Highway Risk Prediction and Factor Evaluation using Convolutional Neural Networks

Xiaodan Zhang, Chengwei Huang, *Member, IAENG,* Yongsheng Chen

***Abstract*—In this paper, we study the highway risk classification and accidents prediction using machine learning models. First, highway facilities data are collected in the form of numbers and text. The facilities information is then processed by encoding and digitalization. Second, stochastic optimization algorithm is used to select features for modeling the risk. Supervised information is provided by human expert. Third, two types of customized convolutional neural networks are introduced to highway road risk modeling. Different feature combinations are evaluated. Finally, experiments are carried out for risk classification, accidents number regression and feature set evaluation. The results show that the proposed highway risk model is effective and may contribute considerably to road risk management.**

***Index Terms*—road risk prediction, feature evaluation, convolutional neural network.**

I. Introduction

Due to the rapid increase of road and traffic construction, traffic safety has become an important problem. More and more research attentions are paid to traffic safety [1], [2], [3], [4]. Predicting the road risk can effectively reduce the number of accidents, casualties and property losses.

Risk prediction is very important for highway safety management. The traffic accident data of a certain section can be used to evaluate the risk level of the road in the section, and improve the road facilities.

In the traditional highway planning, the road risk level assessment is usually carried out manually, and the evaluation model is designed based on expert experience. Although the risk assessments from human experts are reliable in practice, they can not fully reflect the relationship between highway road factors and accident risk. The number of highway road factors considered is small due to the limitation of human experience and computation ability.

The model based on human expert knowledge established for a particular type of highway can not well adapt to various complex situations of the actual highway, and can not accurately reflect the impact of unknown highway driving environment on highway risk level. Even if the relevant parameters are found, the weights between features are often difficult to determine.

It is necessary to design automatic road risk level assessment and accident number prediction methods based on machine learning and data-driven methods to remedy the shortcomings of human expert assessment.

In the existing literature, Hermans et al.[1], [2], studied the road safety problem based on traffic accidents. They proposed a road safety index, and evaluated different factors that influence traffic safety. Miaou and Song[5] studied the Bayesian method for safety in a road network, and they also used spatial information in safety analysis. AlKheder and Al-Rashidi[6] also used a Bayesian approach to study road traffic accidents, especially in the Gulf countries. Their model could be used for accident prediction, and related road factors could be further studied. Tian et al.[7], studied traffic condition from a timing model by analyzing the signals in urban intersections. Their work may be applied to safety management in traffic control. Zhang and Leng[8] proposed to use fuzzy AHP method to study traffic safety using various road factors.

Machine learning algorithms are used for highway data modelling, and the risk levels and accidents can be predicted. Due to the complexity of road factors, it is often difficult to find the optimized parameters for modeling. Conventional machine learning algorithms are often adopted[9], [10], and the deep learning algorithms in this field are lack of study.

In this paper, we use machine learning methods, especially convolutional neural networks, to establish a road risk level classification model and an accidents number prediction model. The analysis of the importance of highway facilities are also carried out in feature evaluation. It can bring considerable returns in the future and is very important to road safety.

The rest of the paper is organized as follows. In Sec.II, the database used in this research is introduced; in Sec.III, the stochastic optimization, shuffled frog leaping algorithm, is applied to feature selection; in Sec.IV deep neural networks are described with details, and applied to highway risk prediction; in Sec.V, the experimental results are provided, including risk classification, accidents prediction and feature evaluation; finally, in Sec.VI, the conclusions are drawn and future research directions are given.

II. Data collection and Pre-Processing

A. The Database Description

The database comes from domestic highway in China without privacy or sensitive information. It includes road facility factors as features, and risk levels as labels. A total of 78 factor parameters are considered, including: main line speed limit, central isolation type, right side object attributes, right side guardrail height, horizontal curve curvature, sight distance (main line), slope, whether the ramp acceleration and deceleration lanes are sufficient, sight distance (entrance ramp), skid resistance, accident-prone sections, speed, vehicle flow range, truck ratio, etc.

Manuscript received December 4, 2020; revised November 10, 2021. This work was supported by Innovation Funding Program of Research Institute of Highway, Ministry of Transport(# 2020-9061).

