Accident Analysis and Prevention 95 (2016) 350–357

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Driver drowsiness detection based on non-intrusive metrics considering individual specifics

Xuesong Wang[a,b], Chuan Xu[c,d,*]

[a] School of Transportation Engineering, Tongji University, Shanghai 201804, China
[b] Road and Traffic Key Laboratory, Ministry of Education, Shanghai 201804, China
[c] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 610031, China
[d] National United Engineering Laboratory of Integrated and Intelligent Transportation, Chengdu 610031, China

ARTICLE INFO

Article history:
Received 16 March 2015
Received in revised form 1 August 2015
Accepted 7 September 2015
Available online 1 October 2015

Keywords:
Drowsiness detection
Multilevel ordered logit model
Non-intrusive
Driving behavior
Eye feature
Driving simulator

ABSTRACT

Objectives: Drowsy driving is a serious highway safety problem. If drivers could be warned before they became too drowsy to drive safely, some drowsiness-related crashes could be prevented. The presentation of timely warnings, however, depends on reliable detection. To date, the effectiveness of drowsiness detection methods has been limited by their failure to consider individual differences. The present study sought to develop a drowsiness detection model that accommodates the varying individual effects of drowsiness on driving performance.
Methods: Nineteen driving behavior variables and four eye feature variables were measured as participants drove a fixed road course in a high fidelity motion-based driving simulator after having worked an 8-h night shift. During the test, participants were asked to report their drowsiness level using the Karolinska Sleepiness Scale at the midpoint of each of the six rounds through the road course. A multilevel ordered logit (MOL) model, an ordered logit model, and an artificial neural network model were used to determine drowsiness.
Results: The MOL had the highest drowsiness detection accuracy, which shows that consideration of individual differences improves the models' ability to detect drowsiness. According to the results, percentage of eyelid closure, average pupil diameter, standard deviation of lateral position and steering wheel reversals was the most important of the 23 variables.
Conclusion: The consideration of individual differences on a drowsiness detection model would increase the accuracy of the model's detection accuracy.

1. Introduction

Drowsy driving is a serious threat to road safety. In 2009, over 800 fatalities and 30,000 injuries from car crashes were attributed to drowsy driving in the United States (NHTSA, 2011). In Europe, an estimated 20% of traffic crashes are caused by drowsy driving (Maycock, 1997). The problem seems more serious in China where, in 2007, 1768 fatalities were attributed to drowsy driving (Road and Transport Authority, 2009). Considering that, as of 2012, China had over 85,000 km of expressways (Central Intelligence Agency, 2010), and the annual growth in kilometers exceeded 14% during 2002–2011, the problem requires immediate attention. A recent study using a naturalistic driving method estimated the increase in crash risk associated with drowsy driving to be four to six times greater than when driving while alert (Klauer et al., 2006). Besides increasing crash risk, drowsy driving crashes are often more severe than other crashes because they frequently occur on high speed expressways, and are frequently run-off-the-road crashes with no braking prior to impact.

Unlike drinking and driving, drowsy driving does not provide an objective measure of its occurrence, and therefore enforcement cannot be used to counter this problem (Radun et al., 2012). An alternative would be to notify the driver if he or she becomes too drowsy to drive safely. This requires the reliable detection of drowsy driving – a problem that has been extensively researched. Based on the type of data used, drowsiness detection can be conveniently separated into the two categories of intrusive and non-intrusive methods. Intrusive methods, such electroencephalograms (EEGs) (Li et al., 2012) or electrocardiograms (EKGs)

[*] Corresponding author at: School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 610031, China.
E-mail address: xuchuan@home.swjtu.edu.cn (C. Xu).

http://dx.doi.org/10.1016/j.aap.2015.09.002
0001-4575/© 2015 Elsevier Ltd. All rights reserved.

(Patel et al., 2011), show good detection accuracy, however, are limited to the research laboratory. In contrast, methods based on non-intrusive measures detect drowsiness by measuring driving behavior and sometimes eye features, and so are useful for real world driving situations.

To date, non-intrusive methods have been less reliable than intrusive methods, partly because individual driver differences in non-intrusive measures have prevented identification of a common point at which drowsiness impairs driving (Ingre et al., 2006b; Jo et al., 2014). In the previous research, individual differences of non-intrusive methods are frequently mentioned, for both driving behavior and eye features. For driving behavior, it was reported that an important measure to drowsiness is the standard deviation of lateral position (SDLP), and this research found that, for the same drowsiness level, different drivers have different SDLPs. (Ingre et al., 2009). Another study finds individual differences in driver's lane departure behavior (Ingre et al., 2006a), and individual differences in the standard deviation of steering wheel movements is also reported (Thiffault and Bergeron, 2003). Regarding eye features, individual differences of blink duration (Ingre et al., 2009; Hamada et al., 2003) and percentage of eyelid closure (PERCLOS) (Wierwille et al., 1994) were also observed in many studies. However, most drowsiness detection methods such as decision trees, logistic regression, Bayesian networks (Yang et al., 2010), artificial neural networks (Patel et al., 2011), and support vector machines (Hu and Zheng, 2009) have not properly handled the problem of differences in the manifestation of drowsiness among individuals. Ignoring such differences reduces the accuracy and reliability of these models, especially for the important non-intrusive measures.

