, New York, NY (2006)
48. Calinski, T., Wong, M.A.: A dendrite method for cluster analysis. Commun. Stat. 3, 1–27 (1974)
49. Alexiadis, V.: Next generation simulation program,the next generation simulation program. Institute of Transportation, (2004)
50. Refaeilzadeh, P., Tang, L., Liu, H.: Cross-validation. Encyclop. Database Syst. 5, 532–538 (2009)

How to cite this article: Zheng, X., Yang, P., Duan, D., Cheng, X., Yang, L.: Real-time driving style classification based on short-term observations. IET Commun. 16, 1393–1402 (2022).
https://doi.org/10.1049/cmu2.12405