Driving risk assessment using cluster analysis based on naturalistic driving data

Conference Paper · October 2014

DOI: 10.1109/ITSC.2014.6958104

6 authors, including:

Yang Zheng
Harvard University
84 PUBLICATIONS **1,399** CITATIONS

Jianqiang Wang
Tsinghua University
170 PUBLICATIONS **2,601** CITATIONS

Chenfei Yu
Tsinghua University
6 PUBLICATIONS **121** CITATIONS

Kenji Kodaka
6 PUBLICATIONS **120** CITATIONS

2014 IEEE 17th International Conference on
Intelligent Transportation Systems (ITSC)
October 8-11, 2014. Qingdao, China

Driving Risk Assessment using Cluster Analysis based on Naturalistic Driving Data*

Yang Zheng, Jianqiang Wang, Xiaofei Li, Chenfei Yu, Kenji Kodaka and Keqiang Li

***Abstract*—In addition to the real traffic accident data, naturalistic driving data can allow researchers gain insights into the factors that cause risk/hazard situations. This paper considers a comprehensive naturalistic driving experiment to collect detailed driving data on actual Chinese roads. Using acquired real-world driving data, a near-crash database is built, which contains vehicle status, potential crash object, driving environment and road type, and weather condition. K-means cluster analysis is applied to classify the near-crash cases into different driving risk levels using braking process features, namely maximum deceleration, average deceleration and percentage reduction in the vehicle kinetic energy. The results indicate that the velocity when braking and triggering factors have strong relationship with the driving risk level involved in near-crash cases.**

I. Introduction

Over the last two decades, significant progress has been made in all aspects of vehicle safety system [1]. Efforts that aim to advance a safer vehicle traffic system can mainly be divided into two areas: 1) active safety [2][3], and 2) passive safety [4]. Although many encouraging achievements have been made, the number of road fatalities remains unacceptably high, and traffic accidents are considered as a major public health problem [5]. Because responsibility for traffic accidents involves the vehicle, driver and road, we must not only improve the safety performance of vehicles but also better understand the factors that affect driving risk and identify the factors that result in accidents to make road transportation much safer.

Many research activities have been conducted to seek better understanding of the factors that affect the probability and injury severity of crashes in the hope of providing police countermeasures to reduce the number of crashes [6]. For example, Al-Ghamid et al. showed that the location and cause of accidents were most significantly associated with accident severity using logistic regression based on accident-related data [7]. Chang et al. proposed a classification and regression tree model to establish the relationship among injury severity, driver/vehicle characteristics, environment factors and accident severity using office recorded vehicle accident data [8]. These studies have typically been based on official traffic accident statistics which have two major limitations: 1) lack of detailed driving data; 2) difficult to collect and acquire (usually collected by traffic police agency). Hence, the studies stated above do not consider the relationship between the detailed driving data (e.g. vehicle speed, acceleration, braking and steering information) and accident severity. Recent developments in vehicle instrumentation techniques have made monitoring the naturalistic driving behavior and obtaining detailed driving data both technologically possible and economically feasible. For example, NHTSA sponsored the project '100-Car Naturalistic Driving Study', which is the first large-scale instrumented vehicle study undertaken to collect naturalistic driving data in the United States [9]. With access to naturalistic driving data, many researchers have proposed new methods and gained new insights in traffic safety involving drivers, vehicles and roadways [10]-[13]. Malta et al. focused the pedal signals and driver speech to better understand the driver behavior under potential threats using a large real-world driving database [10]. Aoude et al. developed SVMs and a hidden Markov model for driver behavior classification at intersections and validated the proposed algorithms using naturalistic intersection data [11].

This paper focuses on the analysis of the factors that affect the driving risk using naturalistic driving data. In this study, we first conducted a comprehensive naturalistic driving experiment to collect detailed driving data on actual Chinese roads and then built a near-crash database through designing a novel data transcription protocol. The driving risk level under near-crash cases is represented by the braking process characteristics. The K-means cluster method is adopted to classify the near-crash cases into different risk level groups based on these three braking process features. The results indicate that the velocity when braking and triggering factors have the largest influence on the driving risk level.

