Dean&Francis
ISSN 2959-6157

Research Progress on Deep Learning-Based Emotion Recognition and State Monitoring in Intelligent Cockpits

Pengyu Chen[1, *]

Wenzhuo Yang [2] **and**

Ziwen Wang [3]

[1] International School Affiliated to Nanjing University, Nanjing, 210009, China
[2] Wanhui Academy Dujiangyan, Dujiangyan, 611800, China
[3] Jiangshan High School of Zhejiang Province, Jiangshan, 324100, China
*Corresponding author: yosoroyeah@outlook.com

Abstract:

As the intelligent cockpit evolves towards a human-machine emotional interaction center, driver emotion recognition has become a key task for enhancing active safety and interaction experience. This paper systematically studies the applicability and optimization paths of multimodal fusion emotion recognition methods based on deep learning in the intelligent cockpit environment. The research findings show that the visual methods (such as YOLO, MobileNetV3) have the advantages of non-intrusiveness and high real-time performance (response time < 40 ms), but are susceptible to changes in lighting and facial occlusion, and have insufficient recognition for high-risk emotions (such as anger); the physiological signals (electroencephalogram, electrooculogram) have strong objectivity, but have high equipment costs, poor wearing comfort, and limited generalization and robustness due to individual differences; the speech method (combined with psychological acoustic model) has the characteristics of natural interaction and resistance to expression disguise, but is greatly affected by in-vehicle noise and has decreased recognition stability. To address the above problems, this paper proposes improvement strategies such as lightweight network structure, sample enhancement for high-risk emotions, and hardware-algorithm collaborative anti-interference, which significantly improve the model's adaptability and real-time performance in extreme environments. The research shows that a single modality is difficult to meet the complex requirements of the intelligent cockpit, and in the future, multi-modal deep fusion should be adopted to achieve the coordinated optimization of accuracy, robustness, and real-time performance while ensuring user experience, providing key support for building an integrated intelligent cockpit emotion computing framework of perception - understanding - intervention.

Keywords: Driver emotion recognition; multimodal fusion; intelligent cockpit; deep learning; real-time performance.

ISSN 2959-6157

1. Introduction

With the deep integration of the automotive industry and artificial intelligence technology, the intelligent cabin system is gradually evolving from a "functional integration platform" to a "human-machine emotional interaction hub", significantly enhancing the driving and passenger comfort, and providing a new paradigm for active traffic safety. However, despite the continuous progress of vehicle active safety technologies, the global traffic accident rate remains high. Numerous studies have shown that the driver's emotional state-such as anger, anxiety, fatigue or distraction - significantly weakens their perception, judgment and operation capabilities, and is one of the key human factors that trigger traffic accidents [1]. Therefore, real-time and accurate monitoring and intervention of the driver's emotional state have become the core task for intelligent cabins to achieve a human-centered safety protection system.

Currently, driver emotion recognition mainly relies on three types of signal sources: physiological signals (such as electroencephalogram, electrocardiogram, and electrooculogram), visual signals (such as facial expressions, eye state, and head posture), and voice signals [2]. Terzi et al. pointed out that the driver's emotional regulation ability is a key human factor affecting traffic safety, indicating that monitoring and regulating the driver's emotional state is of great significance for improving traffic safety levels [3]. Xue et al. further pointed out that existing research mainly focuses on physiological and visual signals. Physiological signals can objectively reflect emotional changes, but they rely on contact sensors, which have problems such as high cost, uncomfortable wearing, and limited application in vehicle systems [4]. Visual signals have the advantages of non-contact and rich information, but are easily interfered by environmental factors such as light changes, occlusion, and head posture deviation. In contrast, voice signals contain rich emotional cues in natural interaction, but are easily affected by vehicle noise, speaking content, and individual acoustic differences. In recent years, breakthroughs in deep learning technology have provided powerful tools for multimodal emotion recognition, which can effectively integrate heterogeneous data, improve model robustness and generalization ability [5,6]. In this context, this paper systematically explores the deep learning-based multimodal emotion recognition methods for vision, physiology and voice, focusing on analyzing their applicability, performance boundaries and engineering implementation challenges in the restricted environment of intelligent cabins, and proposing targeted optimization paths. This research not only helps to promote the transformation of emotion perception technology from the laboratory to real driving scenarios, but also provides important theoretical support and practical references for building an integrated perception - understanding - intervention intelligent cabin emotional computing framework.

