Driving risk classification methodology for intelligent drive in real traffic event

Sun, Chuan; Li, Bijun; Li, Yicheng; Lu, Zhenji

DOI
10.1142/S0218001419500149

Publication date
2019

Document Version
Final published version

Published in
International Journal of Pattern Recognition and Artificial Intelligence

Citation (APA)
Sun, C., Li, B., Li, Y., & Lu, Z. (2019). Driving risk classification methodology for intelligent drive in real traffic event. *International Journal of Pattern Recognition and Artificial Intelligence*, *33*(9), [1950014]. https://doi.org/10.1142/S0218001419500149

International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 33, No. 9 (2019) 1950014 (20 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0218001419500149

Driving Risk Classification Methodology for Intelligent Drive in Real Traffic Event

Chuan Sun*

School of Electromechanical and Automobile
Engineering, Huanggang Normal University
Huanggang 438000, P. R. China

The State Key Laboratory of Information
Engineering in Surveying, Mapping and Remote Sensing
Wuhan University, Wuhan 430079, P. R. China
sunchuan33@foxmail.com

Bijun Li

The State Key Laboratory of Information
Engineering in Surveying, Mapping and Remote Sensing
Wuhan University, Wuhan 430079, P. R. China
lee@whu.edu.cn

Yicheng Li

Automotive Engineering Research Institute
Jiangsu University, Zhenjiang 212013, P. R. China
liyucheng070@163.com

Zhenji Lu

Faculty of Mechanical, Maritime and Material
Engineering, Delft University of Technology
Mekelweg 2, 2628 CD Delft, The Netherlands
z.lu@tudelft.nl

Received 29 September 2018
Accepted 15 November 2018
Published 28 February 2019

To solve the problem that existing driving data cannot correlate to the large number of vehicles in terms of driving risks, is the functionality of intelligent driving algorithm should be improved. This paper deeply explores driving data to build a link between massive driving data and a large number of sample vehicles for driving risk analysis. It sorted out certain driving behavior parameters in the driving data, and extracted some parameters closely related to the driving risk; it further utilized the principal component analysis and factor analysis in spatio-temporal data to integrate certain extracted parameters into factors that are clearly related to the specific

*Corresponding author.

driving risks; then, it selected factor scores of driving behaviors as indexes for hierarchical clustering, and obtained multi-level clustering results of the driving risks of corresponding vehicles; in the end, it interpreted the clustering results of the vehicle driving risks. According to the results, it is found that cluster for different risks proposed in this paper for driving behaviors is effective in the hierarchical cluster for typical driving behaviors and it also offers a solution for risk analyses between driving data and large sample vehicles. The results provide the basis for training on safe driving for the key vehicles, and the improvement of advanced driver assistance system, which shows a wide application prospect in the field of intelligent drive.

Keywords: Vehicle driving risk; driving data; classification methodology; intelligent drive.

1. Introduction

When a vehicle is running on the roads, there is a close correlation between driving behavior and driving safety, which does have a direct impact on road transportation safety. However, different high-risk driving behaviors have caused a massive amount of traffic accidents in recent years for the complex traffic environment worldwide.[21] Moreover, how to avoid high-risk driving behavior has been a focus in the field of automatic driving. With the constant development of intelligent vehicles, it has been possible to improve the advanced driver assistance system (ADAS) in data mining based on a continuous accumulation of both online and offline driving data.[13]

Global position system (GPS), geographic information system (GIS) and wireless communication (WC) technology have been used widely to promote the development of intelligent vehicles. Relevant regulations and standards are improved gradually at the same time.[2,5] Nowadays, a full-time monitoring on vehicles, especially on commercial vehicles, has become realizable in most provinces and cities in China in terms of existing technologies and laws and regulations.[27] Through a full-time monitoring on commercial vehicles, various vehicle monitoring data and videos and pictures in data recorders were available. However, the majority of those data usually become historical data at the end for their long-time storage in databases without a full utilization. Therefore, there was only a monitoring on the basic driving conditions for vehicles, such as their traveling routes, monitoring on fuel consumption and loadings, which has been an invisible waste in data sources.

In the big data era, popular technologies like cloud platforms, cloud computing and data warehouse emerge as per today's requirement.[10,28] In a word, it is possible to build a clear correlation among driving data, driving behaviors and driving safety with an effective utilization and integration of various vehicles data through data mining, which also provides technological services in driving safety for intelligent vehicles in the future.[25,26] Meanwhile, the Society of Automotive Engineers (SAE) has divided vehicle automation into six levels. In Level 0 (no automation), the driver has complete control of driving with no assistance from the system. The next four semiautomated driving levels are Level 1 (driver assistance), Level 2 (partial automation), Level 3 (conditional automation), and Level 4 (high automation); the driver and the system share control of the vehicle at these levels. The highest level of vehicle automation is Level 5 (full automation) in which the vehicle can drive anywhere autonomously without help from the driver.[3] However, considering the technical limitations and economic and social development issues, fully autonomous driving may still be decades away. Research on autonomous driving systems has concentrated mainly on the semiautomated driving levels (Levels 1–3), particularly on intermittent autonomous driving during special traffic events.[14,23]

As the most concerned point in the intelligent driving system, driving risk for vehicles provides an introduction to the application of "cluster" in data mining to intelligent transportation areas based on various typical driving data for vehicles collected through existing driving platforms for intelligent vehicle. It aims to carry out hierarchical cluster on driving risk for various sample vehicles to pick out the vehicles with higher driving risk accurately and to conclude a methodology about the clear correlation among driving data, driving behaviors and driving safety. Results in this paper provide the basis for training on safe driving for the key vehicles, and the design of ADAS, which shows a wide application prospect in the field of intelligent driving.

The driving risk classification methodology discussed in the paper belongs to the field of pattern recognition (PR). Its application scenario is the road environment of unpiloted driving in the future, which provides preliminary basis for the highly intelligent driving in the future, and belongs to research on artificial intelligence (AI) in the automotive field. This paper is organized as follows: Section 2 presents a methodology of driving risk classification; in Sec. 3, the case study in real traffic event is described; Section 4 details the application and discussion; Section 5 gives some concluding remarks and discusses the possible improvements.

2. Methodology

There are different degrees of driving risk when various vehicles run on the roads. These help us decide the degree of driving risk through dynamic model analysis for vehicles. On the one hand, a continuous accumulation of massive driving data has built up a data warehouse for driving risk analysis. Increasing development of spatio-temporal mining technologies, on the other hand, also provides a steady theoretical basis. Therefore, it is realizable to provide a methodology for implementing hierarchical cluster on driving risk for vehicles. First of all, to build this methodology successfully, it should extract the characteristic parameters of driving behaviors for vehicles.