II. Naturalistic Driving data and Database Design

To build a firm research foundation for driving risk assessment and enhanced driving safety, two components are essential: 1) actual driving data and 2) careful experimental design. In contrast to field operational tests, data collection is performed through naturalistic and low intervention method in actual traffic condition. This section introduces the experimental equipment and experiment design, describes

*Research supported by National Natural Science Foundation of China (Grant No. 51175290) and the joint project of Tsinghua and Honda.

Yang Zheng, Xiaofei Li and Keqiang Li are with the State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing 100084,China.(e-mail:zhengy13@mails.tsinghua.edu.cn,lixf11@mails.tsinghua.edu.cn, likq@tsinghua.edu.cn.)

Jianqiang Wang is with the State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing 100084, China. (Corresponding author; phone: (+86-10-62795774; e-mail: wjqlws@tsinghua.edu.cn)

Chenfei Yu was with the State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing 100084, China. She is currently with Pan Asia Technical Automotive Center Co., Ltd., Shanghai 201201, China. (e-mail: yuchenfei@tsinghua.org.cn)

Kenji Kodaka is with Honda R&D Co., Ltd. Automobile R&D Center, Tochigi 321-3393, Japan. (e-mail: Kenji_Kodaka@n.t.rd.honda.co.jp.)

data transcription protocol and then builds the near-crash database.

A. Data-collection Equipment

The naturalistic driving experiments were conducted on a Honda Crosstour. The vehicle was provided with instruments to collect driver, vehicular and road data under real-world conditions. The data-collection system installed in the experimental vehicle included GPS, vehicle sensors, two driving recorders (DR) and four CCD cameras (Fig. 1). The four cameras recorded detailed video scenes including 1) Forward view, 2) Right-side forward view, 3) Left-side forward view, 4) Driver's facial expression. One DR recorded the vehicle speed obtained by GPS, brake signal, steering signal, three-axis acceleration information and detailed video collected by the facial-expression and forward view cameras. The other DR recorded the video collected by both left-side and right-side forward view cameras for the convenience to code the incidents.

In our study, we focused on the risk factors and driver behaviour under near-crash scenario in the naturalistic driving experiment. Near-crash case means the driver must perform emergency braking operation; otherwise a real crash will occur. For the experimental data collection, a near-crash case means that the deceleration of the experimental vehicle reaches a threshold value instead of happening actual accidents. Hence, the data-collection system recorded the vehicle state (speed, brake signal, steering signal and three-axis acceleration) and four video scenes when a large deceleration was detected. The recording time started approximately from 10 s before the triggering point to 5 s after the triggering point, which means that a typical near-crash case has approximately a 15-s signal and video sequence. Fig. 2 shows examples of the recorded driving signals.

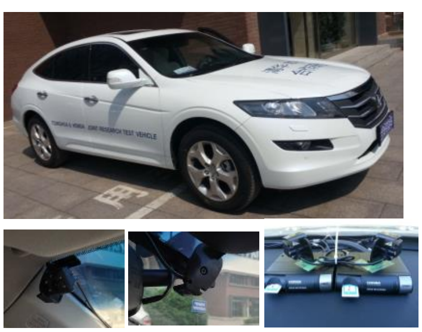

Figure 1. Experimental vehicle and equipment

B. Experiment Design

The naturalistic driving route contained all road types: inner-city highway, city ring road, inter-city road (mixed traffic conditions) and rural road (poor road structure and crowded living quarters). A total of 31 drivers, who have signed the informed consent form, participated in these naturalistic driving experiments at their normal driving state. The experiment lasted for 60 days at 6–7 h/day, which resulted in an approximately 400-h naturalistic driving time and over 8500-km naturalistic driving range. The schedule of the entire experiment plan is listed in TABLE I, and TABLE II lists the naturalistic driving distance on the different road types.

Among the 31 drivers, 9 were female and 22 were male with regular driving license. They were 43 years old, on average (age range from 25 to 67 years) and have possessed a driving license for a mean period of 16 years (ranging from 3 to 48 years).