2. The Main Content of the Research

Among all driver behavior recognition technologies, video images obtained through visual methods have become the mainstream approach and primary data source in this field due to their intuitiveness, ease of acquisition, and rich information content. The traditional visual-based emotion state discrimination method involves monitoring the facial area of the input image, detecting key points and constructing feature vectors to obtain the relationship between each part of the facial area and the geometric features in the image [7]. For the visual detection method, its main characteristic lies in the non-contact acquisition of visual information, which results in a good user experience and rich visual signal information, basically meeting the judgment of the driver's emotion. However, the visual method is also prone to environmental interference, and at the same time, the driver may consciously control or disguise their expressions, leading to system misjudgment [8].

However, merely relying on visual sensors for recognition cannot accurately identify various driver behaviors in complex and changing environments. Therefore, in the research on methods for detecting driver emotions during driving, physiological signals, as objective responses to physiological changes in the human body, are also an important reference indicator in the process of driver emotion recognition. The physiological signal detection method refers to directly detecting the changes in electrical physiological signals or other biological signals generated within the driver's body due to emotional states, to objectively reflect the autonomic nervous system changes triggered by emotions. Among them, electroencephalogram (EEG) signals and electrooculogram (EOG) signals are two important reference standards. Based on the physiological signal detection method, it is obviously quite objective and difficult to disguise, has high precision, and is not easily affected by the environment. At the same time, contact measurement is also a major disadvantage of the physiological signal detection method, with extremely poor comfort and even affecting normal driving [9].

In addition to visual signals and physiological signals for perceiving the driver's emotions, due to certain limitations, voice signals also play an important role in detecting the driver's emotions. The voice signal detection method refers to collecting the driver's speaking voice signals through the vehicle microphone and extracting

acoustic features (such as pitch, speech rate, energy, spectrum, etc.) from them to analyze their emotional state. For the voice signal detection method, it is usually used as supplementary information to improve the accuracy of other information such as visual signals and physiological signals. Similar to the visual method, the voice signal method is also typically a non-contact detection method, so it is easy to deploy. However, it is more susceptible to environmental noise interference and individual voice differences [10].

Therefore, based on the content of the above discussion, Fig. 1 presents a summary of the methods and characteristics for driver emotion recognition.

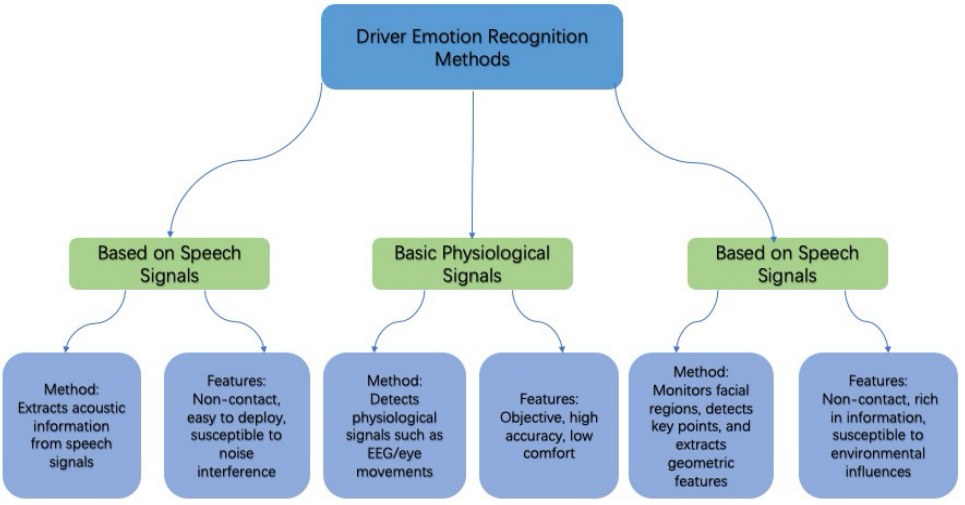

Fig. 1 ummary of Methods and Characteristics for Driver Emotion Recognition

3. Analysis and Discussion

3.1 Analysis of Emotion Recognition Methods Based on Visual Data

3.1.1 Related models

The driver emotion recognition method based on visual data has core advantages in the convenience of data acquisition and the intuitiveness of results. With the integration of deep learning technologies (CNN + YOLO), it effectively makes up for the accuracy shortcomings of traditional visual methods.