Xiaodan Zhang is an Associate Research Fellow of Ministry of Transport Research Institute of Highway(RIOH) , Beijing 100088, China(corresponding author, tel:86-10-82019506; fax:010-62370567; e-mail:zhangdaqing_925@163.com).

Yongsheng Chen is a Research Fellow of Ministry of Transport Research Institute of Highway(RIOH) , Beijing 100088, China(e-mail:ys.chen@rioh.cn).

Chengwei Huang is a Principal Investigator of Zhejiang Lab, Hangzhou, China(e-mail:huangchengwei@zhejianglab.com.

The data is recorded every 100 meters (a road section), and there are 108492 records in total. The data is presented in data frame (table) formats. In the data table, the first column represents the risk level of road safety facilities in this road section, which is encoded as one to five levels. The first level is the safest and the fifth level is the most dangerous. The rest columns represent the feature of safety facilities.

Each data unit in the table is represented by text description or numbers. When it is recorded, the data inspector can choose from a small set of fixed text phrases to describe the safety facilities of a road section. By looking at the highway monitoring pictures, inspector can select the most suitable item from the menu to fill in the database. Therefore, each column corresponds to a set of symbols, and each unit is a random variable.

B. The Feature Encoding

The text description is not used for modeling directly. Its value conforms to the probability distribution(e.g. mixture of Gaussians). These text and symbols can be digitally encoded. Numerical values are replaced by 0-N numbers and others are encoded into one-hot-vectors.

In the original data, some features have clear physical meanings, for example, speed limits. For this type of feature, the speed contains risk information. The faster the speed, the higher the risk. Encoding them numerically can avoid losing the original meaning.

III. Stochastic Feature Optimization

Feature selection and optimization is a key step for many pattern recognition problems[11]. Previous works have explored various conventional feature optimization methods[12], [13].

A. Shuffled Frog Leaping Algorithm

In this section, we will introduce the basic principles in Shuffled Frog Leaping Algorithm(SFLA), which is an effective stochastic optimization method[14], [15]. For feature selection in a d-dimensional target searching space, D frogs are randomly generated to compose initial population. The i^{th} frog individual is represented as $X = (x_{i1}, x_{i2}, \ldots, x_{id})$. Frogs are arranged from good to bad according to their fitness values to divide the whole population into N populations. Among them, the frog ranking fist is assigned into the 1^{st} population, one ranking second into the 2^{nd} population, and so on and so forth until all frogs have been assigned. Every sub-population is used for local depth search. The update operation is applied only to the current worst individual X_w, with the update strategy shown in Eq.1.

$$X_w^{new} = X_w + rand(seed)(X_b - X_w) \tag{1}$$

where $rand(seed)$ represents random number uniformly distributed between 0 and 1, X_b is the best individual in the current sub-population. X_w^{new} represents the individual after updating.

According to the update strategy in Eq.1, if the fitness value of X_w^{new} is better, X_w will be replaced. If it isn't improved, replaces X_w with the global best individual X_g, and repeat the update strategy until meet the iteration limit. After the local depth search of all sub-populations have been finished, all frogs in are mixed and reordered into new sub-populations. The local depth search is repeated until meet the limit of iterations.

B. Feature Optimization for Road Factors

After the road factors are encoded, the feature dimensional space can be optimized. Feature optimization can improve the network convergence and reduce the computational complexity. When sample size is insufficient, the probability model between the road factors and the risk level is difficult to achieve. Feature optimization may overcome the insufficiency of training samples. In this section, We use SFLA to optimize the encoded road factors, before the modelling process.

First, the population is initialized. Each individual represents a combination of feature dimensions. Generally, the feature dimension is reduced from the original dimensions (corresponding to the environmental factors of the highway) to 10-20 dimensions (considering the risk level of the highway is divided into five levels).

Second, when calculating the individual fitness, data-driven method is adopted to avoid explicit definition of fitness function. Manually annotated labels are adopted to test the classification accuracy, which is used as fitness value.