An ideal fatigue-detection system should rank favorably on several specific non-intrusive criteria (Balkin et al., 2011). To address and increase the accuracy for a drowsiness detection model based on non-intrusive measures, a multilevel logit model was built based on both driving behavior measures and eye features, which detect drowsiness by using individual-specific criteria. To compare detection accuracy, two non-individual specific models (i.e., ordered logit model and artificial neural network) were established. Driving behavior, eye features, and subjective drowsiness were scaled and collected in a high-fidelity driving simulator experiment.

2. Method

2.1. Participants

Sixteen male participants aged 24–40 (mean 32.8, SD 5.0) with valid Chinese drivers licenses were recruited from students (two) and staffs (fourteen) at Tongji University. They were required to be in good health, have no sleep related disorders, and not to have taken any pharmaceuticals within one month prior to entering the study. Subjects who had a history of motion sickness were screened out. All subjects provided written consent and were paid about 200 RMB Yuan, depending on the total time in the laboratory. During the experiment, one subject's eye movement data was lost because of a technical problem. One subject fell asleep before completing the task, however, his data up to that point was used.

2.2. Apparatus

The Tongji University driving simulator is shown in Fig. 1. This simulator, currently the most advanced in China, incorporates a fully instrumented Renault Megane III vehicle cab in a dome mounted on an 8 degree-of-freedom motion system with an *X*–*Y* range of 20 m × 5 m. An immersive 5 projector system provides a front image view of $250^\circ \times 40^\circ$ at 1000 × 1050 resolution refreshed at 60 Hz. LCD monitors provide rear views at the central and side mirror positions. For this study, SCANeR^{TM} studio software represented the simulated roadway and controlled a force feedback system that acquired data from the steering wheel, pedals and gear shift lever.

Eye movement data were recorded using a Smarteye® eye tracking system. The system uses four cameras located in the front of the vehicle to record the driver's eye movements at a 60 Hz sampling rate.

2.3. Procedure

2.3.1. Experiment design

All participating drivers were presented with the same conditions in the same order. The driving course, diagrammed in Fig. 2, simulated a 20 km rural highway circle with six lanes and 3.75 m width: composed of the straight segments numbered 1, 3, 5, 9, 11, 13; two circle curves each with a 700-m radius (segments numbered 7, 15); and several transition curves (numbered 2, 4, 6, 8, 10, 12, 14, 16). The length of each straight line was 2 km, and only the straight line road segments were used for the analysis. Grass and trees were placed beside the highway, as well as a few small villages along the straight segments.

To assess driving performance at high levels of drowsiness, night shift workers were selected and tested just after shift completion around 8:00 a.m. Upon arrival at the simulator facility, participants were asked to complete questionnaires on their basic driver information and current levels of drowsiness. To increase the reliability of their self-assessment of drowsiness levels, drivers were provided with clear explanations of the Karolinska Sleepiness Scale (KSS). Then drivers spent 5 min familiarizing themselves with driving the simulator. At about 8:30 a.m., they received the main test, in which each subject was asked to drive and respect road rules for 1 h.

In order to induce drowsiness,

- The driving task was reduced to one lane, eliminating the need to change lanes;
- Drivers were required to drive at a constant speed (120 km/h), eliminating the need for manual gear changes;
- No radio or music was played;
- No environmental disturbances (e.g., crosswinds) were introduced;
- All driving was during daytime periods and no tunnel or weather changes occurred, eliminating the need to adjust headlights;
- Only occasional and uneventful traffic was present.
- After the main test, participants were asked to complete post-experiment questionnaires on their levels of drowsiness.

2.3.2. Measurement of drowsiness – participants' assessments

To track drivers' drowsiness changes during the 1 h driving task, the participants were asked to report their Karolinska Sleepiness Scale (KSS) level at the midpoint of the driving course. To increase the reliability of drowsiness level, a carefully KSS explanation before the experiment was implemented to the driver. KSS uses a 9-point ordinal scale, but it is not necessary to distinguish all nine levels. It is, however, necessary to identify the drowsiness level associated with a high crash risk. Several studies (Yang et al., 2010; Åkerstedt and Gillberg, 1990) suggest that serious behavioral and physiological changes do not occur until KSS ≥ 7. In addition, the description of KSS 7 is sleep, but no effort to keep alert, while KSS 8 and 9 are described as need some or great effort to keep alert. This difference in the level of effort needed to keep alert may affect the crash risk, and justifies dividing these KSS levels. Therefore, drowsiness in this study is categorized into three drowsiness levels (DL) as follows:

- Level 1 (DL = 1): KSS range from 1 to 6, no drowsiness or low-level drowsiness;

Fig. 1. Tongji advanced driving simulator.

