Multimodal Emotion Recognition Empowers Fatigue Driving Detection

Can Gao [1], Yuan Gao [2] and Haoyu Yang [3, *]

[1] School of Information Engineering, Nanhang Jincheng College, Nanjing, 211156, China

[2] SWJTU-Leeds Joint School, Southwest Jiaotong University, Chengdu, 611756, China

[3] School of Information and Electronics, Beijing Institute of Technology, Beijing, 102488, China

* Corresponding Author Email: HYang24@uclan.ac.uk

Abstract. The probability of traffic accidents caused by fatigue is not small, so the reliable fatigue reminder device is an important way to ensure driving safety. Therefore, this paper introduces six kinds of multimodal fatigue driving detection techniques with research significance: (1) a method based on heart rate and Percentage of Eyelid Closure(PERCLOS); (2) a method based on hybrid electroencephalography(EEG) and eye tracking; (3) attention-based approach; (4) a method based on multimodal feature coupled mode; (5) a method based on Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture; (6) a non-invasive method based on heart rate, eye and facial features. Instead of simply listing the existing methods, this paper systematically analyzes the performance of these methods in terms of accuracy, robustness, hardware cost and practical deployment feasibility. The results show that the accuracy of most models is more than 96%. In addition, this paper also critically discusses the main limitations of current methods, and proposes corresponding improvement directions. Finally, this article discusses the application prospects of these technologies in the automotive and transportation fields and clarifies the advantages and disadvantages of these six technologies. This article provides a reference for designing low-cost and easy-to-use in vehicle fatigue driving detection solutions in the field of automotive, bus, aircraft and commercial vehicle fleet. It is helpful to promote the practical application of fatigue detection technology in vehicles and aircraft, thereby improving road safety and reducing the occurrence of fatigued driving accidents.

Keywords: Fatigue Driving Detection, Multimodal Fusion, Attention Mechanism, Intelligent Transportation Systems.

1. Introduction

Driver fatigue is a significant contributing factor to traffic accidents. Globally, the proportion of accidents caused by driver fatigue has been on the rise year by year, accounting for approximately 20% to 30%, which severely undermines traffic safety [1]. With the widespread adoption of advanced driver-assistance systems (ADAS), drivers tend to over-rely on these systems, leading to reduced vigilance. As a result, fatigue-related driving behaviors have become increasingly concealed and difficult to detect, posing severe threats to road safety [2]. However, traditional detection methods mostly rely on single physiological signals or vehicle behavior data, which come with limitations such as low safety of equipment, poor environmental adaptability, and a single dimension of recognition. These methods are prone to misjudgment under complex road conditions or individual differences, and it is difficult to satisfy the needs of real-time accurate monitoring [3]. Therefore, in recent years, scholars have focused on the development of multimodal emotion recognition technology [4]. This technique fuses multi-dimensional data such as facial expressions, speech features, eye movements and physiological signals to achieve cross-modal information complementary. When one mode of data is disturbed by the environment, other modes can provide redundant information to ensure system reliability. At the same time, multimodal systems can capture physiological signal changes such as increased EEG theta waves and decreased alpha activity, which often precede visible fatigue-related behavioral characteristics. This allows for the early warning of driver fatigue [5]. As a crucial application of affective computing, multimodal emotion recognition

technology will promote the development of fields such as intelligent vehicles and ADAS, providing an efficient and reliable technical approach to enhance the active safety capabilities of ADAS.

This study aims to systematically organize, analyze and propose optimizations for six multimodal driver fatigue recognition technologies. By comprehensively collecting and screening relevant research literature in recent years, this study analyzes and evaluates existing multimodal driver fatigue recognition models. It focuses on the fusion strategies, accuracy, robustness and applied models or algorithms of different modal combinations. The study systematically identifies the shortcomings of these six technologies and puts forward theoretical optimization schemes. This helps researchers quickly understand the structural characteristics, modalities, algorithms, and performance of the six methods, thereby reducing the time required for literature review and comparative analysis. On this basis, the study further analyzes how to apply these technologies to optimize the driver fatigue detection methods in the automotive industry and the transportation sector, providing ideas for the improvement of relevant models.