2.1. *Driving behavior parameters for vehicles*

Various driving behaviors can be divided into different driving models in accordance with the progress and intention of operation, such as car-following, lane-changing, going straight and turning. In this respect, it is necessary to classify or integrate each type of driving model to make an assessment on the driving behaviors but it is overcomplicated. However, the final purpose is to output driving data regardless of the driving model. With a focus on the driving risk, some characteristic parameters that are closely related to the driving behaviors, especially to the driving risk for

Table 1. The main characteristic parameters for driving behaviors.

	Time Ratio				
Driving Cycle	Overspeed	Idle Speed	Cruise	Operating Range	Kinetic Energy
	η_{speed}	η_{idel}	η_{drive}	s	W
Speed	Average	Standard Deviation	Variance	Quadratic Sum	Maximum
	$\overline{v}$	σ_v	v_{var}	v_{ss}	$v_{\max}$
Acceleration	Positive Acceleration			Acceleration (Vector)	
	Average	Standard Deviation	Maximum	Average	Standard Deviation
	$\overline{a^+}$	σ_{a+}	$a^+_{\max}$	$\overline{a}$	σ_a
	Negative Acceleration			Others	
	Average	Standard Deviation	Maximum	Variance	Quadratic Sum
	$\overline{a^-}$	σ_{a-}	$a^-_{\max}$	a_{var}	a_{ss}

vehicles should be extracted out of a massive amount of driving data. Its purpose is to outline a correlation between driving data and driving behaviors in order to avoid a complicated analysis of each type of driving model. As shown in Table 1, there are 20 main characteristic parameters in total for driving behaviors for reference.

2.2. *Index parameters extraction*

In consideration of the properties of driving data and the main influential factors for driving risk, the following parameters were chosen from Table 1 as the characteristic parameters for hierarchical cluster of driving risk for vehicles.

(1) The proportion of time when running speed exceeds the 80% of limited speed: η_{speed}.

Speed usually plays an important role in the driving safety for vehicles. A high speed would decline drivers' ability to deal with curve roads and obstacles but extend the braking distance, which means a sharp reduction in reaction time for drivers when faced with risk. For instant, it shows that the accident rate would increase by two times for each additional 5 km/h when speed exceeds 60 km/h in which the accident severity would also increase exponentially.[12] Therefore, different maximum speeds for safe driving should be set in line with different road environments. Moreover, based on researches about running speed and limited speed,[1] it is thought that drivers would show a speeding-prone behavior potentially in this paper when running speed exceeds the 80% of the limited speed.

(2) Average speed: $\overline{v}$, speed standard deviation: σ_v.

There is a close relation between average speed and traffic accidents. The higher the average speed is, the more easily traffic accidents would happen.[22] Under a highly running speed, a slight acceleration would lead to a sharp increase in accidents, which has a great impact on the driving safety. In addition, the speed standard deviation reflects the discrete distribution of speed and it positively correlates with accident rate.

(3) Acceleration standard deviation: σ_a, average positive acceleration: $\overline{a^+}$, positive acceleration standard deviation: σ_{a+}.

An acceleration value reflects drivers' operation on the accelerator pedal and brake pedal. Similar to the speed standard deviation, a characteristic pattern of the acceleration standard deviation also reflects the discrete distribution. Driver's operations on accelerator pedal are dealt digitally in the form of a positive acceleration value, which has a direct impact on driving behaviors, such as its start and stop, acceleration and deceleration and ride comfort. Meanwhile, it is closely related to both traffic environment and driving behaviors. It is common that a radical driver tends to press on the accelerator fiercely and accelerate or decelerate suddenly.[15]

(4) Average negative acceleration (deceleration): $\overline{a^+}$, negative acceleration standard deviation: σ_{a+}.

Drivers' operation on the brake pedal or release on accelerator would lead to a change in negative acceleration (braking deceleration), which is a reflection on the drivers' operation on the accelerator and brake pedal.[6] In general, braking deceleration is corresponding to how urgent the braking behavior is. It is a great threat to driving safety through a series of damages, such as failure in evasive action to the front vehicles and rear-end collision. In terms of trucks, it also means a negative impact on loads. Furthermore, it can lead to abrasions in braking equipment for vehicles, such as braking wheel hubs and tires, even heat fading to make brakes out of functions.

In summary, there are eight index parameters in total for a hierarchical cluster of driving risk. In terms of index parameters without being picked up, the reasons can be referred based on the analysis. The first kind of indexes is related to the fuel consumption instead of driving risk for vehicles, such as kinetic energy (W), time proportion of idle speed (η_{idel}) and time proportion of drive (η_{drive}). Main characteristics of the second kind of indexes have been included in aforementioned ones although those indexes are closely related to the driving risk. Correlation between this kind of indexes and driving risk are so close that they can be replaced by each other, such as speed, the quadratic sum of acceleration (ss) and variance (var). A repeated choice of this kind of indexes has little impact on the result but reduces efficiency. The third group of indexes is characterized with lack of information and severe accidents. Therefore, those indexes are not persuasive enough in research like maximum speed ($v_{\max}$), maximum acceleration ($a^+_{\max}$) and maximum deceleration ($a^-_{\max}$). The last group of indexes does not make a difference for the sake of their characteristics (superposition of positive and negative values), such as $\overline{a}$ average of acceleration (vector).

2.3. *Driving risk classification*

From the perspective of driving risk, several sample vehicles should be divided into different groups with the same driving risk based on their running conditions. It aims to implement a reasonable management on those vehicles and design suitable parameters for ADAS.

Cluster analysis is an effective solution for this problem. According to cluster analysis, things can be classified into different groups based on their properties. The properties can be single, various and even combined in which things with similar properties can be divided into a group while different things can be classified into different groups based on their differences. In accordance with the basic principle of cluster analysis, the same group of samples or variables should be defined based on the background information and then divided into different groups. During this process, the number of the group will be decreased step by step and reach the same group at the end. For each cluster level, it should meet the requirement of the standard that differences in the same group are small while that in different groups are large. It can have an assessment on the cluster results based on the variance. In general, a small class variance is corresponding to a group of samples with little difference.[8,11]

In this paper, the hierarchical cluster has been adopted. This method has been widely used in different fields, such as weather, finance, and transportation. Steps for classification of driving risk can be referred to as follows[19]:

Step 1. The factor score of corresponding driving behaviors for n vehicles should be set into a group, respectively and it should include n groups. Distance matrix should be built up for each data point. Moreover, the distance can be obtained according to the aforementioned definition.

Step 2. Two groups of data with the closest distance can be classified in the same group. As a result, only $n-1$ groups are left. Repeat step 1 and build up a new distance matrix.

Step 3. According to the standard in step 2, repeat the procedure in step 2 again and again. It will not stop until all factor scores of corresponding driving behaviors for n vehicles are included in the same group.