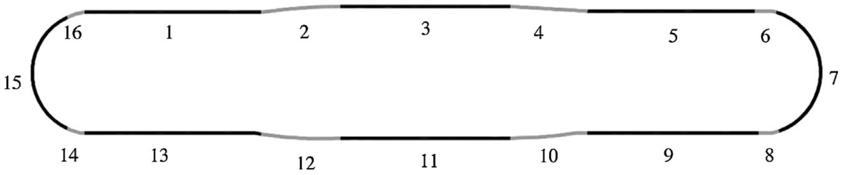

Fig. 2. Ring shaped highway and segment ID numbers.

- Level 2 (DL = 2): KSS is 7, moderate-level drowsiness;
- Level 3 (DL = 3): KSS range from 8 to 9, high-level drowsiness.

2.3.3. *Measurement of the effects of drowsiness*

Vehicle-based signals measuring driving behavior, with a 10 Hz sample frequency, were obtained from the system, including vehicle speed, lateral position and steering wheel angle. Based on these signals, several drowsiness-related behavior indicators were extracted. Eye movement signals including eyelid opening and pupil diameter were also recorded, with a 60 Hz sample frequency. Using the Smarteye® Pro software, eye blink was identified. Eye activity indicators including percentage of eye closure (PERCLOS), average pupil diameter, blink frequency, and average blink duration were calculated. These measures are summarized in Table 1.

2.3.4. *Data analysis*

The unit for analysis is each straight line segment mentioned in Section 2.3.1. Using DL as the dependent variable, and the driving behavior and eye feature metrics described in Table 1 as the independent variables, three models were used: the individual-specific multilevel ordered logit (MOL) model, and two non-individual-specific models, an ordered logit (OL) model and an artificial neural network (ANN) model. In drowsiness detection problem, OL is the basic statistic model modeling the relationship between discrete dependent variable DL and independent variables. ANN is a main method to solve classification problem and widely applied in drowsiness detection problem. However, individual difference was not considered in the two models. To cover the shortage of the above models, MOL was built to detect drowsiness considering individual specifics.

A feed-forward neural network with five units in one hidden layer consisting of the interconnection of neurons only between two adjacent layers was built, and a back propagation training method was applied using IBM SPSS. Type of training is batch and optimization algorithm is scaled conjugate gradient. The total dataset was divided into training (398 samples) and testing (170 samples) datasets to prevent overfitting. The data in the training set was used to build the models, and the data in the validation set was used to test the models. The stopping criterion is one consecutive step with no decrease in error calculated by testing sample. A hyperbolic tangent function was used as the activation function of the hidden layer and a Softmax activation function was used for the output layer.

To ensure each of the two data sets contained the data from every subject at every drowsiness level, the training set was generated by randomly selecting 70% of the data for each subject at each drowsiness level, and the rest of the data was assigned to the validation set. The individual specific multilevel ordered logit model was established first, and then the non-individual specific ordered logit model and neural network model were constructed using the same variables as the multilevel ordered logit model.

3. Results

In the MOL model development, each of the 23 variables was tested for statistical significance and the nonsignificant variables were eliminated. Among those studied, five variables were identified as significant, as judged by 95% credible interval (CI): PERCLOS, Pupil, Blink duration, SWM_Re, LP_stdev. Then, among the significant variables, the Pearson correlation coefficients were examined. Blink durations were highly correlated with PERCLOS, but PERCLOS is more significantly related to DL. Therefore, in the final model, Blink duration was also eliminated.

The results of the MOL model are shown in Table 2. Two eye feature metrics, one steering variability metric and one lane variability metric are used as the explanatory variables in the final model. Among these explanatory variables, the fixed effects of PERCLOS, LP_stdev and SWM_Re are positive, while Pupil was negative. The threshold γ_1 is significant ($t = -2.255$, sig. = 0.025) and γ_2 is nonsignificant ($t = -0.427$, sig. = 0.669). For the random effects of

Table 1
Driving behavior and eye feature metrics.

Metrics	Description of the variables	Mean	S.D.
Driving behavior			
LP_stdev	Standard deviation of lateral position (m)	0.306	0.134
LP_avg	Average of lateral position (m)	0.214	0.269
LD_Area[a]	Sum of lane departure time-space area (m s)	1.627	4.207
LD_TArea[b]	Sum of lane departure time-space area weighted by lane crossing time (m s)	6.129	35.383
LD_Frequency	Lane departure frequency	0.660	1.118
LD_Speed	Lane departure lateral speed (m/s)	0.046	0.099
LD_Tc	Time percentage of lane crossing of the vehicle center	0.002	0.017
LD_Te	Time percentage of lane crossing of the vehicle edge	0.021	0.045
SW_Speed_stdev	Standard deviation of steering angular speed (degree/s)	0.012	0.008
SW_Area_MA[c]	Area surrounded by steering angle and its moving average	0.440	0.257
SWM_Re	Steering wheel reversals	190.485	29.522
SW_Range_1	Percentage of steering speed in 0–2.5 degree/s	0.876	0.085
SW_Range_2	Percentage of steering speed in 2.5–5 degree/s	0.077	0.037
SW_Range_3	Percentage of steering speed in 5–7.5 degree/s	0.024	0.020
SW_Range_4	Percentage of steering speed in 7.5–10 degree/s	0.010	0.012
SW_Range_5	Percentage of steering speed exceeding 10 degree/s	0.013	0.024
Speed	Average speed (km/h)	117.424	6.650
Speed_stdev	Standard deviation of speed (km/h)	2.866	2.860
Speeding_T	Time percentage of speed exceeding the limit speed 120 km/h	0.311	0.372
Eye features			
Blink_Frequency	Average blink frequency per second	0.504	0.318
Blink_duration	Average blink duration (s)	0.402	0.054
PERCLOS	Percentage of eyelid closure	0.132	0.099
Pupil	Average pupil diameter (mm)	3.807	0.894