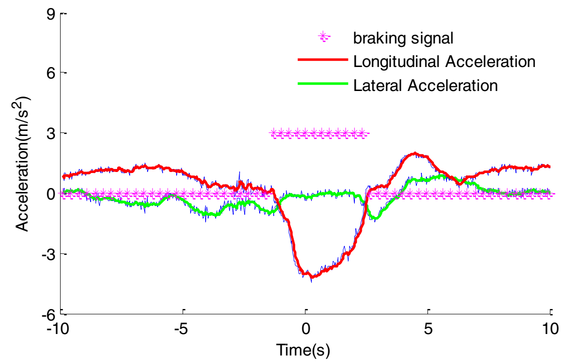

Figure 2. Example of recorded driving signals

TABLE I. SCHEDULE OF THE ENTIRE EXPERIMENT

Time period	Morning	Afternoon	Night
Hours	140	220	50

TABLE II. ROAD TYPES IN THE EXPERIMENT

Road type	1	2	3	4
Kilometres	1800	1210	4100	1650

1: Highway, 2: City ring road, 3: Inner-city road, 4: Rural road

C. Hand Labelling of Near-crash Database

Altogether, we obtained 912 near-crash cases throughout the 60-day naturalistic driving experiment with the 31 drivers. Deciding the protocol of labelling the multi-modal information is critical in properly associating the near-crash driving situation with the recorded driving state signal and video. A novel data transcription protocol that considers a comprehensive cross section of the factors that could affect the drivers and their responses is proposed in this study. The proposed protocol comprises the following five major categories:

1) Vehicle status
2) Potential crash object
3) Driving environment and road type
4) Weather condition
5) Driver information and driver actions

The designed transcription protocol is comprehensive and contains important attributes that describe the potential factors contributing to the driving risk, providing potential for analysing the relationship among the driving risk, driver/vehicle characteristics and road environment. Graduate students with driving license served as volunteer taggers, who manually labelled the recorded 912 near-crash cases according to the designed transcription protocol. Finally, we developed a near-crash database. TABLE III lists the definition of the transcription protocol.

TABLE III. DEFINITION OF TRANSCRIPTION PROTOCOL

Variable	Code	Type	Description
Vehicle Status			
Velocity when braking	V_BRA	Continuous	The vehicle speed when the driver triggers the braking signal or the turn point of acceleration signal (m/s)
Maximum deceleration	D_MAX	Continuous	The maximum deceleration during the emergency braking process (m/s^2)
Time interval of braking	T_IN	Continuous	The time interval between the braking signal trigging and the time point of maximum deceleration
Velocity Reduction	V_RED	Continuous	The vehicle speed reduction form the braking signal trigging to the time point of maximum deceleration
Vehicle status before braking process	V_STA	Qualitative	1, Deceleration process; 2, Acceleration process; 3, Constant speed;
Vehicle maneuver	V_MAN	Qualitative	1, Straight going; 2, Right turn; 3, Left turn; 4, Lane change; 5, Others
Potential crash object			
Crash Object Type	O_TYP	Qualitative	1, Vehicle; 2, Single-track vehicle (motorcycle and bicycle); 3, Pedestrian; 4, Others
Potential crash type	P_CRA	Qualitative	1, Rear end; 2, Conflict during intersection; 3, Jump out; 4, Opposite driving conflict; 5, Cut-in conflict; 6, Others
Triggering factors	T_FAC	Qualitative	0, Non-host vehicle factors; 1, Traffic light; 2, Lane reduction; 3, Lane change; 4, Collision avoidance; 5, Others
Driving environment and road type			
Near crash location	N_LOC	Qualitative	1, Intersection; 2, Non-intersection
Road type	R_TYP	Qualitative	1, Structure road; 2, Normal road; 3, Hybrid road; 4: Rural road
Parking vehicle along the road side	P_PLA	Qualitative	0, No; 1,Yes
Barriers for the opposing traffic flow	B_TRA	Qualitative	0, No; 1,Yes
Barriers for vehicles and pedestrian	B_VEH	Qualitative	0, No; 1, Yes
Weather Condition			
Weather	WEA	Qualitative	1, Sunny;2:Cloudy; 3: Others
Light condition	L_CON	Qualitative	1, Lightness; 2, Little dim;
Driver information and actions			
Gender	GEN	Qualitative	1, Male; 2, Female
Age	AGE	Continuous	The driver's age (years)
Time span with driving license	T_DIR	Continuous	The time period that owning the valid driving license
Steering light	S_LIG	Qualitative	0, No; 1, Yes
Vehicle horns	V_HON	Qualitative	0, No; 1, Yes
Second Task	S_TASK	Qualitative	0, No; 1, Talking; 3, Others