In general, the procedure of the aforementioned hierarchical cluster can be shown visually with hierarchical-clustering-diagrams, namely, tree diagram, which can provide assisting services for the later analysis. In addition, there are some points for a supplement.

(1) Principal component analysis and factor analysis are intermediate in the classification for the driving risk. As the first step, principle component analysis tries to pick up the parameter principle factor of driving behaviors and finds a matching point to the driving risk. However, taking principal component analysis as a basis, factor analysis determines the parameter factors of driving

behaviors for vehicles to be picked up for cluster analysis later. In addition, it can be defined theoretically in consideration of its practical meaning.

(2) According to the cluster analysis, results should be practical instead of single clustering without any meaning. Therefore, it is crucial to have a reasonable explanation and comprehensive description for the results. Moreover, some actual cases are necessary for analysis.

(3) In general, its practicability determines whether the clustering result is accurate or not. Meanwhile, some actual cases are necessary for the test and it would be detailed in the analysis of actual cases.

First of all, eight characteristic parameters of driving behaviors for vehicles are combined linearly by principal component analysis. Several of the whole variances will be available through a series of operation, namely, principle component (PC). In consideration of the actual traffic background, this PC reflects the characteristic information of driving risk. For a further reasonable explanation, the rotate factor (RC) can be obtained by factor rotation. Compared with the PC, RC usually contains much clearer information about driving behaviors. All factors picked up in this manner can be endowed with matching driving behaviors. Corresponding factor scores of driving behaviors for sample vehicles in the driving mode picked up by hierarchical cluster can be chosen as cluster parameter indexes. Then, the driving risk should be dealt with in the hierarchical cluster according to their driving behavior characteristics. Finally, results of clustering should be interpreted and tested.

3. Case Study in Real Traffic Event

3.1. *A case study*

The vehicle field test was carried out on the Wuhan–Shiyan Section of G70 Fuzhou–Yinchuan Expressway in the central west of Hubei. 30 drivers were classified into 30 groups to take the test at 9:00 a.m. and complete it within a day as required. The entire test experienced two months in total. Before the beginning of the test conducted by each group, weather conditions in the following day in areas along the way of the test should be surveyed in advance for the purpose of selecting sunny days of good visibility. Without doubt, days of heavy rain or severe weather were avoided.

According to eight parameters concerning behaviors for driving risk, the driving data for each to-be-chosen vehicle in driving mode should be classified and tabulated. The results can be referred to in Table 2.

According to Table 2, it is obvious that there are various differences visually among parameters of driving behaviors for each sample vehicle. However, it is rather difficult to classify the sample vehicles only according to the details in the further analysis or even the standard for the driving risk. To sum up, it is necessary to deal with them with the method proposed in this paper.

Table 2. Driving behavior parameter data in case study.

	η_{speed}	$\overline{v}$	σ_v	σ_a	$\overline{a^+}$	σ_{a+}	$\overline{a^-}$	σ_{a-}
No.	%	km/h	km/h	m/s^2	m/s^2	m/s^2	m/s^2	m/s^2
D01	0.8	69.7	15.1	0.889	0.733	0.478	1.117	0.639
D02	8.4	79.1	20.7	0.739	0.617	0.436	0.922	0.533
D03	15.4	96.8	21.9	1.058	0.908	0.572	1.247	0.703
D04	14.9	92.2	19.7	0.772	0.639	0.442	1.081	0.619
D05	6.9	75.7	19.1	0.800	0.694	0.486	0.978	0.597
D06	6.4	75.5	14.3	0.772	0.692	0.472	0.925	0.447
D07	6.8	75.0	18.9	0.791	0.686	0.481	0.967	0.590
D08	5.7	72.0	17.5	0.674	0.612	0.402	0.844	0.431
D09	6.5	85.4	17.7	0.900	0.742	0.503	1.092	0.575
D10	2.3	75.9	14.1	0.581	0.503	0.317	0.917	0.483
D11	10.6	93.3	18.2	0.775	0.664	0.550	0.878	0.494
D12	3.9	73.5	18.1	0.800	0.689	0.489	0.958	0.506
D13	6.7	80.2	19.3	0.781	0.689	0.467	0.928	0.528
D14	1.4	66.5	15.1	0.649	0.601	0.406	0.805	0.399
D15	7.8	76.8	18.8	0.803	0.725	0.472	0.919	0.475
D16	3.1	70.7	17.8	0.831	0.744	0.486	0.933	0.564
D17	1.4	67.1	15.3	0.656	0.608	0.411	0.814	0.403
D18	6.5	81.4	18.5	0.753	0.657	0.442	1.006	0.513
D19	4.9	77.6	17.7	0.908	0.742	0.475	1.161	0.647
D20	19.0	98.9	21.9	1.056	0.839	0.547	1.181	0.767
D21	11.7	81.5	21.7	0.681	0.614	0.464	0.836	0.469
D22	8.4	77.6	18.5	0.922	0.756	0.531	1.081	0.597
D23	10.2	87.4	19.4	0.847	0.728	0.473	1.022	0.566
D24	7.5	75.9	17.8	0.936	0.753	0.497	1.153	0.692
D25	10.4	88.2	19.6	0.856	0.736	0.478	1.033	0.572
D26	6.6	82.1	18.7	0.761	0.664	0.447	1.017	0.519
D27	5.8	72.7	17.7	0.681	0.619	0.406	0.853	0.436
D28	2.6	72.0	16.8	0.825	0.664	0.553	0.964	0.653
D29	5.5	89.4	17.8	0.550	0.461	0.300	0.869	0.428
D30	3.9	74.2	17.2	0.869	0.761	0.561	0.992	0.511

3.2. *Factor analysis*

SPSS 19.0 was adopted for the factor analysis on parameters for driving behaviors. The principal component analysis was used for the factor-extracting method, and the varimax method was used for the rotation of the PCs. Statistical Product and Service Solutions (SPSS) is one of the commercial software for statistical analysis, data mining, predictive analysis and decision support and it has been recognized in corresponding fields.[24]

Firstly, according to the analysis of the correlation, the test statistic for each variance Kaiser–Meyer–Olkin is 0.731. It reflects a close correlation among all indexes when the KMO statistic reaches 0.7 and above. A significant result of the test of sphericity is 0.000, which means samples are enough to implement the analysis. Therefore, results of the test on the applicability for factor analysis are acceptable. Analytical results of parameter factors for driving behaviors are shown in Tables 3 and 4.

Table 3. Characteristic value and accumulative contribution.

Method	Component	Total	Variance/%	Cumulative/%
Principal component analysis	PC1	5.129	64.114	64.114
	PC2	1.638	20.472	84.586
	PC3	0.622	7.772	92.358
Factor analysis (factor rotation)	RC1	2.604	32.545	32.545
	RC2	2.464	30.798	63.344
	RC3	2.321	29.014	92.358

Table 4. Factor loading matrix.