[a] $A=\sum_{i=1}^{n} A_i, A_i=\int d_i(t)\Delta t$. While A is LD_Area, and $d_i(t)$ is the lane crossing distance of ith lane crossing at t moment.

[b] $T=\sum_{i=1}^{n} T_i, T_i=t_i\int d_i(t)\Delta t$. While T is LD_TArea, and t_i is the lane crossing duration of ith lane crossing.

[c] $M=\int|S(t)-S_{MA}(t)|\Delta t$. While M is SW_Area_MA, $S(t)$ is the steering wheel angle at t moment, $S_{MA}(t)$ is the moving average of steering wheel angle at t moment, and the relationship between $S_{MA}(t)$ and $S(t)$ is: $S_{MA}(0)=S(0)$, $S_{MA}(t+\Delta t)=\alpha\times S_{MA}(t)+(1-\alpha)\times S(t+\Delta t)$. In order to smooth steering wheel angel, α is 0.95 in this study.

Table 2
MOL and OL model estimated results.

Parameters	Effect estimate		t	Sig.	95% confidence level	
	Mean	S.D.			Lower	Upper
MOL						
Threshold						
γ_1	−3.778	1.675	−2.255	0.025	−7.070	−0.486
γ_2	−0.712	1.666	−0.427	0.669	−3.986	2.562
Fixed effects						
PERCLOS	5.226	2.129	2.455	0.014	1.043	9.410
Pupil	−1.780	0.318	−5.60	0.000	−2.403	−1.156
SWM_Re	0.010	0.005	2.182	0.030	0.001	0.020
LP_stdev	5.227	1.007	5.193	0.000	3.249	7.206
Random effects						
Between-subject variance	3.496					
Within-subject variance	5.195					
ICC	0.402					
OL						
Threshold						
γ_1	−0.545	0.694	−0.786	0.432	−1.907	0.816
γ_2	1.914	0.698	2.742	0.006	0.544	3.283
Fixed effects						
PERCLOS	9.891	1.441	6.865	0.000	7.063	12.719
Pupil	−1.065	0.121	−8.779	0.000	−1.303	−0.827
SWM_Re	0.011	0.003	3.212	0.001	0.004	0.017
LP_stdev	3.606	0.833	4.329	0.000	1.971	5.240

individual differences, the intra-class correlation coefficient (ICC) of the data set (training set) is 0.402, which shows a large between-group heterogeneity and within-group homogeneity. Therefore, it can be inferred that if an ordered logit model were implemented without considering the random effects between subjects, the results may be biased and inaccurate. It is also implied that the drowsiness detection algorithm should vary for different subjects.

The results of the OL model using the same explanatory variables as the MOL model are also shown in Table 2. All the explanatory variables are significant at 95% CI, and the coefficient for each variable shows different values but the same sign as those in the MOL model. The threshold γ_2 is significant ($t=2.742$, sig. = 0.006) and γ_1 is nonsignificant ($t=-0.786$, sig. = 0.432).

For ANN model, PERCLOS, and Pupil, SWM_Re were standardized before input into the model. Based on the ANN model, the importance of each variable was also calculated. The normalized importance of each variable is PERCLOS (100%), Pupil (74.5%), LP_stdev (65.2%), and SWM_Re (41.1%), which implies that eye feature metrics performed better than driving behavior metrics in drowsiness detection. The receiver operating characteristic (ROC) curve of ANN was also formed. The area under the ROC curve for DL = 3 (0.887) is larger than DL = 1 (0.779) and DL = 2 (0.647), which

Table 3
Models accuracy summary.

Model and dataset	Observed	Predicted			Correct	Overall accuracy
		1	2	3		
MOL (Train set)	1	67.90%	31.00%	1.20%	67.90%	68.40%
	2	22.80%	68.90%	8.30%	68.90%	
	3	1.90%	29.60%	68.50%	68.50%	
MOL (Test set)	1	64.17%	30.28%	5.56%	64.17%	64.15%
	2	27.13%	63.33%	9.54%	63.33%	
	3	5.00%	29.68%	65.32%	65.32%	
OL (Train set)	1	59.52%	39.29%	1.19%	59.52%	54.80%
	2	40.00%	51.67%	8.33%	51.67%	
	3	3.70%	43.52%	52.78%	52.78%	
OL (Test set)	1	50.00%	50.00%	0.00%	50.00%	52.70%
	2	30.83%	52.50%	16.67%	52.50%	
	3	1.19%	42.86%	55.95%	55.95%	
ANN (Train set)	1	54.17%	40.28%	5.56%	54.17%	57.80%
	2	26.67%	63.33%	10.00%	63.33%	
	3	9.09%	36.36%	54.55%	54.55%	
ANN (Test set)	1	47.92%	52.08%	0.00%	47.92%	56.04%
	2	30.00%	65.83%	4.17%	65.83%	
	3	12.20%	36.59%	51.22%	51.22%	

means it was easier to distinguish DL = 3 than DL = 1 or 2 than to distinguish any other DL from the remaining DLs.