Zhang Jingming and others optimized the network backbone, neck, and head architectures. In the feature extraction stage, they enhanced the ability to locate the driver's facial area in complex cockpit environments [8]. They eliminated interference from non-target areas through grid division and multi-scale bounding box prediction. In the feature fusion stage, advanced feature extraction technology and a context-adaptive mechanism were introduced to adapt to multiple deployment scenarios. Relying on CNN to extract subtle facial behavior features and training with a dataset of tens of thousands of images of irregular operations, the final mAP@95% reached over 90%. It also has real-time detection capabilities and can stably monitor complex driving environments.

The face detection module of Liu Hong's system incorporates the CBAM attention module and the GSBS module, and adopts optimizations such as the VoVGSCSP module, CARAFE upsampling operator, and BiFPN structure. These improvements reduce the detection time per image from 23.5ms to 16.9ms while increasing the accuracy by 0.5% [7]. For the emotion recognition module, the first layer of the network is adjusted to a slice concatenation operation, and it employs DO-DConv convolution, the GAM attention module, and combines CycleGAN to balance the dataset. In real-vehicle experiments, the system achieves an accuracy of 86.78% in recognizing five basic emotions, with the response time controlled within 40ms, which meets the real-time monitoring requirements of intelligent cockpits.

3.1.2 Analysis of model limitations

The main limitations of the model proposed by Zhang et al. are reflected in two aspects: functionality and generalization. Firstly, its core focus is on driver violation recognition, without involving emotional state discrimination. As a result, it cannot capture potential driving risks caused by negative emotions such as anger and disgust. Its function is limited to behavioral compliance monitoring and has not been extended to driving psychological state early

ISSN 2959-6157

warning. Secondly, its training dataset mainly consists of samples of single violation behaviors, lacking samples of scenarios with superimposed behavior and emotion as well as samples of drivers with different ages and facial features. This leads to insufficient adaptability and generalization ability of the model in diverse scenarios and populations.

The main limitations of Liu Hong's model are reflected in two aspects: emotion coverage and environmental adaptability. Firstly, it only validates five types of emotions—happiness, fear, sadness, surprise, and neutrality—and does not include high-risk negative emotions such as anger and disgust. These two types of emotions are precisely important triggers for traffic accidents, resulting in a gap in the dimension of emotion monitoring. Secondly, in real-vehicle experiments, the model does not fully consider extreme scenarios such as strong light, low light, and dynamic occlusion. Under extreme conditions, the face detection frame is prone to deviation, which causes distortion in emotion feature extraction and a significant decline in recognition accuracy.

3.1.3 Suggestions for Improvement

Expanding emotion recognition functionality: Based on the existing behavior recognition model, add a lightweight emotion classification sub-network (e.g., the optimized MobileNetV3). Through multi-task training, realize the joint monitoring of behavior-emotion to fill the gap in driving psychological early warning.

Expanding the composite scenario dataset: Incorporate behavior-emotion omposite samples (e.g., making a call while angry) and samples of drivers with different characteristics. Combine data augmentation techniques to enhance the model's adaptability to complex scenarios and diverse populations.

Complementing high-risk emotion monitoring: Acquire real samples of anger and disgust through emotion induction experiments, and generate scarce samples using CycleGAN. This improves the coverage of emotion categories and strengthens the extraction of high-risk emotion features to enhance recognition accuracy.

Enhancing robustness in extreme environments: Add an extreme scenario test group, introduce image enhancement algorithms to preprocess input frames, and optimize the feature extraction network to ensure that the recognition accuracy remains above 80% in extreme environments with stable response time.

3.2 Analysis of Emotion Recognition Methods Based on Physiological Data

3.2.1 Related models

The core value of emotion recognition methods based on physiological data (such as electroencephalogram and electrooculogram) lies in the objectivity and accuracy of the recognition results, which is difficult to be replaced by visual methods. Physiological signals directly reflect the driver's neural activities and physiological states, and are not disturbed by subjective expression disguises (such as when a driver deliberately conceals their anger), thus being able to more truly capture potential risk emotions (such as latent fatigue and anxiety).