	Driving behavior parameter								
Component	η_{speed}	$\overline{v}$	σ_v	σ_a	$\overline{a^+}$	σ_{a+}	$\overline{a^-}$	σ_{a-}	Interpretation
PC1	0.731	0.640	0.670	0.933*	0.876*	0.780*	0.858*	0.870*	Acceleration–deceleration
PC2	0.629*	0.668*	0.576*	−0.340	−0.365	−0.359	−0.214	−0.200	Speeding-prone
PC3	0.064	−0.171	0.265	0.019	0.196	0.438*	−0.435*	−0.314	—
RC1	0.924*	0.883*	0.868*	0.209	0.177	0.157	0.227	0.259	Speeding-prone
RC2	0.192	−0.035	0.311	0.729*	0.822*	0.920*	0.330	0.409	Acceleration
RC3	0.213	0.323	0.040	0.641*	0.481	0.240	0.900*	0.813*	Deceleration

Note: * means data with high factor loading. The higher the loading (absolute value) between a variance and a factor is, the closer this factor is to the variance.

According to Table 3, three principal components, PC1, PC2 and PC3 are included, on the premise that cumulative variance is 92.358% (>90%). It is obvious that these three principal components have offered 92.358% of the total original information, which meets the requirement of the principle of factor analysis. Meanwhile, there is no change in the total cumulative contribution rate in the rotation, that is, there is no loss in the total information. In accordance with the results of factor rotation, all variance contribution rates of three rotation factors, RC1, RC2 and RC3, are about 30%. To sum up, RC1, RC2, and RC3 have a significant impact on the characteristics of driving behaviors as well as the greatest contribution to the characteristics of driving behaviors.

Factor loadings mean the loading of the variance in its matching factor. In fact, it is a correlation coefficient between variance and the common factor. It not only shows the dependent relationship between variance and factor but also reflects the significance of variance for its corresponding factor. Based on Table 4, there is a close correlation between PC1 and all parameters concerning acceleration like σ_a, $\overline{a^+}$, σ_{a+}, $\overline{a^-}$ and σ_a. PC2 is closely related to the parameters for speed, such as η_{speed}, $\overline{v}$ and σ_v. However, PC3 is only related to some parameters for acceleration like σ_{a+} and $\overline{a^-}$. To sum up, as a factor, PC1 is closely related to acceleration–deceleration behavior while PC2 to speeding-prone behavior. However, PC3 is meaningless here.

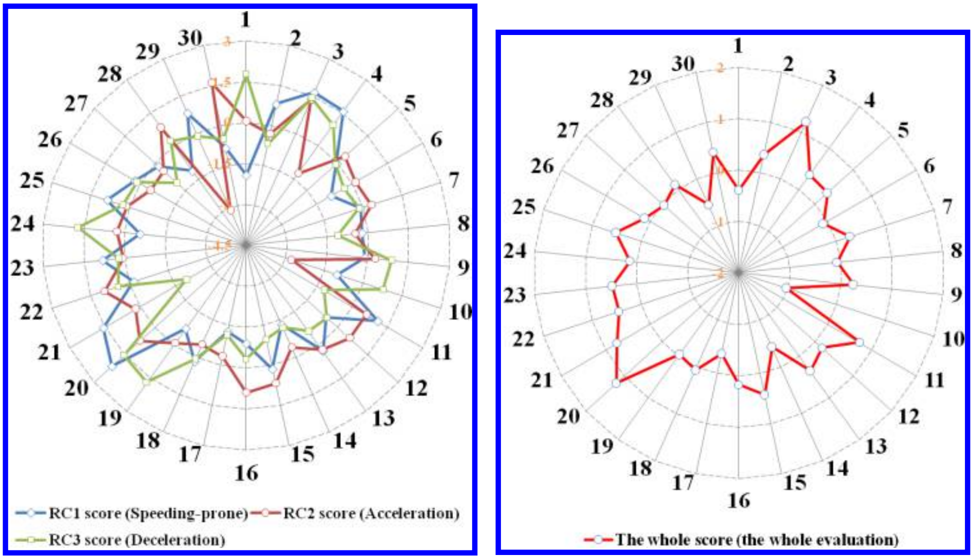
Nevertheless, both PC1 and PC2 fail to define the correlation based on the acceleration and deceleration. Researches on the driving risk are not enough to be persuasive. Therefore, rotation analyses for factors are necessary. There is a close correlation between the rotation factor RC1 and parameters for speed, which is the same as the symbolic meaning of PC2. In addition, RC2 is only closely related to the parameters for positive acceleration like σ_a, $\overline{a^+}$ and σ_{a+}, while RC3 are only closely related to the parameters for deceleration like σ_a, $\overline{a^-}$ and σ_a. To sum up, RC1 shares the same symbolic meaning of PC2; as a factor, RC2 is closely related to acceleration behavior while RC3 to deceleration behavior. All factor analyses for driving behavior have been finished. As a result, eight parameters (physical variance) for driving behavior which are closely related with each other were divided into four factors (specific meaning) with specific information for driving behavior.

3.3. *Hierarchical cluster for driving risk*

According to the symbolic meaning for the driving behavior of the aforementioned principal component (PC1) and rotation factors (RC1, RC2 and RC3), it is realizable to implement the hierarchical cluster for driving behavior characteristics for the sample vehicles in line with their corresponding driving behaviors (acceleration–deceleration behavior, speeding-prone behavior, acceleration behavior and deceleration behavior). This result shows different driving behavior risks of different vehicles being clustered and index of clustering means the factor score of its corresponding sample vehicle. Meanwhile, a mathematical model of the factor score can be obtained through the regression estimate. Each factor is converted into a linear combination for parameter variance and then the factor score can be made available, the whole score of the sample vehicles can be obtained finally. Factor score (the whole score included) is one of the important bases for the evaluation (the whole evaluation included) on the sample vehicles, which is a quantitative ranking reflecting the driving behavior risk for vehicles. In order to implement analyses on the driving behavior risk for 30 sample vehicles visually, a radar map for driving risk is drawn according to the corresponding factor score (see Fig. 1).

According to Fig. 1, the higher the factor score of vehicles is, the riskier the driving behavior is in accordance with the corresponding driving behavior risk for the sample vehicles. Moreover, the whole score is a reflection of the whole driving risk for sample vehicles. For example, it can be seen directly that the speeding-prone behavior of No. 20 vehicle tops in the list of risky behavior while that of No. 10 comes last. For a further analysis, a hierarchical cluster has been implemented for the sample vehicles according to their corresponding factor score. It can be referred to in Table 5.