The summary of drowsiness detection accuracy is shown in Table 3. ANN performed better than the OL model in both the train set and test set, but for each data set (train set and test set), the overall accuracy of the MOL model is the highest among the three models. For the test set, the detection accuracy of MOL varies in a very small range across drowsiness levels (63.33–65.32%), while the other two models vary greatly (OL: 50.00–55.95%; ANN: 47.92–65.83%). The top three possible detection errors for the three models were the same, which are mistaking 1 for 2, mistaking 3 for 2 and mistaking 2 for 1. Mistaking 1 for 3 and mistaking 3 for 1 have the smallest chance of error.

4. Discussion

In order to find out a group of suitable indicators for drowsiness warning, 23 non-intrusive indicators were developed in this study. Eight metrics are based on lane lateral position, eight on steering wheel angle, three on vehicle speed and four on eye features. The MOL results indicate that the indicators group formed by PERCLOS, Pupil, SWM_Re, and LP_stdev are appropriate to detect drowsiness. To test whether the variables selected in the MOL were acceptable, all 23 variables were input into an ANN model with 10 neurons in hidden layers. In order of importance, PERCLOS, Pupil, Blink_duration, SWM_Re, and LP_Stdev were the most important predictors of DL. With the exception of Blink_duration, which was eliminated because of the correlation with PERCLOS, these variables are also significant in the MOL model.

Identification of the most important variables was vital because when detecting drowsiness, the input variables selection is an important consideration. The top three important variables are all eye feature metrics. These results can be interpreted that eye feature metrics perform better than driving behavior metrics in drowsiness detection. The following test also verified this inference. After removing eye feature metrics, the detection accuracy of the ANN model for the test set was reduced from 56.0% to 45.8%, while keeping only eye feature metrics reduced the detection accuracy to just 49.3%. Because of this, some drowsiness detection studies have used only eye feature metrics to detect drowsiness (Jo et al., 2014; Hu and Zheng, 2009). However, the detection accuracy of the ANN model using both metrics has the highest detection accuracy, suggesting that use of driving behavior metrics is also needed. Moreover, the eye features are often measured by camera and image processing, which may not be reliable. Use of the driving behavior metrics can be a supplement to increase the reliability of the detection system.

Of the driving behavior metrics, both lane related and steering related metrics are important in the drowsiness detection models. Among lane related metrics, the lane variability measure (LP_stdev) performed better than other lane departure measures. A possible reason is that the lane departure metric only measures the features of lane departure events, so the lane variability information is missed for the non-departure parts. SWM_Re is the most important variable among steering related metrics, which measures steering variability. Rapid steering wheel movement is suggested as a drowsiness measurement (Sandberg and Wahde, 2008). In this study, however, where it is measured by SWM_Rang_5, it is not significant in the MOL model. Some studies (Forsman et al., 2013) also conclude that lane variability is highly correlated with steering variability; but we calculated the Pearson correlation of SWM_Re and LP_stdev at 0.089 (sig. = 0.059), which indicates the correlation between these variables is small.

Previous research finds there might be a curvilinear relationship between KSS and drowsiness metrics (Ingre et al., 2006a), with a stronger change at high KSS levels when compared with low KSS levels. In this study, among the four significant variables in the MOL model, the mean change between DL 3 and DL 2 is larger than that between DL 2 and DL 1 (see Fig. 3). PERCLOS is a typical example of this change, by which it can be inferred that PERCLOS is sensitive to a high drowsiness level. This figure also might explain why PERCLOS is the variable with the highest importance in the ANN model. Due to the larger change of metrics on higher DLs, it can be inferred that higher drowsiness can be more easily detected, which is also verified by the detection accuracy of the MOL and OL models. Therefore, the correct detection rate for DL = 3 is the highest.

In this study, the OL model can be considered as the basic model, while the MOL and ANN models can be viewed as two improvements on the OL model. The MOL model's improvement is in considering individual differences, while the ANN's improvement is in ensuring higher adaptability for the data. The results show MOL has the highest detection accuracy among the three models, which can be attributed to using a series of individual specific thresholds, achieved by adding a random intercept in the subject