III. DEFINITION AND CLUSTER OF DRIVING RISK

A. *Definition of Driving Risk*

In this paper, driving risk is defined as a potential threat that may cause vehicle crashes or other accidents. Usually, the consequence of driving risk that involves the driver is mainly reflected by the emergency braking operation. Hence, the driving risk level can be represented by the braking process characteristics. Fig. 3 shows the key point used to define a typical deceleration curve of the braking process. The following three features are adopted to represent the driving risk level involved in a typical near-crash case during naturalistic driving:

(1) Maximum deceleration during the braking process a_{min}.

(2) Average deceleration $a_{average}$ from the braking triggering point t_0 to the point of maximum deceleration t_1

(3) Percentage reduction in the vehicle kinetic energy η_E from the braking triggering point t_0 to the point of maximum deceleration t_1

The average deceleration $a_{average}$ can be calculated by the following formula:

$$a_{average} = \frac{1}{t_1 - t_0}\int_{t_0}^{t_1} a(t)dt = \frac{1}{t_1 - t_0}[v(t_1) - v(t_0)]. \quad (1)$$

where $v(t), a(t)$ denotes the vehicle velocity and acceleration respectively. The percentage reduction in the vehicle kinetic energy η_E can be calculated as following:

$$\eta_E = \frac{\frac{1}{2}mv^2(t_0) - \frac{1}{2}mv^2(t_1)}{\frac{1}{2}mv^2(t_0)} = 1 - \left[\frac{v(t_1)}{v(t_0)}\right]^2 \quad (2)$$

where m denote the vehicle mass.

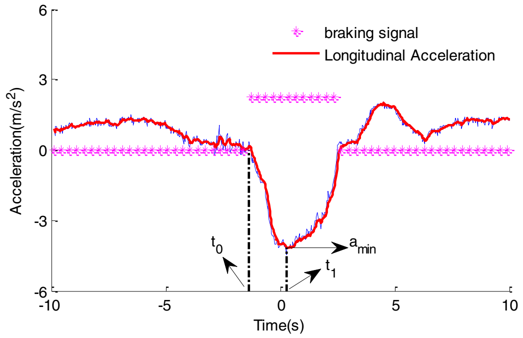

Figure 3. Key features of driving-risk level

B. *Cluster for Driving Risk*

The main criterion in evaluating the driving risk level is the braking process feature, defined as

$$X = \left[\, a_{min}, a_{average}, \eta_E \right]^T. \quad (3)$$

Cluster analysis is a valid and objective approach to classify driving risks in different near crashes into different risk levels and has been used in individual driver risk assessment research [12]. The K-means cluster method, which is popular for cluster analysis in data mining, is employed to classify the driving risks involved in different near-crash cases into different risk groups based on the feature X. Using a pre-determined number of clusters, the K-means cluster method partitions the observations into k clusters, where each observation belongs to a cluster whose mean is closest to its value [14]. The K-means method minimises the within-cluster sum of squares:

$$\arg\min_S \sum_{i=1}^{k} \sum_{X_j \in S_i} \left\| X_j - \mu_i \right\|^2, \quad (4)$$

where $X = [X_1, X_3, \cdots, X_n]$ is the set of observed data, which represents the braking process feature $X_i = \left[\, a_{min}, a_{average}, \eta_E \right]_i^T$ in the context of this paper; $S = [S_1, S_3, \cdots, S_n]$ represents the set of k clusters and μ_i denotes the mean point of cluster set S_i.

The driving risk level under each near-crash case is classified into one of the three clusters: 1) low-risk group, 2) moderate-risk group, 3) high-risk group. Near crashes in the clusters with the highest maximum deceleration are considered to be high driving-risk group. The output of the cluster analysis is shown in Fig. 4. TABLE V summarises the statistical characteristics of these three driving-risk groups. The number distribution of the different risk groups follows a pyramid structure, which means that the high-risk group has minimum near-crash cases, whereas the low-risk group has the largest number of near-crash cases. We can see that the maximum deceleration of the high-risk group is more than two times that of the low-risk group, and the maximum deceleration of the moderate-risk group is also much higher than that of the low-risk group.