It is found that there is certain coupling relation among the clustering results of those five types of risks. For example, both No. 3 and No. 20 vehicles show high risk in those five types of risk cluster while No. 1 vehicle reflects a middle risk in terms of acceleration-deceleration behavior but low risk in terms of speeding-prone behavior. It is these differences that show characteristic differences of various driving behavior risk for sample vehicles.



(a) Each factor score

(b) The whole score

Fig. 1. Radar map of driving risk level for sample vehicle.

Table 5. Cluster analysis of driving behavior risk.

Risk Level	Acceleration–Deceleration	Speeding-prone	Acceleration	Deceleration	The Whole Behavior
I	10,14,17,29	1	10,29	8,11,14,15,17,21,27	1,10,14,17,29
II	6,8,27	6,10,12,14,16,17,19,24,28,30	2,4,8,14,17,18,26,27	2,5,6,7,12,13,16,29,30	6,8,19,27
III	1,2,5,7,11,12,13,15,16,18,21,26,28,30	5,7,8,9,13,15,18,22,26,27	1,5,6,7,9,13,19,21,23,24,25	18,22,23,25,26,28	7,9,12,16,18,24,26,28
IV	4,9,19,22,23,24,25	2,11,23,25,29	11,12,15,16,20,22,28	4,9,10	2,4,5,11,13,15,21,22,23,25,30
V	3,20	3,4,20,21	3,30	1,3,19,20,24	3,20

Application of the results can be referred to as follows. It is necessary to carry out key monitoring and investment on No. 3 and No. 20 vehicles throughout the entire process. For vehicles with fierce speeding-prone behavior, such as No. 4 and No. 21 vehicles, a targeted education on driving safety can remind them of a reasonable speed. There is a high risk in the deceleration behavior of No. 1, No. 19 and No. 24 vehicles. It is crucial of them to avoid a fierce operation on the brake pedal and slow down gradually in advance based on the actual running conditions.

In conclusion, the hierarchical cluster has been implemented on the driving risk for vehicles. On the one hand, sample vehicles with highly driving risk can be picked out based on the four kinds of typical driving behaviors (acceleration–deceleration behavior, speeding-prone behavior, acceleration behavior and deceleration behavior). On the other hand, sample vehicles with highly driving risk can be chosen out of massive samples through the whole driving behavior, which reflects the risk of driving behavior for vehicles as a whole. In addition, a targeted monitoring can be launched on sample vehicles for a specific driving behavior with a specific driving risk in accordance with the distribution of driving risk. Although there are only 30 sample vehicles, methods for research are still practical within a large sample for vehicles. In practice, it is significant to launch a targeted monitoring on the key vehicles, including carrier vehicles, commuter vehicles, and special vehicles. Moreover, relevant training on driving safety and the design of the corresponding ADAS are necessary.

3.4. *Accuracy and validation of the classification model*

Cluster analysis is an exploratory method for data mining and there is no general strict and uniform standard for it. It depends on the analytical objects. However, there are some general principles as follows. Clustering results should be characterized with an extensively but clearly discriminated dimension so that it is easy to make reasonable conclusions and interpretations for each type of groups with the digital profession as a basis. An evaluation can also be launched on the results of the hierarchical cluster for the driving risk according to the sameness of the results. In other words, clustering can be carried out on the operation mode of the same drive for the test in another method and then it is easy to conclude the differences between two pieces of results. It is a methodology generated from data mining theory and it has been an effective method for test on clustering results so far.[18,24]

Results of the hierarchical cluster for driving risk have drawn a clear image of risk distribution for four kinds of typical driving behaviors, namely, acceleration–deceleration behavior, speeding-prone behavior, acceleration behavior and deceleration behavior. The interpretation for each behavior is reasonable in which all kinds of driving risks for vehicles have been covered without a single accumulation in any clustering result. The whole driving risks for sample vehicles originated from the same data source are clustered in different cluster methods. A comparison between those results and that of the hierarchical cluster has been made to conclude the differences in each clustering result.

In general, cluster methods consist of the hierarchical cluster, K-means, Self-Organizing Map (SOM) and Fuzzy c-means (FCM).[4,20] K-means is a typical algorithm for cluster based on distance, which makes it suitable in the cluster for massive data. Different from K-means, SOM is an algorithm for cluster based on neural networks and it is characterized with Unsupervised Learning. As an algorithm for the cluster, FCM determines the dependence on a cluster for each data point based on the dependent degree. Taking results of the hierarchical cluster as the reference, comparative standards are shown as follows: Total number of samples with differences, that is, the total number of samples with difference from that of the hierarchical cluster; operation time, that is, time consumption for the solution of the cluster; average consistency, that is, the average of the consistency of each method with the hierarchical cluster. Results of clustering comparison for the whole driving

Table 6. Clustering comparison for the whole driving risk.

Clustering Method	Difference Amount	Run Time	Average Consistency
Hierarchical Cluster	—	0.125 s	—
K-means	0	0.138 s	100%
FCM	5	0.371 s	87.04%
SOM	8	1.035 s	78.72%

risk for sample vehicles based on the aforementioned methods can be referred to in Table 6.

It is found that results by K-means are totally the same with that by the hierarchical cluster in terms of the group number, recorded distribution in groups and group characteristics. However, there are some differences among FCM, SOM and cluster in this paper. Clustering results still have a good performance in classification and interpretation to some extent. It is a fact that some driving risks for vehicles are actually within two groups. There are some differences in classification according to different methodologies. Therefore, it leads to some differences. Nevertheless, it is reasonable as a whole without serious differences among those four kinds of methods. In terms of practical time, the first three kinds of methods are closely the same as each other while SOM is practical for a longer time.