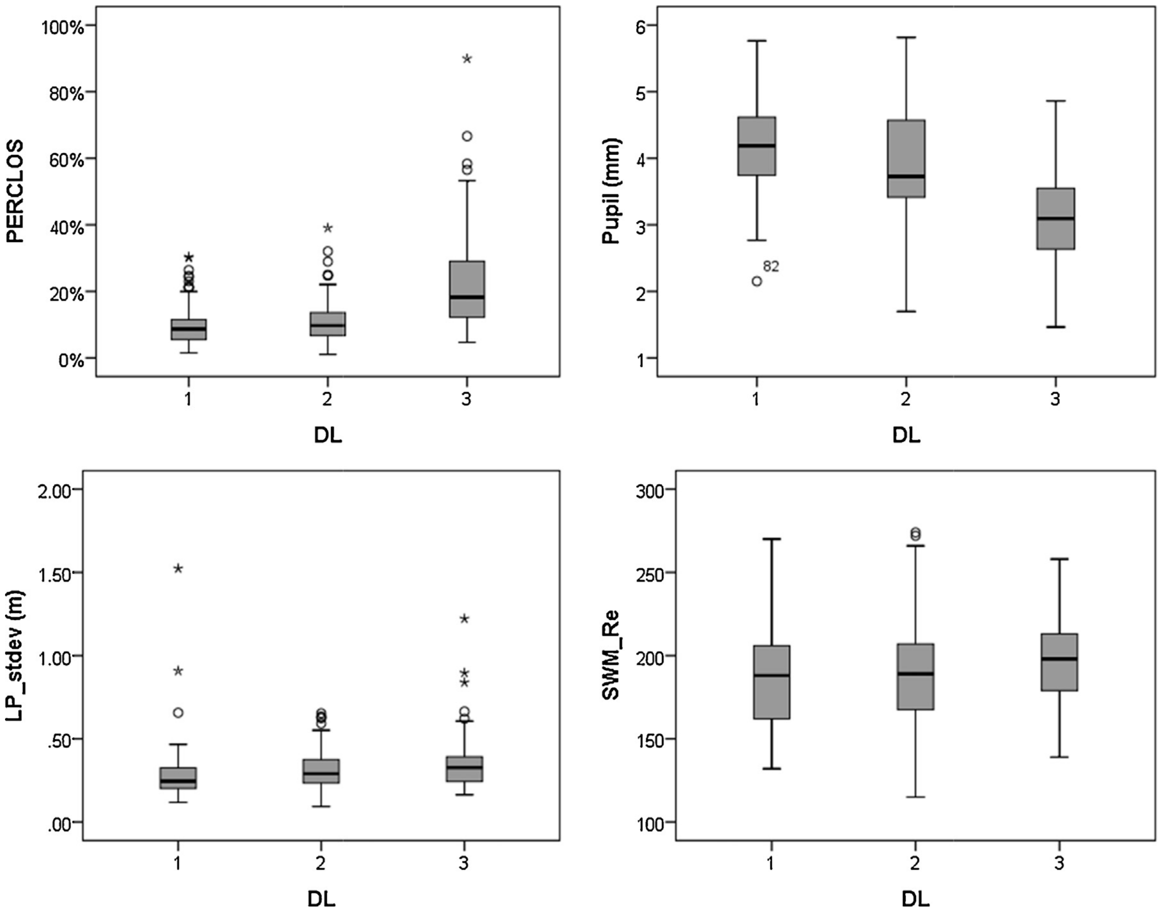

Fig. 3. Boxplot of four important variables.

level of MOL. That is, the improvement by using an individual specific logit model is larger than by using a more complex algorithm with higher flexibility. However, in MOL, this random intercept for each driver is hard to predict. It can be decided only by a procedure such as training. Driving experience, age, and other characteristic variables of the driver were analyzed as explanations of individual differences, but no strong results were found. Yet the individual difference is still the main obstacle to increasing the drowsiness detection accuracy.

Some research that has classified drowsiness in only two levels (alert or drowsy) has achieved high detection accuracy. But two levels of drowsiness are not enough to support warning the driver before the crash risk becomes critical. Therefore, three levels of drowsiness were used in this study. However, this study confirms that the detection accuracy is highly related with lower numbers of drowsiness levels. If we used two levels of drowsiness (Level 1: KSS 1–7; Level 2: 8–9), the detection accuracy increases from 64.15% to 88.6% for the MOL model with the same variables, and the detection accuracy increases from 56.04 to 83.3% for ANN. It can be inferred that the more drowsiness levels to be classified, the lower detection accuracy we would get in the models. Therefore, when comparing the detection accuracy among detection models, the way to classify drowsiness levels should also be considered.

5. Conclusion and recommendations

Twenty-three non-intrusive metrics including driving behavior and eye feature metrics were evaluated in a simulated shift-work study with a motion-based high-fidelity driving simulator in a controlled laboratory environment. Based on these metrics, an MOL was developed to detect three levels of drowsiness. For comparison, two non-individual specific models, OL and ANN, were also established.

The MOL has the highest detection accuracy. This may be attributed to using a series of individual specific criteria, which was achieved by adding a random intercept at the subject level. Among the 23 variables, PERCLOS, Pupil, LP_stdev and SWM_Re were significant in MOL and OL, and were also confirmed in ANN. Metrics of eye features performed better (showed higher importance) in the drowsiness detection models than other metrics, which was also verified using ANN by comparing the detection accuracy between eye features only and driving behaviors only. We also found higher DLs are more easily detected because of higher heterogeneity between adjacent DLs.