2. Multimodal Fatigue Detection: Theoretical Foundations and Methodological Innovations

In the realm of fatigue driving detection, multimodal data fusion has firmly established itself as a pivotal approach aimed at enhancing both the accuracy and reliability of fatigue detection systems. Multimodal methods can use the complementary nature of different data types to overcome the limitations of single - modality approaches by integrating data from multiple sources, such as physiological signals, behavioral cues, and vehicle-related parameters. This chapter delves into six representative multimodal fatigue detection methods. Moreover, these methods using different data types and fusion strategies. Furthermore, this chapter provides a comprehensive analysis of the contributions of these approaches to the field, highlighting their unique strengths and areas that need further improvement.

2.1. Fusion of Heart Rate and PERCLOS

This method uses non-invasive technology to collect physiological and behavioral data. The physiological data are extracted by remote sensing photoelectric volumetric recording (RPPG) signal, RGB video and unsupervised learning techniques, and the behavioral data are obtained by using eyelid features to obtain PERCLOS value. In the process of processing, an independent 1D CNN models are used to process each mode data respectively, and then the AdaBoost algorithm is used to fuse the results of each model, which significantly improves the overall robustness of the fatigue detection system.

This non-invasive design is a major breakthrough because it eliminates the impact of the discomfort of wearable devices on the normal driving behavior of drivers. In the complex driving environment, there are interference factors such as light change or vehicle vibration, so it is very important to optimize the model to ensure reliable detection. Finally, this method is particularly suitable for non-invasive priority scenes, but its performance may be limited in high noise environments [6].

2.2. Fusion of Hybrid EEG and Eye-Tracking

This innovative method combines EEG data which is collected by 32-channel cap with eye movement data obtained from the eye-tracking glasses and uses CNNs and Multi-Layer Perceptrons (MLPs) to encode neural activity and visual attention patterns respectively. In addition, the method introduces a cross-modal prediction alignment module to enhance the fusion of features from different modes, and adopts a 1D attention mechanism to filter features which are the most relevant for fatigue detection.

Comparing to single-modality methods and those based on canonical correlation analysis (CCA), this approach achieves superior accuracy during within-session, cross-session, and cross-subject tasks.

However, because the method relies on two different devices, the data acquisition process becomes complicated, which not only increases the setup time, but also may cause errors. At the same time, the pre-training process required by the model further increases the complexity of the system, so that a more concise design scheme must be adopted in practical application. Therefore, this method is effective in a controlled laboratory environment, but in real driving scenes, its complex data acquisition requirements may take challenges [7].

2.3. Attention-Based Approach

In order to reduce the interference caused by light changes and head posture, this method employs multiple cameras, including RGB, Near-Infrared (NIR), and depth cameras, and combines the attention mechanism in the attention network. In addition, the technology successfully completes yawning detection and fatigue level discrimination by extracting facial features through weighted and fused views of different cameras, and achieves an accuracy rate of 96%.

Although this method can maintain excellent performance under different environmental conditions, this scheme is overly dependent on facial expression features, resulting in the failure to fully integrate other important physiological signals. Therefore, in order to achieve more comprehensive fatigue detection, it is necessary to integrate more modal data. As a result, this method is suitable for scenes where facial expression is the main fatigue indicator, but other important fatigue-related signals may be missed [8].

2.4. Optimized Driver Fatigue Detection Method Using Multimodal Neural Networks

This model dynamically couples EEG, Electrocardiogram (ECG), and facial features through a neural network. In this network, different modalities mutually influence each other's contributions. For example, physiological signals can guide the extraction of facial features, and vice versa. In the DROZY data set, the model achieved an accuracy of 98.41%, fully demonstrating its powerful ability in capturing fatigue patterns.

Moreover, the dynamic feature interaction is a key innovation that enhances the model's predictive power. However, the high cost of computation and complex model structure leads to significant challenges to real-time applications. Therefore, the researchers need to optimize its efficiency and make it more suitable for real driving scenes. As a result, this method shows great potential in high precision fatigue detection, but it needs to be optimized for real-time performance [9].