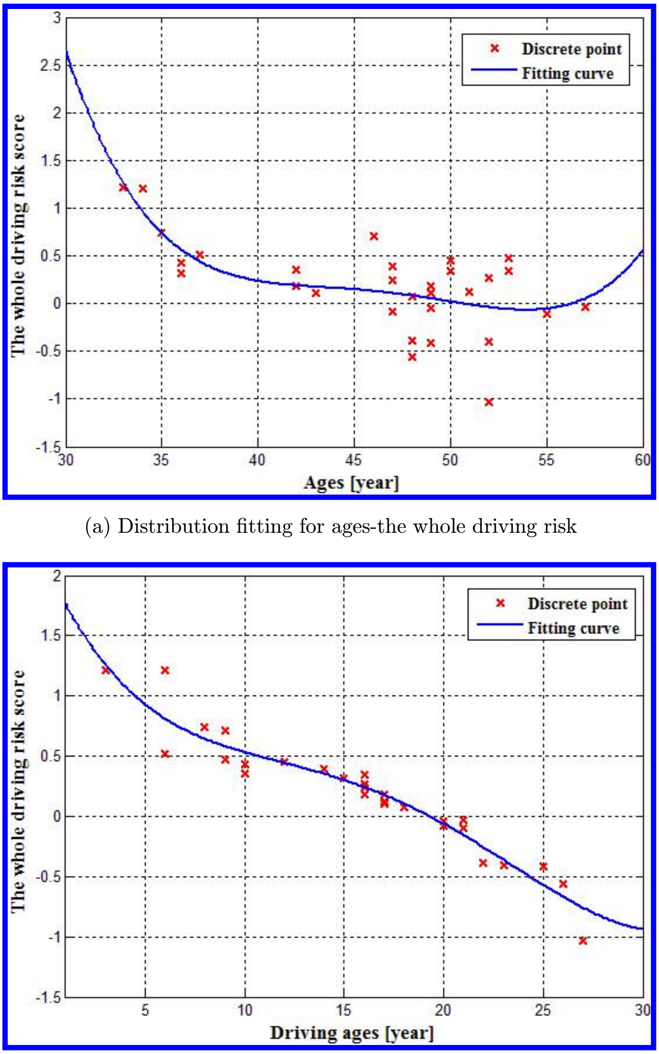
However, each algorithm also owns limitations in practice. In terms of the hierarchical cluster, it is impossible to train and correct by self on classification or combination because it is characterized with classification by the hierarchy. K-means shows high dependence on the determination of the original point, which determines the final stability of the results. For FCM, it is necessary to set the number of clusters, which limits the quality of the best solution as a whole. As neural networks by artificial simulation, SOM is slow for solutions, especially dealing with massive data. To sum up, a comprehensive thinking is needed for the reasonability of clustering object and result interpretation to choose a suitable method in practice.

4. Application and Discussion

4.1. *Correlation between drivers' factors and driving risk*

In consideration of the difference in the personalities of different drivers, it is common that there are some differences in the operation on vehicles and consciousness on the surroundings when they drive in the same traffic environment.[7] It results from the fact that differences in personal recognition and behaviors by personal characteristics lead to a difference in driving risks. There is a range in terms of drivers' factors, such as gender, age, driving age, vision and driving style. All these factors have an equal impact on drivers' performance in the entire driving. In view of the qualitative or quantitative influence of the aforementioned factors on drives, as well as the inherent correlation between individual factors of drivers and driving risk, there is further research needed.

There are 30 drivers in total, including professional drivers, private car owners and so on, for the test on the operation mode in driving. According to statistics, there are 26 male drivers and 4 female drivers. Their ages vary from 33 to 57 (standard deviation: 6.61) while the driving age varies from 3 to 27 ((standard deviation: 6.61). A polynomial function fitting method has been adopted to fit the relation between the abscissa and the ordinate in which it takes individual factors including age and driving age as the abscissa while the whole driving risk for a vehicle as the ordinate. A fitting curve for the relation between the individual factors and driving risk for vehicles can be referred to in Fig. 2.



(a) Distribution fitting for ages-the whole driving risk

(b) Distribution fitting for driving ages-the whole driving risk

Fig. 2. Distribution correlations between drivers' factors and driving risk.

According to Fig. 2(a), the driving risk for vehicles decreases with age as a whole. From the figure, it can be seen that it comes to its stability after 40 and it reaches the lowest point between 50 and 50 while there will be a slow increase after that. The reason is that the young drivers who are characterized with shorter driving ages and lack of driving experience tend to be adventurous in driving. Different from the young, the old ones usually bear much more worry for degeneration on both physiological and psychological functions with age, which results in an increase in their driving risks. Based on Fig. 2(b), the driving risk for vehicles will decrease steadily as driving age increases. The driving risk of the student drivers, whose driving ages are less than 3 years, is extremely high. There are four female drivers in the test, and they are Nos. 12, 14, 15 and 24. According to a comprehensive analysis on their whole driving risk for vehicles, it is found that the results mainly fall in the middle level, that is, driving risk for females is neither the highest nor the lowest.

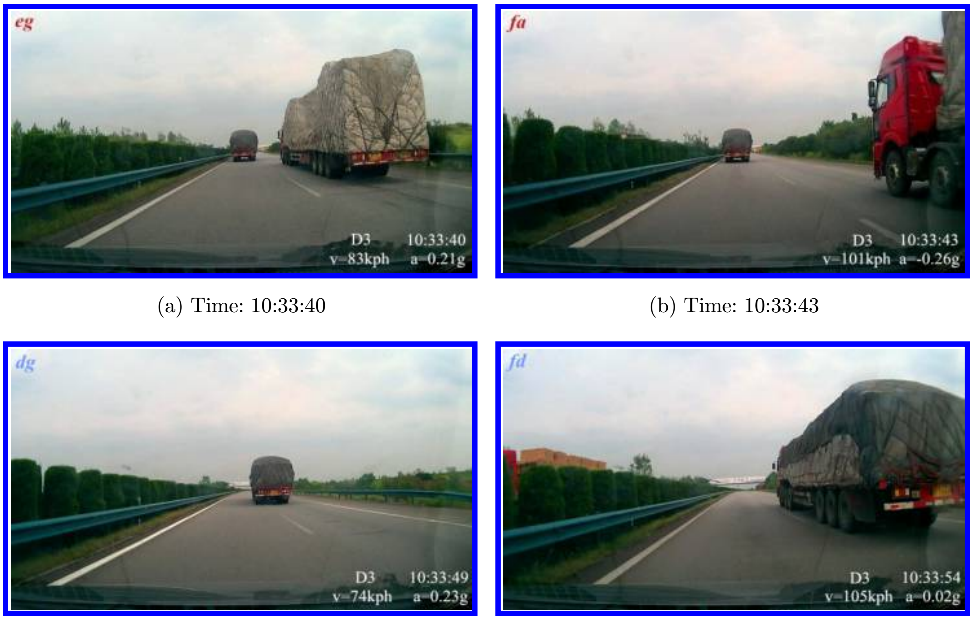
To sum up, there are some conclusions: The driving risk for vehicles decreases steadily with age as a whole and there is a slow increase after 55; the driving risk for vehicles decreases continuously and it comes to stability finally; the driving risk for vehicles of female drivers falls in the middle level without any extremely high or low risk.

According to some statistical analyses on traffic accidents, it shows that the traffic accidents usually happen among the young drivers with short driving ages.[17] Compared with the male drivers, the female drivers tend to engage themselves into slight traffic accidents more easily. However, chances for serious traffic accidents among the male drivers are far higher than that among the female drivers. Results of the distribution of the rate for traffic accidents on roads concerning age, driving age and gender have offered some support for the conclusion in this paper. Conclusions about the correlation between the individual factors of drivers and the driving risk for vehicles proposed in this paper are generally consistent with the previous statistical analyses on traffic accidents.