The technical pathway proposed by Shi Jinxuan et al. (Fourier Transform + 1D Convolutional Encoder + Recurrent Neural Network) further amplifies the advantages of physiological data [9]. Its advantages are mainly reflected in three aspects: First, Fourier Transform can effectively extract the frequency-domain features of EEG signals (e.g., enhanced α-waves correspond to a relaxed state, and enhanced β-waves correspond to a tense state) and the time-domain features of EOG signals (e.g., increased blinking frequency corresponds to fatigue). Second, the 1D Convolutional Encoder addresses the issue of insufficient sensitivity to emotional changes when using a single physiological signal (e.g., EEG alone) through feature fusion. Third, the Recurrent Neural Network (RNN) can capture the dynamic evolution process of emotions (e.g., the gradual transition from calm to anger), avoiding misjudgments caused by transient signal fluctuations.

3.2.2 Analysis of model limitations

1) Limitations in device portability and population coverage

Strong device dependence: The collection of EEG and EOG signals relies on multi-lead contact electrodes (for example, EEG requires multiple leads in the occipital and temporal lobes, while EOG requires special electrodes on the forehead). The equipment is relatively large in size and has a cumbersome wearing process, which is inconsistent with the requirements for lightweight and convenience in vehicle-mounted scenarios, making it difficult to adapt to the cockpit space of different vehicle models.

Insufficient adaptation to individual differences: The experimental data were based on 23 subjects with an average age of 23 years, failing to fully cover populations of different age groups (e.g., middle-aged and elderly individuals) and physiological characteristics (e.g., skin sensitivity, poor electrode contact caused by hair loss). As a result, the model's recognition performance declines in subjects whose feature distributions deviate from the average level, leading to limited generalization.

2) Common Limitations at the Technical Implementation Level

Difficulty in balancing real-time performance and cost:

Although recognition accuracy has been optimized through convolutional encoders and LSTM, multi-modal feature extraction (e.g., EEG frequency band division, eye movement feature statistics) and preprocessing (e.g., LDS smoothing) consume significant computing power. Using high-computing-power hardware will further increase costs, while adapting to low-computing-power in-vehicle chips may lead to reduced real-time performance (e.g., response latency exceeding 50ms).

Weak robustness in complex environments: Experiments are based on a simulated driving environment (SEED-VIG dataset) and fail to fully simulate scenarios in real-world driving such as jolts (which cause unstable electrode contact) and electromagnetic interference (e.g., interference of physiological signals by in-vehicle electronic devices). This easily leads to a decrease in the signal-to-noise ratio (SNR) of signals, which in turn affects the accuracy of emotion recognition.

3.2.3 Suggestions for improvement

1) Equipment and Cost Optimization:

Lightweight hardware adaptation: Replace traditional multi-lead contact electrodes with wireless dry electrodes (e.g., the NeuroSky MindWave single-electrode EEG device), reduce the number of electrodes to 3-5 key leads (e.g., forehead, occipital lobe), and decrease device size and wearing complexity. Meanwhile, integrate the signal acquisition and preprocessing modules into one unit through chip integration technology, reducing hardware costs to within 2-3 times that of visual devices.

Simplifying the preprocessing workflow: Based on existing feature importance analysis (e.g., differential entropy features with 2Hz frequency band resolution yield the optimal performance), screen core features (e.g., EEG β-wave power spectrum, EOG blink amplitude) to reduce redundant computations. Replace some complex modules with lightweight algorithms (e.g., replacing LSTM with GRU using the MobileNet architecture), and improve real-time performance while ensuring recognition accuracy (correlation coefficient ≥ 0.92), with response latency controlled within 30ms.

2) Improvement of Generalization and Robustness:

Expanding the diverse dataset: Increase subject samples covering different age groups (20-60 years old), physiological characteristics (wearing glasses, skin sensitivity), and driving scenarios (urban congestion, highway cruising). Combine transfer learning technologies (such as domain adaptation) to reduce the recognition error of the model by more than 15% on subjects whose individual characteristics deviate from the mean, thereby enhancing generalization ability.

Integration of anti-interference technologies: Add an electromagnetic shielding layer on the hardware side to reduce in-vehicle electronic interference; design an adaptive noise filtering algorithm (e.g., real-time denoising based on wavelet transform) on the software side; and develop an electrode contact status monitoring module for jolting scenarios. When poor contact is detected, a signal compensation mechanism is automatically triggered to ensure that the recognition accuracy remains above 85% in complex environments.