Based on this analysis, employing a user-specific method to detect driver drowsiness is recommended in order to address the inaccuracies caused by individual differences. In this study, group characteristics (such as age and gender) of participants were controlled. However, if it can be demonstrated that group characteristics exist, we can build group-specific models that would simplify the model training process by group determinants. Therefore group characteristics are recommended for study. Also, because establishment of drowsiness level in this research was subjective, more accurate measures should be applied, for example, using EEG to determine drowsiness level.

Acknowledgements

Supported by Road and Traffic Engineering Key Laboratory of Tongji University, Ministry of Education.

Appendix 1. Multilevel ordered logit model

Multilevel ordered logit models are often presented as cumulative logit models. Suppose an ordered DL_{ij} is the drowsiness level for *i*th subject on the *j*th road segment. A latent continuous variable DL^*_{ij} is established as the unobserved measure of DL_{ij}. DL^*_{ij} is related to DL_{ij} by a series of latent thresholds. Differing from the ordered logit model, the multilevel model accounts for each subject's individual performance by using a set of variable thresholds specific to each subject: γ_{k_i} ($k = 1, 2$), see Eq. (2).

$$\mathrm{DL}_{ij} = \begin{cases} 1 & \text{if} \quad DL^*_{ij} < \gamma_{1_i} \\ 2 & \text{if} \quad \gamma_{1_i} < \mathrm{DL}^*_{ij} < \gamma_{1_i} \\ 3 & \text{if} \quad \gamma_{2_i} < \mathrm{DL}^*_{ij} \end{cases} \tag{1}$$

The DL^*_{ij} can be written in the same form as the regular linear regression model.

$$\mathrm{DL}^*_{ij} = \theta_{ij} + \varepsilon_{ij} \quad \text{and} \quad \theta_{ij} + \sum_{p=1}^{p} \beta_p x_{p_{ij}} \tag{2}$$

where $x_{p_{ij}}$ is the explanatory variable for *i*th subject on *j*th segment. ε_{ij} is the disturbance term, which is assumed as a logistic distribution as the cumulative density function. Thus, the cumulative response probabilities of the ordinal DL may be denoted as:

$$P_{ij(k)} = \Pr(\mathrm{DL}^*_{ij} \le k) = F(\gamma_{ki} - \theta_{ij}) = \frac{\exp(\gamma_{k_i} - \theta_{ij})}{1 + \exp(\gamma_{k_i} - \theta_{ij})}, \quad k = 1, 2 \tag{3}$$

$$\mathrm{Logit}(P_{ij(k)}) = \log\left[\frac{P_{ij(k)}}{1 - P_{ij(k)}}\right] = \log\left[\frac{Pr(\mathrm{DL}^*_{ij} \le k)}{Pr(\mathrm{DL}^*_{ij} \ge k)}\right] = \gamma_{k_i} - \theta_{ij}, \quad k = 1, 2 \tag{4}$$

In order to accommodate differences among subjects, the thresholds γ_{ki} were specified as random effects.

$$\gamma_{k_i} = \gamma_k + b_i, \quad k = 1, 2 \tag{5}$$

where the intercept γ_k represents a constant component for thresholds for all subjects. A random effect component b_i is formulated to accommodate the between-subject heterogeneities.

An intra-class correlation coefficient (ICC) is normally defined to examine the proportion of specific subject-level variance:

$$\mathrm{ICC} = \frac{\sigma_B^2}{\sigma_b^2 + \sigma_w^2} \tag{6}$$

where σ_w^2 is within group variance and σ_b^2 is between group variance. A value of ICC close to zero indicates there is a very small variation between the different subjects, and a model without multilevel structure is adequate for the data. Otherwise, a multilevel model would be preferred.

Appendix 2. Artificial neural network model

The artificial neural networks (ANNs), a popular class of computational intelligence models, has been widely applied to drowsiness detection, partly because of its ability to work with massive amounts of multi-dimensional data, its modeling flexibility, and its generally good predictive ability.

In this study, we built a feed-forward neural network with one hidden layer consisting of the interconnection of neurons only between two adjacent layers. A back propagation training method was used. Before modeling, the following standardization procedure was carried out for each metric:

$$x_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{7}$$

A basic computational element is called a node. Each node receives input from an external source or from other nodes. Each input has an associated weight (w_{ij}), which can be modified to model synaptic learning by the process of training. The input of the *j*th node in the hidden layer is calculated as follows:

$$Z_j = \sum_i w_{ij}(x_i + b_j) \quad i = 1, 2, \ldots, N \quad j = 1, 2, \ldots, M \tag{8}$$

where Z_j is the input to *j*th node in the hidden layer, w_{ij} is the weight of *i*th node in the input layer to *j*th node in the hidden layer, x_j is the value of *i*th node in the input layer, b_j is bias value for the *j*th node in the hidden layer, *N* is the number of nodes in the input layer, and *M* is the number of nodes in the hidden layer.