4.2. *Event study for high-risk driving*

High-risk driving events can be preliminarily captured from Table 5. Then, video surveillance data are rapidly positioned and the vehicle surveillance videos in a period during which high-risk driving events occur in a concentrated manner are also exported according to a time slot on the vertical coordinates.[16] Combining the surveillance video, emphasis of analysis is laid upon high-risk driving events. Resultantly, analysis efficiency and accuracy are improved on the one hand; on the other hand, difficulties in vehicle surveillance video extraction, capture and positioning are eliminated. As shown in Fig. 3, it is a vehicle surveillance video of high-risk driving events captured in this time slot.

As a 10 min long (40–50 min) surveillance video of vehicle No. 3 has been acquired, concentrated searching and positioning of high-risk driving events in this period can be carried out to capture a typical event among them, that is, the behavior that the



(a) Time: 10:33:40

(b) Time: 10:33:43

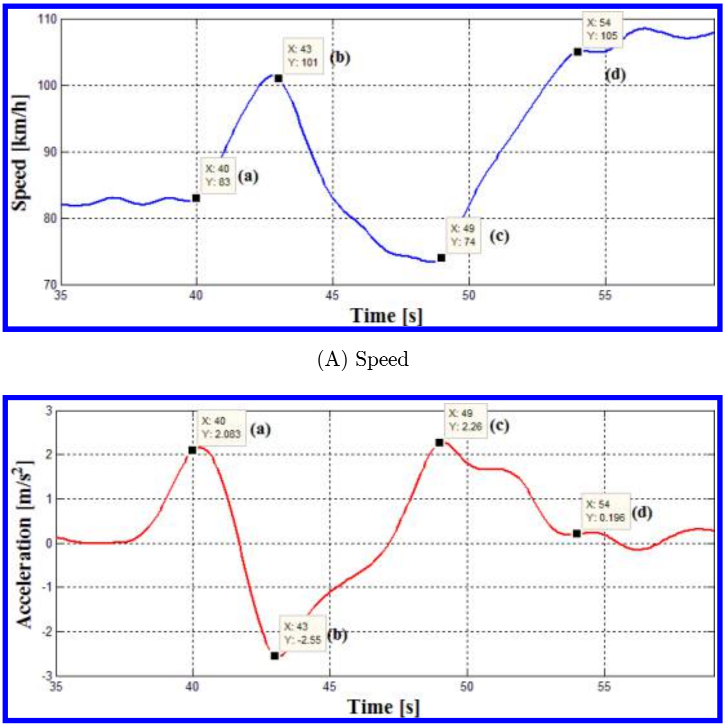
(c) Time: 10:33:49

(d) Time: 10:33:54

Fig. 3. Vehicle surveillance video of high-risk driving events.

target vehicle overtakes two large semitrailers twice consecutively on the expressway without incurring any accidents. However, it has been deemed as high-risk driving operation in nonconformity with safety specifications. Event procedure is presented in Fig. 3. (a) [10:33:40] The target vehicle moves on an overtaking lane (the left lane) attempting to overtake a semitrailer on the running lane (the right lane); however, there has already been a semitrailer ahead just completing the overtaking without changing the lane. In this case, the speed of the target vehicle is 83 km/h and the acceleration is 0.21 g. (b) [10:33:43] The target vehicle fulfills overtaking with a speed arriving at 101 km/h; however, it is found that the car ahead changes the lane slowly and there also exists a phenomenon of excessively close car following. Therefore, a measure of braking intervention is taken with deceleration of 0.26 g. (c) [10:33:49] After braking, the speed of the target vehicle reduces to 74 km/h waiting for the completion of lane changing conducted by the car ahead; then, the target vehicle begins to accelerate for the second time to overtake the car ahead and the acceleration is restored to 0.23 g. (d) [10:33:54] The target vehicle accelerates to 105 km/h and overtakes steadily. The entire event continues for 25 s approximately. Time domain plot for speed and acceleration of high-risk driving events is given in Fig. 4 that has been subjected to proper denoising, filtering and smoothing.

Figure 4 reveals time-domain variations of key driving data (speed and acceleration) related to such a period of continuous driving events of high risks. Four sections intercepted are presented and denoted as (a), (b), (c) and (d) clearly



(A) Speed

(B) Acceleration

Fig. 4. Time domain plot for speed and acceleration of high-risk driving events.

signifying dramatic fluctuations of speed and acceleration in the process of consecutive overtaking. Although they both fall into the range of security threshold, it is still difficult for the driver to effectively control the vehicle in the event of an emergency.

Through analysis, it becomes clear that the case that the target vehicle overtakes two heavy-duty semitrailers on the expressway consecutively falls into the category of high-risk driving behavior. Such events possess two points of high risks. First, if the target vehicle overtakes under the circumstance that the car ahead has not completed lane changing, the collision risk can substantially increase. Second, if the overtaking is targeted at two heavy-duty semitrailers, their long bodies make overtaking distance and time increase invisibly; in addition, such two semitrailers have been fully and high-loaded, their rollover may incur dramatically severe outcome. For this reason, drivers of three vehicles conducted necessary and reasonable driving operations. While the target vehicle applied brake in time to keep a proper distance, the car ahead rationally changed its lane after overtaking and another vehicle on the right decelerated when it to as being overtaken (inferred by video observation). As a result, the entire event did not incur any traffic accidents in spite of high risks.

Nevertheless, education and warning about driving security specifications should be reinforced for the driver of vehicle No. 3.

5. Conclusions and Recommendations

There is a further research on the hierarchical cluster for the driving risk in this paper. In order to find solutions for the difficulties to match the existing driving data and the driving risk for large sample vehicles, eight parameters for driving risk have been picked out from 20 general parameters in a total of driving behavior for vehicles based on the accumulative data. A total of 4 factors are chosen out of the aforementioned eight parameters through a combination of principal component analysis and factor analysis. Those four factors have been proven in direct relation to the typical driving behavior (including the whole driving behavior). After that, it is necessary to cluster the vehicles in the operation mode for the test based on their factor scores for driving behavior with the hierarchical cluster. Moreover, a reasonable interpretation has been given for the clustering results of driving behavior for vehicles and then an evaluation also been offered in three aspects. Finally, an analysis on a driving scene with high-driving risk has been implemented for further research on the correlation between individual factors of drivers and the driving risk. According to the results, it is found that the cluster for different risks proposed in this paper for driving behaviors is effective in the hierarchical cluster for typical driving behaviors for vehicles and it also offers a solution for risk analyses between driving data and large sample vehicles. These results provide effective support for the monitoring on key vehicles, training on driving safety and the design of ADAS in practice. Moreover, it shows a wide application prospect in the field of intelligent driving at the same time. The methods proposed in this paper have a good performance for solutions. However, in order to reach a completely autonomous work and constant online processing for different driving data, there is a need for further programming and debugging.