The feature vector $\mathbf{f}$ contains the total number (M) of road factors, and each feature dimension of the feature vector is selected from the above road factors. The sub-set feature vector $\mathbf{f_j}$ contains any combination of feature dimensions not more than M. Each sub-set vector $\mathbf{f_j}$ represents a possible combination of features, thus forming the feature set $\{\mathbf{f_j}\}$ to be optimized. The process of generating the optimal feature combination by using the SFLA is as follows, as reported in our previous work[16]:

1)From our database of road factor and risk level, high confidence data set is selected as the validation data for SFLA.

2)Data-driven method is used to calculate the cost function. Shuffled frog leaping algorithm generates different individuals, each representing a feature vector $\mathbf{f_j}$. Each time, through the classification model, the classification rate is tested and used as the fitness value of each individual.

3)According to the fitness value, the individuals with high values are selected and retained in the next iteration to eliminate the individuals with low values.

4)When the feature discrimination reaches the preset accuracy, the iteration is completed and the optimal feature combination is preserved. Otherwise, steps 2) and 3) are repeated.

After optimization, the dimension of feature vector is reduced.

IV. Deep Neural Networks

A. Deep Neural Network

Deep Neural Network(DNN) has been applied to various previous works and often outperformed conventional methods[17], [18], [19]. In this work we aim at the application of machine learning methods in road risk management.

The structure of the deep neural network is shown in the following figure: The structure diagram consists of input

layer v, multiple hidden layer h_k and output layer. Only the adjacent layer nodes are connected, and there is no connection between the nodes in the same layer. Each connection has a weight value w.

The training of the whole deep neural network is mainly composed of unsupervised and supervised training. The training process is to train a deep generative model based on restricted Boltzmann machine. After the initial network parameters are obtained by layer-by-layer greedy training. Back-propagation based on the minimum mean square error is used to update the parameters of the DNN.

B. Multi-Path Deep Neural Network

In our previous work[16], we reported a DNN model with application in road safety application. In this paper, we extend it to feature selection evaluation and use it to compare with a novel multi-scale CNN model that we proposed to use for road safety classification.

The input is an optimized feature vector, and the output is the risk level. It improves the traditional deep neural network structure as follows: The output L_i of each layer of the neural network is connected to the final fully connected layer, and the fusion vectors are merged: $L = [L_1, L_2, \dots]$, ReLu (Rectified Linear Unit) activation function is used. The cost function type adopts cross-entropy cost function.

C. Multi-scale Convolutional Neural Network

Inspired by the previous work[20], a multi-scale residual network is used. It had successful application in data modeling due to its ability to be adaptive to local patterns. In this paper we extend its application in road safety classification with 1D convolution.

Conventional machine learning algorithms have drawbacks in learning complicated representations. Deep neural network, especially convolutional neural network may provide a stronger ability to discover data relations. Using multi-scale receptive field, the residual network can better model the relationship between risk level and highway data.

The advantage of residual network is the ability to learn the features of the original input through the short-cut connection. The short-cut connection can bring deeper network structures to learn the representation. As shown in Fig.1, we build several residual blocks and fully connected layers to estimate the risk probabilities for highway data. We adopt non-causal convolution to process the road data, in which both direction of the road can affect the current risk level.

V. Experimental Results

In our experiments, we test machine learning models for solving three problems. The first experiment is the risk level classification problem. The second experiment is the accidents number prediction problem. The third experiment is risk factor evaluation problem.

A. Highway Risk Assessment

In this experiment, we use the method described in Sec.II to process the data, encode the data and fuse the multi-source data to obtain the 78-dimensional features. The output has five risk levels from level 1 to level 5.

Firstly, the feature is optimized by the shuffled frog leaping algorithm, and then the optimized feature is used as the input of the deep neural network, and the output is the road risk levels. In this experiment, we tested various machine learning algorithms, and the results are shown in Tab.I.

To better understand the results of risk classification models, multi-classification results are converted into two-class classification statistics. Ten-fold cross-validation is also performed on the data set.

B. Highway Accident Prediction

In this experiment, we use Multi-Path DNN and Multi-Scale CNN to perform regression analysis of accident data.

First we convert the number of accidents into the estimated number of accidents per 100 meters according to the traffic accidents records.

Then, we use the supervised information to test the model ability of accidents prediction. The out put of the model is the estimated number of accidents per 100 meters, centered at a given location.