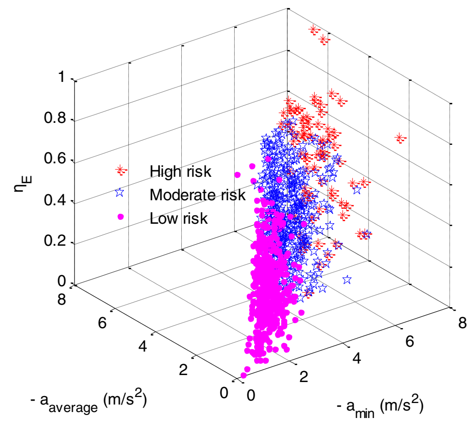

Figure 4. K-means cluster results

IV. Cluster Result Analysis

A. *Data Distribution on Driving risk Level*

According to the proposed data transcription protocol (shown in TABLE III.), the distribution of near-crash cases on the driving risk levels in terms of 19 potential risk factors is shown in TABLE IV.

The frequency information listed in TABLE IV indicates that traffic light in the fifth potential risk variable T_FAC is an important factor on the driving risk level because a relatively high proportion of near-crash cases caused by sudden changes of the traffic light occurs in the moderate- and high-risk groups (55.3% and 35.0%, respectively). Meanwhile, the proportion of near-crash cases caused by other triggering factors conforms to the overall distribution of the driving-risk levels. From the sixth potential risk variable N_LOC, we can find similar statistical results where near-crash cases that occur at the intersection are relatively higher in moderate- and high-risk groups (44.5% and 12.6%, respectively) than those outside the intersection area (38.0% and 5.2%, respectively). The other meaningful findings listed in TABLE IV show that the higher the braking speed is, the higher is the proportion of near-cash cases in the moderate- and high- risk groups. The proportions in the moderate- and high-risk groups are, respectively, 46.4% and 13.9% when the speed at the braking point lies in the range from 10 to 20 m/s, whereas those when the speed at the braking point lies from 0 to 10 m/s are, respectively, 34.7% and 2.8%, as shown in the 19th potential risk variable V_BRA.

B. *Risk factors affecting Driving Risk*

In this section, the variable importance obtained form decision tree is used to quantify the influence of potential risk factors on driving risk level. This analysis is performed using SPSS software. For detail description of decision tree and variable importance ranking, please refer to [15].

TABLE VI lists the normalized importance of the potential risk factors. We can easily see that the two variables, namely, velocity when braking (V_BRA), triggering factor (T_FAC), have the largest influence on the driving-risk level, which are conformed to aforementioned analysis.

TABLE IV. Distribution of driving-risk levels by potential risk variables

Num	Variable Code	Description	Count	Driving risk level LR 52.0%	MR 40.2%	HR 7.8%
1	V_STA	Deceleration process	265	48.7%	43.8%	7.5%
		Acceleration process	531	55.4%	37.5%	7.2%
		Constant speed	116	44.0%	44.8%	11.2%
2	V_MAN	Straight going	778	51.0%	40.4%	8.6%
		Right turn	38	65.8%	31.6%	2.6%
		Left turn	41	65.9%	34.1%	0.0%
		lane change	46	45.7%	50.0%	4.3%
		Other	9	44.4%	44.4%	11.1%
3	O_TYP	Vehicle	596	55.0%	40.4%	4.5%
		Single-track vehicle	98	72.4%	21.4%	6.1%
		Pedestrian	69	60.9%	37.7%	1.4%
		Others	149	22.1%	53.0%	24.8%
4	P_CRA	Rear end	349	51.3%	45.0%	3.7%
		Conflict during intersection	70	61.4%	32.9%	5.7%
		Jump out	65	60.0%	36.9%	3.1%
		Opposite driving conflict	46	67.4%	28.3%	4.3%
		Cut-in conflict	191	63.4%	30.4%	6.3%
		Others	191	63.4%	30.4%	6.3%
5	T_FAC	Non-host vehicle factors	723	57.7%	37.6%	4.7%
		Traffic light	103	9.7%	55.3%	35.0%
		Lane reduction	9	77.8%	22.2%	0.0%
		Lane change	33	48.5%	48.5%	3.0%
		Collision avoidance	26	57.7%	42.3%	0.0%
		Others	18	57.7%	42.3%	0.0%
6	N_LOC	Intersection	317	42.9%	44.5%	12.6%
		Non-intersection	595	56.8%	38.0%	5.2%
7	R_TYP	Structured road	285	46.0%	43.5%	10.5%
		Normal road	238	46.2%	43.3%	10.5%
		Hybrid road	251	62.5%	31.9%	5.6%
		Rural road	138	55.1%	43.5%	1.4%
8	P_VEH	No	586	48.1%	42.8%	9.0%
		Yes	326	58.9%	35.6%	5.5%