3.3 Analysis of Emotion Recognition Methods Based on Speech Data

3.3.1 Related models

The driver emotion recognition method based on speech data has core advantages in non-invasiveness and convenience of data acquisition—it does not require the driver to wear additional equipment, and can collect speech signals in real time only through the in-vehicle microphone. Moreover, speech signals can directly reflect the driver's emotional fluctuations (e.g., increased pitch and faster speech rate when angry, and low intonation when sad) and are not disturbed by facial disguise, making it particularly suitable for real-time monitoring scenarios in intelligent cockpits. By integrating deep learning and psychoacoustic technology, this method has further overcome the limitations of traditional speech features, such as weak anti-interference ability and single recognition dimension, and has demonstrated good practicality in complex driving environments.

Li Xiang et al. simulated the auditory mechanism of the human ear based on a psychoacoustic model: first, they enhanced the fatigue-sensitive frequency band of 200 Hz–3 kHz through frequency weighting of the outer and middle ears; then, they converted the linear frequency into 109 critical sub-bands using Bark scale mapping, thereby achieving a more refined characterization of emotion-related frequencies [10]. To address feature errors caused by speech frame asynchrony, the study designed a frame alignment scheme of envelope cross-correlation+frequency-domain cross-correlation and extracted 16-dimensional perceptual features (including modulation differences, noise masking ratio, harmonic structure, etc.) to quantify emotionally abnormal sounds. The experiment was based on 6,208 driving speech samples (covering three states: normal, mild fatigue, and severe fatigue). By adopting the methods of multi-reference sample comparison and FSVM-DBN (Fuzzy Support Vector Machine + Dynamic Bayesian Network) fusion decision, the average recognition accuracy of the model reached 92.4%, among which the recognition accuracy for severe fatigue was as high as 96.1%, with the precision and recall rates being 90.8%

ISSN 2959-6157

and 93.1% respectively—significantly outperforming the traditional detection method that combines linear features with MFCC.

3.3.2 Analysis of model limitations

Contradiction between real-time performance and complexity: The model needs to complete calculations for 109 Bark sub-bands and multi-classifier fusion. The detection time on a 4-core CPU platform ranges from 0.65 to 1.8 seconds, which meets basic real-time requirements. However, when the number of reference samples exceeds 9, the computational load increases significantly, making it difficult to adapt to embedded in-vehicle devices with limited computing power.

Insufficient emotion transferability: The model is mainly designed for fatigue state recognition. Its psychoacoustic features (e.g., noise masking ratio) have low discriminability for emotions such as anger and disgust. If directly transferred to multi-emotion recognition tasks, it is prone to fatigue-anger misjudgments, resulting in weak generalization ability across different emotion types.

3.3.3 Suggestions for improvement

Lightweight Adaptation to In-Vehicle Hardware: Screen psychoacoustic features to retain 8–10 core features (e.g., modulation differences, harmonic structure); compress the number of parameters using model quantization (INT8 quantization) and pruning techniques; replace the DBN fusion decision with a lightweight voting mechanism. On the premise that the reduction in recognition accuracy does not exceed 2%, shorten the detection time to less than 0.5 seconds to meet the requirements of embedded devices.

Cross-Emotion Transfer Optimization: Construct a multi-emotion psychoacoustic database, supplement frequency band weights for emotions such as anger and disgust (e.g., enhance the 1–2 kHz frequency band for anger); introduce transfer learning in model training, use fatigue recognition as a pre-training task, and fine-tune parameters of emotion-related feature layers to improve recognition accuracy across different emotion types. The goal is to reduce the fatigue-anger misjudgment rate to below 5%.

4. Summary

This paper presents a systematic investigation of deep learning-based driver emotion recognition methods within the context of intelligent cockpits. The study commences by analyzing the significant impact of emotional fluctuations on driving safety, followed by a comprehensive review of existing emotion recognition techniques utilizing visual, physiological, and speech signals, alongside their inherent limitations. Subsequently, the technical implementation of three mainstream approaches is elaborated in detail.