Based on the number of accidents locations, we can estimate the density distribution, and use it as the target for regression.

The CNN model is especially suitable for model the relations between neighboring locations. In practice, the road factors are similar in close locations, and need to be processed by one-dimensional convolution for better representation.

Using Multi-Path DNN the Mean Squre Error(MSE) of the prediction is 1.38 while using Multi-Scale CNN the MSE reaches 1.01.

C. Risk Factors Evaluation

In order to better understand the factors related to risk, in this evaluation, we rank the top 10 feature dimensions related to highway road factors. As shown in Fig.2, using the best performance Multi-Scale CNN model, we provide the mean error rates by removing one feature at each testing time.

Based on the experts knowledge, two subset of features are drawn that are considered the insignificant factors, and the significant factors. These features are shown in Tab.II.

By removing insignificant feature set and observing the results, we can validate the important road factors. This is meaningful for the data collection in practice, as the insignificant factors could be skipped in manually annotation to save time and cost. The risk management for highway becomes more efficient when focusing on more important road factors.

We can see from Tab.III, after removing the insignificant factors, the model accuracy decreases very little. After removing the significant factors, the accuracy decreases considerably. This proves that the evaluation of the risk factors is reliable. By removing the important factors the overall performance of the system has changed considerably.

Multi-Scale CNN has shown a better performance, and it has a stronger ability to model the relations between the road factors and the risk locations. Extended experimental results on Multi-Scale CNN model are shown in Tab.IV and Tab.V. We can see that Set B is considerably more important than Set A. The performance drop introduced by removing factors from Set B is bigger.

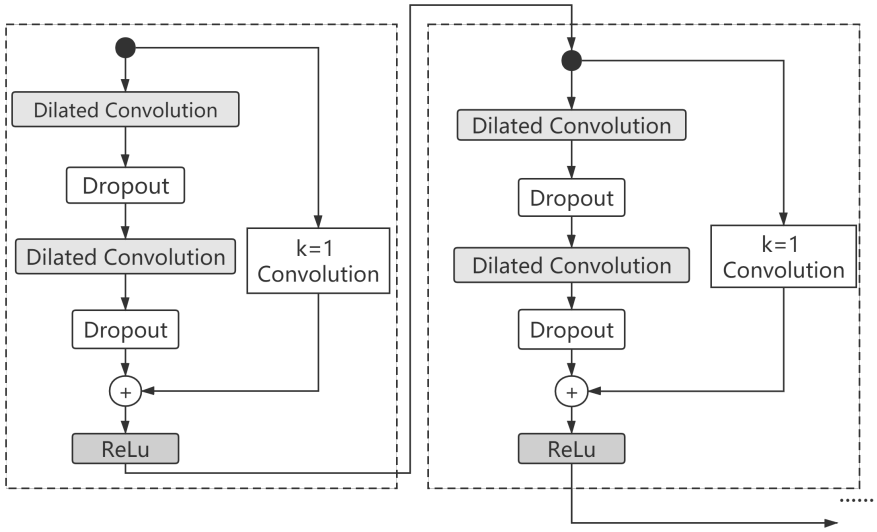

Fig. 1. Residual blocks in the network structure.

TABLE I
ACCURACY FOR EACH PAIR OF TWO-CLASS RISK LEVELS CLASSIFICATION

Two-class Pairs	SVM	XGBoost	GMM	KNN	Multi-Path	Multi-Scale
Lv1-Lv2	82.3%	94.5%	81.6%	83.4%	94.9%	95.0%
Lv1-Lv3	85.4%	97.5%	88.9%	81.3%	98.9%	99.1%
Lv1-Lv4	86.1%	98.1%	95.4%	85.6%	99.4%	99.6%
Lv1-Lv5	93.2%	97.2%	97.7%	89.3%	98.8%	98.7%
Lv2-Lv3	84.5%	87.5%	86.2%	80.1%	90.9%	93.6%
Lv2-Lv4	87.5%	90.2%	90.1%	85.4%	94.2%	95.5%
Lv2-Lv5	93.4%	97.2%	94.4%	81.7%	99.3%	99.3%
Lv3-Lv4	77.8%	80.1%	80.8%	71.2%	82.3%	86.3%
Lv3-Lv5	81.2%	87.3%	92.8%	83.2%	98.8%	99.5%
Lv4-Lv5	90.0%	93.2%	92.2%	85.6%	98.8%	98.3%
Average	**86.1%**	**92.3%**	**90.2%**	**82.7%**	**95.6%**	**96.5%**