2.5. Fusion of CNN–LSTM Architecture

This hybrid model deals with vehicle-related data, facial data, and physiological data such as Fatigue-Associated Symptom Index (FASI) and Neurological Alertness Unified Test (NAUT) through combining CNN (for feature extraction) and LSTM networks (for capturing temporal dependencies). In addition, the technique uses a stacked architecture including convolutional, pooling, LSTM, flattening, and dense layers, and uses the Adam optimization algorithm to achieve 96% accuracy.

The model is excellent in processing time series data and can effectively capture the dynamic evolution process of fatigue state. However, in the complex driving scene with large noise interference, its performance will be significantly reduced, which highlights the need for adaptive mechanism to enhance model stability in noisy environment. Therefore, although this method is suitable for long-term fatigue trend analysis, it still needs to be improved for noise interference in real driving environment [10].

2.6. Fusion of Heart Rate, Eye, and Facial Features

Combining infrared cameras, RPPG, and Multi-task Cascaded Convolutional Networks (MTCNN), this method enhances feature extraction by using an improved Pan-Tompkinn algorithm and 1D-MTCNN. Moreover, employing Bidirectional Long Short-Term Memory (BiLSTM) networks to

model bidirectional temporal dependencies. Therefore, this method achieves 98.2% accuracy and 95% Heart Rate Variability (HRV) identification under complex conditions.

Its high robustness in real-world settings makes it a promising approach. However, its dependency on infrared technology may limit its scalability, especially in cost-sensitive applications. Therefore, exploring cost-effective alternatives is essential for wider adoption. Finally, this method is well-suited for high-precision fatigue detection in relatively well-equipped vehicles but needs to address cost issues for mass-market application [11].

2.7. Synthesis and Future Directions

These methodologies collectively display the immense potential of multimodal fusion technology in addressing fatigue detection challenges. Key trends include a focus on non-invasiveness to prioritize driver comfort without sacrificing accuracy, the use of attention mechanisms to dynamically focus on fatigue-relevant features, and temporal modeling via LSTM/BiLSTM to capture sequential dependencies.

However, there are still trade-offs between complexity, computational cost, and environmental adaptability. Therefore, future research should focus on achieving hybrid simplicity, balancing model depth with real-time feasibility. In addition, it is also crucial to develop cross-modal synergy by creating unified frameworks that seamlessly integrate diverse data types. Additionally, improving scalability by reducing reliance on specialized hardware (e.g., EEG caps) will enable broader applicability. By addressing these gaps, the field can advance toward practical, high-performance fatigue detection systems tailored to real-world driving conditions.

3. Comparative Analysis of Multimodal Fatigue Detection Methods

The results of the six methods show that multimodal fusion is significantly better than single mode on the whole, and their accuracy rates are generally over 96%. However, there are obvious differences in fusion level, robustness and application value among different methods. The detailed comparison of these six methods is summarized in Table 1.

Table 1. The detailed comparison of these six methods

Method	Application	Research	Advantage	Major defect
Heart rate + PERCLOS	Relatively Limited	Relatively limited	High comfort; Easy to deploy	Shallow fusion, low noise robustness
Hybrid EEG+eye-tracking	Relatively low	Relatively the highest	High environmental adaptability; Accurate feature highlighting	High cost, sensitive to noise and placement
AMMF	Relatively moderate-to-high	Relatively high	Simple deployment; Low cost; High robustness	High computational complexity
Multimodal feature coupled mode	Relatively Limited	Relatively the second-highest	High integration depth, high accuracys	High computational cost and complexity
CNN + LSTM architecture	Relatively moderate	Relatively moderate	High stability; Strong dynamic performance	Sensitive to noise, limited generalization
non-invasive	Relatively the highest	Relatively high	Best cost-effectiveness, easy deployment	Shallow Shallow cross-modal interaction