Num	Variable Code	Description	Count	Driving risk level LR 52.0%	MR 40.2%	HR 7.8%
9	B_TRA	No	372	58.9%	37.4%	3.8%
		Yes	540	47.2%	42.2%	10.6%
10	B_VEH	No	537	55.9%	37.8%	6.3%
		Yes	375	46.4%	43.7%	9.9%
11	WEA	Sunny	727	51.2%	41.3%	7.6%
		Cloudy	147	55.8%	34.7%	9.5%
		Others	38	52.6%	42.1%	5.3%
12	L_CON	Lighted	796	52.3%	40.2%	7.5%
		Slightly dim	116	50.0%	40.5%	9.5%
13	GEN	Male	661	51.0%	40.5%	8.5%
		Female	251	54.6%	39.4%	6.0%
14	AGE	≤ 30	145	50.3%	41.4%	8.3%
		31-40	291	54.0%	39.9%	6.2%
		41-50	232	48.7%	40.9%	10.3%
		51-60	202	56.9%	35.6%	7.4%
		$60 \leq$	42	38.1%	57.1%	4.8%
15	T_DIR	≤ 10	305	50.8%	40.0%	9.2%
		11-20	380	56.1%	37.1%	6.8%
		21-30	157	46.5%	44.6%	8.9%
		$30 \leq$	70	47.1%	48.6%	4.3%
16	S_LIG	No	784	51.3%	40.3%	8.4%
		Yes	128	56.3%	39.8%	3.9%
17	V_HON	No	859	51.1%	41.0%	7.9%
		Yes	53	66.0%	28.3%	5.7%
18	S_TASK	No	784	52.0%	40.4%	7.5%
		Talking	125	51.2%	39.2%	9.6%
		others	3	66.7%	33.3%	0.0%
19	V_BRA	$(0,10]$	501	62.5%	34.7%	2.8%
		$(10,20]$	388	39.7%	46.4%	13.9%
		$(20,+\infty)$	23	30.4%	56.5%	13.0%

Note: Num denotes the index of potential risk variables, and LR: low-risk group; MR: moderate-risk group; HR: high-risk group

TABLE V. CHARACTERISTIC OF DRIVING RISK GROUPS

Risk groups	Number of near crash cases	Percentage	Mean of features of the braking process		
			a_{min} (m/s^2)	$a_{average}$ (m/s^2)	η_E
Low-risk group	474	52.0 %	-1.931	-1.027	30.9 %
Moderate-risk group	367	40.2 %	-3.278	-1.717	56.6 %
High-risk group	71	7.8 %	-5.385	-3.125	66.1 %

TABLE VI. IMPORTANCE OF THE POTENTIAL FACTORS

Variables	Normalised importance
V_BRA	100.0%
T_FAC	96.7%
O_TYP	82.9%
P_CRA	75.9%
AGE	11.7%

1) Velocity when Braking

As shown in TABLE VI, velocity when braking (V_BRA) is the most important potential risk variable. Intuitively, the higher the vehicle speed is, the greater is the kinetic energy of the lone-driver-vehicle system. If a potential threat or sudden change of the object status occurs in the driving environment, the lone-driver-vehicle system will become more unstable and risky, meaning that the driving risk level involved in a near-crash case will be high as the vehicle velocity increases. Some other research studies have indicated that driving speed is an important factor in road safety [16]. Elvik et.al pointed out that speed not only affects the severity of a crash but also is related to the risk of being involved in a crash [17].