TABLE II
INSIGNIFICANT FEATURES AND SIGNIFICANT FEATURES BASED ON EXPERT KNOWLEDGE

Rank	Insignificant Features Set A	Significant Features Set B
1	re-construction price	flat curve curvature
2	left side land usage	slop
3	right side land usage	visual distance (main line)
4	right road side object distance	visual distance(ramp exit)
5	left shoulder vibration belt	anti-slip
6	right shoulder vibration belt	ramp acceleration lane sufficiency
7	road side parking	central isolation type
8	tire break marks	traffic amount range
9	road construction	main line speed limit
10	tunnel lighting intensity	truck ratio
11	speed surveillance camera	right guardrails height
12	variable message board	frequent accidents section
13	total traffic influence	speed
14	sign of tunnel emergency parking	right road side object character

TABLE III
MEAN ERROR WHEN REMOVING DIFFERENT FEATURE SETS

Test Case	SVM	XGBoost	GMM	KNN	Multi-Path	Multi-Scale
Removing Set A	14.9%	9.0%	10.8%	8.9%	5.9%	4.7%
Removing Set B	17.2%	14.4%	15.7%	12.3%	12.9%	11.5%

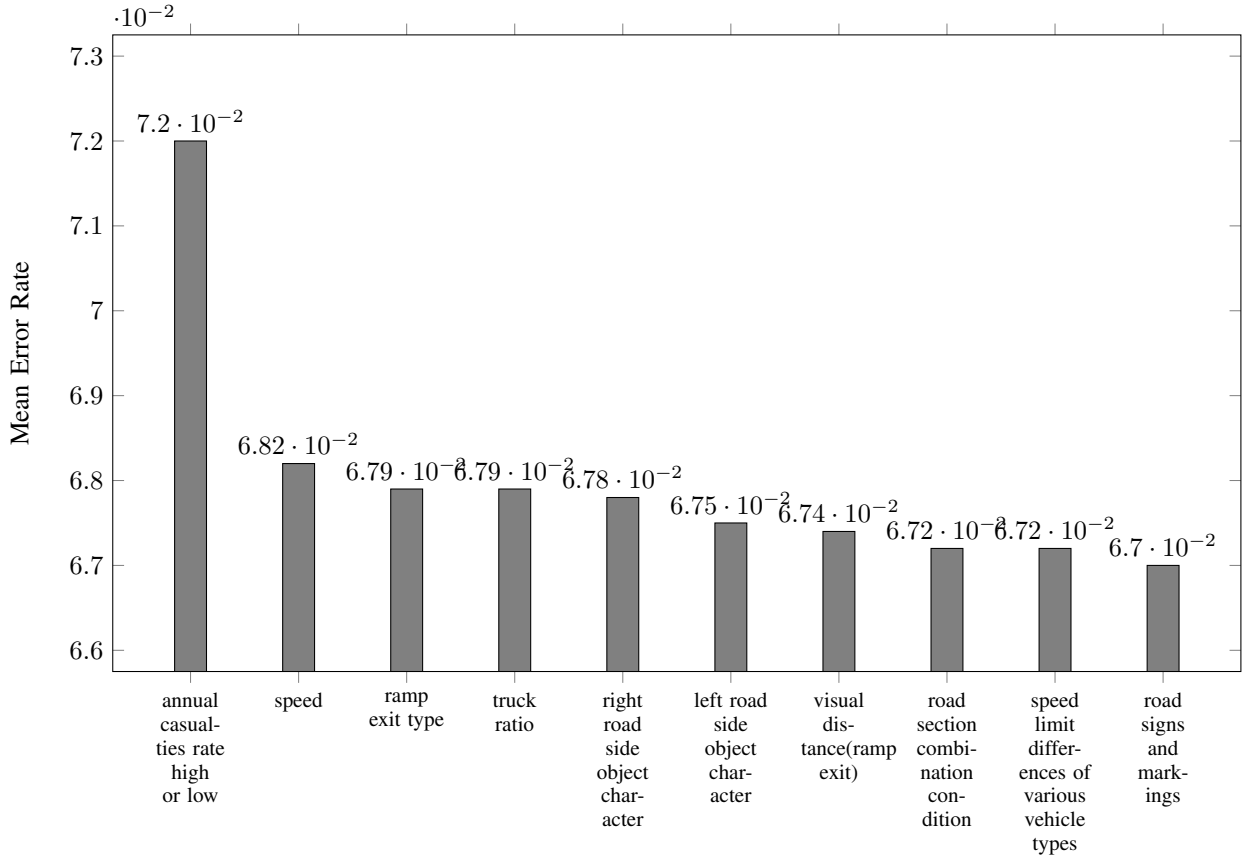

Fig. 2. Ranking of highway risk factors.

TABLE IV
CONFUSION MATRIX OF MULTI-SCALE CNN RESULTS(REMOVING SET A).

	Level 1(Predicted)	Level 2(Predicted)	Level 3(Predicted)	Level 4(Predicted)	Level 5(Predicted)
Level 1	86.0%	6.5%	3.8%	2.6%	1.1%
Level 2	5.3%	76.9%	7.7%	7.7%	2.4%
Level 3	2.0%	8.1%	67.3%	15.1%	7.5%
Level 4	2.7%	3.0%	15.0%	75.5%	3.8%
Level 5	2.3%	2.0%	8.4%	3.6%	83.7%

TABLE V
CONFUSION MATRIX OF MULTI-SCALE CNN RESULTS(REMOVING SET B).

	Level 1(Predicted)	Level 2(Predicted)	Level 3(Predicted)	Level 4(Predicted)	Level 5(Predicted)