For the visual pathway, lightweight convolutional neural network (CNN) architectures, particularly the YOLO series and MobileNetV3, achieve high-precision and low-latency facial emotion recognition. This is accomplished through optimization strategies incorporating attention mechanisms and feature pyramids. Regarding the physiological pathway, one-dimensional convolutional encoders integrated with recurrent neural networks process multimodal physiological signals, including electroencephalogram (EEG) and electrooculogram (EOG). These models effectively extract emotion-related features across temporal, spectral, and spatial domains, thereby ensuring the objectivity of recognition outcomes. For the speech pathway, the integration of psychoacoustic models with deep sequential networks enables the effective capture of emotional sensitivity within speech signals, leveraging analysis of frequency bands and prosodic features.

A critical analysis reveals distinct advantages and limitations inherent to each approach concerning performance, applicability, robustness, and engineering feasibility. Visual methods offer ease of deployment but remain susceptible to environmental interference such as lighting variations and occlusions. Physiological methods provide high accuracy and reliability yet entail significant implementation costs and user discomfort due to sensor requirements. Speech-based methods facilitate natural human-machine interaction but exhibit limited robustness in noisy cabin environments commonly encountered during driving.

Notably, the significant complementarity observed among these three modalities in practical intelligent cockpit scenarios underscores the potential of multimodal fusion strategies for enhanced emotion recognition. The innovation and significance of this research are manifested in three principal contributions. First, the study achieves a systematic integration and comparative analysis within a unified framework specifically tailored to intelligent cockpits. It comprehensively evaluates the performance characteristics and identifies critical bottlenecks across the three major emotion recognition modalities—visual, physiological, and speech—addressing a key gap in existing literature which predominantly focuses on single-modality reviews. Second, the research adopts an improvement-oriented perspective targeting practical deployment challenges. It proposes feasible enhancement strategies, including lightweight network design for resource-constrained environments, high-risk emotion sample augmentation to address incomplete emotion coverage, and hardware-algorithm co-optimization techniques for robust

noise resistance. These strategies collectively advance the technology from a merely functional state towards seamless and user-friendly integration. Third, the study strongly emphasizes the critical importance of multimodal fusion for future development. It advocates for breaking down modality barriers through deep learning-driven fusion of multisource information, thereby significantly enhancing recognition accuracy, real-time performance, and environmental robustness while maintaining an optimal user experience. This direction provides a clear pathway for advancing emotional intelligence capabilities within next-generation intelligent cockpits. This comprehensive analysis establishes a robust foundation for the development of practical and reliable driver emotion monitoring systems.

Authors Contribution

All authors contributed equally to this work.

References

[1] Alfaras M S and Karan O. A Review of Advancements in Driver Emotion Detection: Deep Learning Approaches and Dataset Analysis. 2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), FEZ, Morocco, 2024, pp. 1-9.

[2] Zhou Dongmei, Cheng Yougjian, et al. Drivers' Comprehensive Emotion Recognition Based on HAM. Sensors, 2023, 23, Art. no. 8293.

[3] Terzi R, Sagiroglu S, Demiraglu M U. Big data perspective for driver/driving behavior. IEEE Intelligent Transportation Systems Magazine, 2018, 12(2): 20-35.

[4] Xue H. Research on Driver Speech Emotion Recognition and Regulation Method. Chongqing: Chongqing University, 2022.

[5] Xiang Guoliang, Yao Song, et al. A multi-modal driver emotion dataset and study: Including facial expressions and synchronized physiological signals. Engineering Applications of Artificial Intelligence, 2024, 130, Art. no. 107772.

[6] Jha S, Marzban M F, Hu T, Mahmoud M H, Dhahir N A and Busso C. The Multimodal Driver Monitoring Database: A Naturalistic Corpus to Study Driver Attention," iIEEE Transactions on Intelligent Transportation Systems, 2022, 23(8) :10736-10752.

[7] Liu H. Research and Application of Driver Emotion State Judgment Algorithm Based on Visual Features. Guangzhou: South China University of Technology, 2024.

[8] Zhang J, Xu G. Research on Driver Behavior Recognition Method Based on Deep Learning. Automation Application, 2025, 66(10): 59-62+68.

[9] Shi J. Research on Driver Fatigue Detection Method Based on Intelligent Processing of Physiological Signals. Taiyuan: Shanxi University, 2023.

[10] Li S, Li G, Shi J. Driver Fatigue Detection Based on Speech Psychological Analysis. Chinese Journal of Scientific Instrument, 2018, 39(10): 166-175.