2) Triggering Factors

Triggering factor (T_FAC) is the second important potential variable with a 96.7% normalized importance. TABLE VI shows that traffic light in the fifth potential risk factor variable T FAC has significant effect on the driving risk level involved in near-crash cases because a relatively high proportion of near-crash cases caused by sudden changes in the traffic light occurs in moderate- or high-risk groups (55.3% and 35.0%, respectively). This result agrees with other previous studies in vehicle accidents that result from dilemma zone at signalized intersection [11][18].

V. CONCLUSIONS

In this study, we obtained 912 near-crash cases through a 60-day naturalistic driving experiment employing 31 drivers. A comprehensive transcription protocol, which contains important attributes describing the conditions contributing to driving risk, was designed to provide the possibility of analyzing the relationship between driving risk, driver/vehicle characteristics and road environment. In this paper, the driving risk level under near-crash cases has been represented by the braking process characteristics, namely, 1) maximum deceleration, 2) average deceleration and 3) percentage reduction in the vehicle kinetic energy. K-means cluster is used to cluster different driving risk levels involved in a near-crash case using the braking process features. The results indicate that the velocity when braking and triggering factors have the largest influence on the driving risk level, which, to some extent, are in accordance with some previous studies.

REFERENCES

[1] D. Caveney, "Cooperative vehicular safety applications," *Control Systems, IEEE*, 2010, pp. 38-53.

[2] M. Nagai, "The perspectives of research for enhancing active safety based on advanced control technology," *Vehicle System Dynamics*, 2007, pp. 413-431.

[3] D. Zhang, K. Li, J. Wang, "A curving ACC system with coordination control of longitudinal car-following and lateral stability," *Vehicle System Dynamics*, 2012, pp. 1085-1102.

[4] A. Jarašūniene, G. Jakubauskas, "Improvement of road safety using passive and active intelligent vehicle safety systems," *Transport*, 2007, pp. 284-289.

[5] DTM-China (Ministry of Public Security, Department of Traffic Management). Annual report of road traffic accidents statistics in P.R. China. Scientific Research Institute of Traffic Management, Ministry of Public Security, Beijing (in Chinese). 2010.

[6] D. Lord, F. Mannering, "The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives," *Transportation Research Part A: Policy and Practice*, 2010, pp. 291-305.

[7] A. S. Al-Ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity," *Accident Analysis & Prevention*, 2002, pp. 729-741.

[8] L.Y. Chang, H.W. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques," *Accident Analysis & Prevention*, 2006, pp. 1019-1027.

[9] S.G. Klauer, T.A. Dingus, V.L. Neale, et al. "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data," 2006.

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

[11] G.S. Aoude, V.R. Desaraju, L.H. Stephens, "Driver behavior classification at intersections and validation on large naturalistic data set," *Intelligent Transportation Systems, IEEE Transactions on*, 2012, pp. 724-736.

[12] F. Guo, Y. Fang, "Individual driver risk assessment using naturalistic driving data," *Accident Analysis & Prevention*, 2013, pp. 3-9.

[13] F. Guo, S.G. Klauer, J. M. Hankey, "Near crashes as crash surrogate for naturalistic driving studies," *Transportation Research Record: Journal of the Transportation Research Board*, 2010, pp. 66-74.

[14] L. Kaufman, P. J. Rousseeuw, "Finding groups in data: an introduction to cluster analysis," John Wiley & Sons, 2009.

[15] L. Breiman, J. Friedman, C.J. Stone, et al. "Classification and regression trees," *CRC press*, 1984.

[16] L. Aarts, I. Van Schagen, "Driving speed and the risk of road crashes: A review," *Accident Analysis & Prevention*, 2006, pp. 215-224.

[17] R. Elvik, P. Christensen, A. Amundsen, "Speed and road accidents. An evaluation of the Power Model," *TØI report*, 2004.

[18] H.Rakha, I. El-Shawarby, J.R. Setti, "Characterizing driver behavior on signalized intersection approaches at the onset of a yellow-phase trigger," Intelligent Transportation Systems, IEEE Transactions on, pp. 630-640.