Level 1	73.5%	12.7%	6.1%	5.4%	2.3%
Level 2	10.0%	50.7%	17.1%	16.5%	5.7%
Level 3	4.5%	17.2%	42.8%	21.0%	14.5%
Level 4	5.2%	7.1%	23.1%	58.6%	6.0%
Level 5	4.9%	3.9%	15.9%	8.3%	67.0%

VI. CONCLUSION

In this paper, using highway risk data, combined with deep neural network and shuffled frog leaping algorithm, we compared two novel deep neural network models with the other conventional machine learning methods.

The classification model and regression model are established respectively. The classification model is used for road risk level assessment and the regression model is used for road accident prediction. The verification results show that the proposed method is effective.

In our future work, we will study more deep learning models to improve the abilities of risk prediction, and we will collect more highway risk data to faciliate the study.

REFERENCES

[1] E. Hermans, F. Van den Bossehe, G. Wets, "Combining road safety information in a performance index," *Accident Analysis and Prevention*, vol.40, no.4, pp.1337-1344, 2008.

[2] E. Hermans, F. Van den Bossehe, G. Wets, "Uncertainty assessment of the road safety index," *Reliability Engineering and System Safety*, vol.94, pp.1220-1228, 2009.

[3] X. Zhang, Y. Chen, M. Liu, et al., "Acoustic Traffic Event Detection in Long Tunnels Using Fast Binary Spectral Features" *Circuits, Systems, and Signal Processing*, vol.39, no.6, pp.2994-3006, 2020.

[4] V. Anchukov, A. Alyukov, S. Aliukov, "Stability and Control of Movement of the Truck with Automatic Differential Locking System" *Engineering Letters*, vol.27, no.1, pp.131-139, 2019.