The output of a node is decided by its input as well as the activation function. Different activation functions such as sigmoid functions, hyperbolic tangent functions, and logistic functions can be used. A hyperbolic tangent function was used as the activation function of the hidden layer in our study. It was calculated as follows:

$$H_j = f(Z_j) = \tanh(Z_j) = \frac{e_j^{2z} - 1}{e_j^{2z} - 1} \tag{9}$$

where H_j is the output of *j*th node in hidden layer.

In our study, we want the outputs of ANN to be interpretable as probabilities for a categorical target variable (DL), for those outputs to lie between 0 and 1, and to have a sum of 1. Therefore, a Softmax activation function is used for the output layer, which is written as follows:

$$O_k = \frac{\exp(Z_k)}{\sum_{m=1}^{c} \exp(Z_m)} \tag{10}$$

where O_k is the output of *k*th node in the output layer, *c* is the number of categories for the target variable.

In the training process, the network output, in general, may not be equal to the desired output. Therefore, the output error is calculated as the difference between the network output and the desired output. If the output error does not satisfy the tolerance level, the network modifies the connection weights (w_{ij}) according to the value of the output error; then, training data is inputted again to the network and the network output is calculated. The training cycle is continued until the network achieves the desired tolerance level.

References

Åkerstedt, T., Gillberg, M., 1990. Subjective and objective sleepiness in the active individual. Int. J. Neurosci. 52 (1/2), 29–37.

Balkin, T.J., Horrey, W.J., Graeber, R.C., Czeisler, C.A., Dinges, D.F., 2011. The challenges and opportunities of technological approaches to fatigue management. Accid. Anal. Prev. 43 (2), 565–572.

Central Intelligence Agency, United States, 2010. The World Factbook, Retrieved 2013.

Forsman, P.M., Vila, B.J., Short, R.A., Mott, C.G., Van Dongen, H., 2013. Efficient driver drowsiness detection at moderate levels of drowsiness. Accid. Anal. Prev. 50, 341–350.

Hamada, T., Ito, T., Adachi, K., Nakano, T., Yamamoto, S., 2003. Detecting method for drivers' drowsiness applicable to individual features. Proc. Intell. Transp. Syst. 2, 1405–1410.

Hu, S., Zheng, G., 2009. Driver drowsiness detection with eyelid related parameters by support vector machine. Expert Syst. Appl. 36 (4), 7651–7658.

Ingre, M., ÅKerstedt, T., Peters, B., Anund, A., Kecklund, G., Pickles, A., 2006b. Subjective sleepiness and accident risk avoiding the ecological fallacy. J. Sleep Res. 15 (2), 142–148.

Ingre, M., Åkerstedt, T., Peters, B., Anund, A., Kecklund, G., 2006a. Subjective sleepiness, simulated driving performance and blink duration: examining individual differences. J. Sleep Res. 15 (1), 47–53.

Jo, J., Lee, S.J., Park, K.R., Kim, I.J., Kim, J., 2014. Detecting driver drowsiness using feature-level fusion and user-specific classification. Expert Syst. Appl. 41 (4), 1139–1152.

Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J.D., Ramsey, D.J., 2006. The Impact of Driver Inattention on Near-crash/Crash Risk: An Analysis using the 100-car Naturalistic Driving Study Data. No. HS-810 594.

Li, W., He, Q.C., Fan, X.M., Fei, Z.M., 2012. Evaluation of driver fatigue on two channels of EEG data. Neurosci. Lett. 506 (2), 235–239.

Maycock, G., 1997. Sleepiness and driving: the experience of UK car drivers. Accid. Anal. Prev. 29 (4), 453–462.

NHTSA's National Center for Statistics and Analysis, 2011. Traffic Safety Facts: A Brief Statistical Summary. NHTSA, U.S. Department of Transportation, DOT HS 811 449.

Patel, M., Lal, S.K.L., Kavanagh, D., Rossiter, P., 2011. Applying neural network analysis on heart rate variability data to assess driver fatigue. Expert Syst. Appl. 38 (6), 7235–7242.

Radun, I., Ohisalo, J., Radun, J., Rajalin, S., 2012. Law defining the critical level of driver fatigue in terms of hours without sleep: criminal justice professionals' opinions and fatal accident data. Int. J. Law Crime Justice 40 (3), 172–178.

Sandberg, D., Wahde, M., 2008. Particle swarm optimization of feed forward neural networks for the detection of drowsy driving. In: IEEE International Joint Conference on Neural Networks, pp. 788–793.

The Ministry of Public Security of the People's Republic of China, 2009. Road and Transport Authority. Road Traffic Crash Statistics 2001–2008.

Thiffault, P., Bergeron, J., 2003. Fatigue and individual differences in monotonous simulated driving. Pers. Individ. Differ 34 (1), 159–176.

Wierwille, W.W., Wreggit, S.S., Kirn, C.L., Ellsworth, L.A., Fairbanks, R.J., 1994. Research on Vehicle-based Driver Status/Performance Monitoring; Development, Validation, and Refinement of Algorithms for Detection of Driver Drowsiness, DOT HS 808 247.

Yang, G., Lin, Y., Bhattacharya, P., 2010. A driver fatigue recognition model based on information fusion and dynamic Bayesian network. Inf. Sci. 180 (10), 1942–1954.