Acknowledgments

This study is sponsored by the National Natural Science Foundation of China (41671441, 41531177, U1764262); Hubei Provincial Natural Science Foundation of China (2018CFC863); the Research Project of Hubei Provincial Department of Education (Q20182905); the National Undergraduate Training Program for Innovation and Entrepreneurship (201710514011); the Scientific Research Project of Huanggang Normal University. Special thanks go to Professor Xiaomin Chen for providing the paper development.

References

1. L. Aarts and I. Van Schagen, Driving speed and the risk of road crashes: A review, *Accident Anal. Prev.* **38**(2) (2006) 215–224.

2. M. M. Anjum *et al.*, Collision detection of vehicle and coverage of using GPS and GSM technology, *Int. J. Eng. Sci.* **7**(3) (2017) 4989–4992.
3. M. Bahram, M. Aeberhard and D. Wollherr, Please take over! an analysis and strategy for a driver take over request during autonomous driving, in *Intell. Vehicles Symp. (IV)* (IEEE, 2015), pp. 913–919.
4. C. Bai, D. Dhavale and J. Sarkis, Complex investment decisions using rough set and fuzzy c-means: An example of investment in green supply chains, *Eur. J. Oper. Res.* **248**(2) (2016) 507–521.
5. K. Bhalla *et al.*, Rapid assessment of road safety policy change: Relaxation of the national speed enforcement law in Russia leads to large increases in the prevalence of speeding, *Injury Prev.* **21**(1) (2015) 53–56.
6. J. Brady and M. O'Mahony, Development of a driving cycle to evaluate the energy economy of electric vehicles in urban areas, *Appl. Energ.* **177** (2016) 165–178.
7. M. J. Cassidy and J. R. Windover, Driver memory: Motorist selection and retention of individualized headways in highway traffic, *Transport. Res. A-Pol.* **32**(2) (1998) 129–137.
8. L. Chen *et al.*, Fault detection based on AP Clustering and PCA, *Int. J. Pattern. Recognit. Artif. Intell.* **32**(2) (2018) 1850001.
9. B. C. Cronk, How to use SPSS®: *A Step-by-Step Guide to Analysis and Interpretation* (Routledge, 2017).
10. Z. Deng, J. Zhang and T. He, Automatic combination technology of fuzzy CPN for OWL-S web services in supercomputing cloud platform, *Int. J. Pattern. Recognit. Artif. Intell.* **31**(7) (2017) 1759010.
11. N. Gillis, D. Kuang and H. Park, Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization, *IEEE T. Geosci. Remote* **53**(4) (2015) 2066–2078.
12. C. Goldenbeld and I. van Schagen, The credibility of speed limits on 80 km/h rural roads: The effects of road and person (ality) characteristics, *Accident Anal. Prev.* **39**(6) (2007) 1121–1130.
13. C. H. Jang *et al.*, Design factor optimization of 3D flash lidar sensor based on geometrical model for automated vehicle and advanced driver assistance system applications, *Int. J. Auto. Tech.-Kor.* **18**(1) (2017) 147–156.
14. K. Kircher, A. Larsson and J. A. Hultgren, Tactical driving behavior with different levels of automation, *IEEE T. Intell. Transp.* **15**(1) (2014) 158–167.
15. N. Kovácsová, T. Lajunen and E. Rošková, Aggression on the road: Relationships between dysfunctional impulsivity, forgiveness, negative emotions, and aggressive driving, *Transport. Res. F-Traf.* **42** (2016) 286–298.
16. M. L. Lee *et al.*, High risk of near-crash driving events following night-shift work, *Proc. Natl. Acad. Sci. USA* **113**(1) (2016) 176–181.
17. D. T. Levy, Youth and traffic safety: The effects of driving age, experience, and education, *Accident Anal. Prev.* **22**(4) (1990) 327–334.
18. D. Li, S. Wang and D. Li, *Spatial Data Mining* (Springer Berlin Heidelberg, 2015).
19. A. Liu *et al.*, Hierarchical clustering multi-task learning for joint human action grouping and recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(1) (2017) 102–114.
20. A. Majumder, L. Behera and V. K. Subramanian, Emotion recognition from geometric facial features using self-organizing map, *Pattern Recognit.* **47**(3) (2014) 1282–1293.
21. T. O. Nævestad, R. O. Phillips and B. Elvebakk, Traffic accidents triggered by drivers at work — A survey and analysis of contributing factors, *Transport. Res. F-Traf.* **34** (2015) 94–107.
22. R. Sánchez-Mangas *et al.*, The probability of death in road traffic accidents. How important is a quick medical response? *Accident Anal. Prev.* **42**(4) (2010) 1048–1056.

23. W. Sheng *et al.*, An integrated manual and autonomous driving framework based on driver drowsiness detection, in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)* (IEEE, 2013), pp. 4376–4381.
24. I. H. Witten *et al.*, *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, 2016).
25. L. Yan *et al.*, Driving mode decision making for intelligent vehicles in stressful traffic events, *Transport. Res. Rec.* **2625** (2017) 9–19.
26. L. Yan *et al.*, Driving risk status prediction using Bayesian networks and logistic regression, *IET Intell. Transp. Symp.* **11**(7) (2017) 431–439.
27. S. Yin and Z. Huang, Performance monitoring for vehicle suspension system via fuzzy positivistic C-means clustering based on accelerometer measurements, *IEEE-ASME T. Mech.* **20**(5) (2015) 2613–2620.
28. J. Zhao *et al.*, A novel clustering-based sampling approach for minimum sample set in big data environment, *Int. J. Pattern. Recognit. Artif. Intell.* **32**(2) (2018) 1850003.

Chuan Sun received his Ph.D degree in Vehicle Engineering from Wuhan University of Technology, China, in 2017. He has been working as Assistant Professor in Huanggang Normal University since 2017. His research interests include vehicle safety and driving behavior.

Bijun Li is currently a Professor at Wuhan University, China. He received his Ph.D. degree from Wuhan University, China, in 2008. His research interests include intelligent vehicle and automobile navigation.

Yicheng Li received the Bachelor degree in Electrical Engineering and Automation from Hebei University of Architecture in 2011, and the Master degree and Ph.D. degree in Intelligent Transportation Engineering from Wuhan University of Technology, in 2014 and 2018. He is currently an assistant professor in Jiangsu University. His research interests include intelligent transportation systems, computer vision and image processing, 3D data processing.

Zhenji Lu joined the Department of Biomechanical Engineering at TU Delft in August 2014 as a Marie Curie Fellow in the HF Auto project. He is studying human behavior in highly automated driving during transient manoeuvres such as merging, splitting, platoon entry, platoon exit, authority transitions between manual and highly automated driving, and unexpected situations, such as sensor and computer failure.