[5] S. P. Miaou, J. J. Song, "Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence," *Accident Analysis and Prevention*, vol.37, no.4, pp.699-720, 2005.
[6] S. AlKheder, M. Al-Rashidi, "Bayesian hierarchical statistics for traffic safety modelling and forecasting" *International Journal of Injury Control and Safety Promotion*, pp.1-13, 2019.
[7] X. Tian, F. Ruan, H. Cheng, et al., "A signal timing model for improving traffic condition based on active priority control strategy" *Engineering Letters*, vol.28, no.1, pp.235-242, 2020.
[8] C. Y. Zhang and X. Y. Leng, "Urban traffic safety evaluation model based on fuzzy analytic hierarchy process" *Advances in Transportation Studies*, special issue 1(SI 1), pp.51-60, 2020.
[9] J. Liu, Q. Zhang and Z. Luo, "Dynamical analysis of fuzzy cellular neural networks with periodic coefficients and time-varying delays" *IAENG International Journal of Applied Mathematics*, vol.46, no.3, pp.298-304, 2016.
[10] Z. Qiao, L. Zhou and J. Z. Huang, "Sparse linear discriminant analysis with applications to high dimensional low sample size data" *International Journal of Applied Mathematics*, vol.39, no.1, pp.600-606, 2009.
[11] C. Wu, C. Huang, H. Chen, "Expression recognition using semantic information and local texture features" *Multimedia Tools and Applications*, vol.77, no.9, pp.11575-11588, 2018.
[12] C. Huang, B. Song, L. Zhao, "Emotional speech feature normalization and recognition based on speaker-sensitive feature clustering" *International Journal of Speech Technology*, vol.19, no.4, pp.805-816, 2016.
[13] C. Wu, C. Huang, H. Chen, "Text-independent speech emotion recognition using frequency adaptive features" *Multimedia Tools and Applications*, vol.77, no.18, pp.24353-24363, 2018.
[14] H. Yu, C. Huang, X. Zhang, et al., "Shuffled frog-leaping algorithm based neural network and its application in speech emotion recognition" *Journal of Nanjing University of Science and Technology*, vol.35, no.5, pp.659-663, 2011.
[15] Xiaodan Zhang, Chengwei Huang, et al., "Recognition of practical speech emotion using improved shuffled frog leaping algorithm" *Chinese Journal of Acoustics*, vol.33, no.4, pp.441-441, 2014.
[16] X. Zhang, C. Huang, Y. Chen, "Method of Highway Risk Assessment and Accident Quantity Prediction Based on Multi-Source Heterogeneous Data and Deep Neural Network," *Resilience and Sustainable Transportation Systems*, Reston, VA, American Society of Civil Engineers, pp.118-125, 2020.
[17] C. Huang and H. Jiang, "Image indexing and content analysis in children's picture books using a large-scale database" *Multimedia Tools and Applications*, vol.78, no.15, pp.20679-20695, 2019.
[18] Y. Xie, R. Liang, Z. Liang, et al. "Speech emotion classification using attention-based LSTM" *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.27, no.11, pp.1675-1685, 2019.
[19] X. Xu, C. Huang, et al. "Graph learning based speaker independent speech emotion recognition" *Advances in Electrical and Computer Engineering*, vol.14, no.2, pp.17-23, 2014.
[20] G. Zhou, Z. Li, M. Fu, Y. Feng, X. Wang, C. Huang, "Sequence-to-Sequence Load Disaggregation Using Multi-Scale Residual Neural Network," *IEEE Transactions on Instrumentation and Measurement*, vol.70, pp.1-10, 2020.

Xiaodan Zhang received Ph.D. degrees from Southeast University, China, in 2013. Her major research areas are traffic safety and smart freeway, and she is specialized on modern intelligent data analytics based on Big Data.

She works as an associate research fellow with Ministry of Transport Research Institute of Highway (RIOH) since August 2013.

Chengwei Huang (M'19) received his undergraduate degree in 2006, and Ph.D. for speech emotion recognition in 2013, both from Southeast University (China). His main research interests include affective computing, signal processing and machine learning.

He worked on face recognition at University of Houston. He worked on human-computer interaction as an associate professor in Soochow University. He is currently with Zhejiang Lab, focusing on AI industrial applications.

Yongsheng Chen received two Ph.D. degrees, from Ryerson University, Canada, in 2013, and Beijing University of Technology, China, in 2001. His major research areas are traffic safety and smart freeway, and he is specialized on modern intelligent data analytics based on Big Data.

He works as a Research Fellow with Ministry of Transport Research Institute of Highway (RIOH) since September 2018, after serving a four-year term as the Chief Research Scientist of RIOH since August 2014.