The multimodal feature coupled model which adopts feature-layer fusion achieves 98.41% on the DROZY data set. Furthermore, its advantage lies in realizing the dynamic interaction between physiological signals and facial features in the feature extraction stage, and reducing information loss through weight adjustment and deep fusion. This design allows the method to achieve the highest level of accuracy of all methods. However, the model structure is complex, the calculation is large,

and the requirements for computing power and energy consumption are high, so this method faces deployment challenges in real-time applications. In contrast, the non-invasive approach (heart rate + eye + facial features) achieved 98.2% on the MDAD data set, which was slightly lower than the feature coupling model, but the gap is not huge. Its key advantage is that the enhanced Pan-Tompkins algorithm is used for heart rate signal enhancement, and the BiLSTM is combined to catch temporal features, which significantly improves the quality of non-invasive signals. However, due to the lack of complex feature layer interaction, it still has some deficiencies in cross-modal information utilization, so the accuracy is slightly lower than that of the feature coupled model. Nevertheless, this method has more advantages in practical deployment and large-scale commercial promotion because of its non-contact collection mode, low hardware requirements and good comfort. In contrast, the accuracy of EEG and eye-tracking methods is lower than that of the previous two methods, because EEG signals are highly dependent on electrode contact quality, which is susceptible to noise and electrode position deviation interference. Furthermore, timing errors can also occur in eye signals when cross-mode alignment is performed. However, this method has strong environmental adaptability and can highlight key features well, reduce redundancy. The accuracy of the attention-based approach (AMMF) (96.61%) is close to that of EEG and eye-tracking methods, but because the latter rely on invasive sensors, deployment costs and difficulties are high. Therefore, AMMF has more advantages in the application level due to its easy deployment. However, because of insufficient fusion depth and time series modeling, the accuracy of this method is lower than that of multimodal feature coupled model and non-invasive method. The method of CNN and LSTM architecture also achieves 96%, with the advantage of being able to combine spatial and temporal features to maintain stability in dynamic driving scenarios. However, due to its sensitivity to physiological signal noise and limited generalization, compared with the previous methods, this method has some disadvantages in complex environments. Heart rate and PERCLOS method is insufficient to adapt with complex driving environments because of its single-modality and shallow fusion characteristics, So its accuracy rate is only 87.1 percent. Nevertheless, the technology is non-invasive (does not require wearable devices), avoiding discomfort and requiring only a normal camera, so that this method is also somewhat competitive.

From the perspective of cost-effectiveness and deployment, non-invasive methods are the most feasible in practical applications due to their near-top accuracy, and low hardware requirements. Although hybrid EEG and eye tracking method represents the theoretical precision limit, it is more suitable for scientific research benchmark or high-risk scenarios due to its high cost. In conclusion, the revelation of this chapter comparison is: feature coupled model represents the performance limit, attention mechanism ensure stability in complex environment, non-invasive method has the most advantages in industrial implementation, which is also the preferred direction for real-time systems.

4. Challenges and Optimization Ideas

Although the methods mentioned above have excellent performance, each of them has its limitations. This section will discuss into these issues and propose targeted optimization ideas.

The first challenge involves hardware cost and deployment: The non-invasive multimodal fusion method relies on infrared cameras (e.g., Oni S500), which require additional deployment and result in high hardware costs. The attention-based multimodal multi-view fusion method depends on specific hardware (e.g., Qualcomm SliM 3D sensor, Intel RealSense camera), leading to relatively high costs. The multimodal fusion method based on heart rate and PERCLOS relies on a single RGB camera. In complex environments, such as extreme light conditions, the camera is susceptible to interference, affecting detection stability. Moreover, the feature extraction process is complex and has certain requirements for the computing capacity of the hardware.

To address these issues, several optimization ideas are proposed. For the non-invasive multimodal fusion method, alternative sensing strategies can be explored—for instance, replacing infrared cameras with low-cost near-infrared-enhanced RGB cameras combined with lightweight super-

resolution algorithms to maintain data accuracy under varying illumination. For the attention-based multimodal multi-view fusion method, a hybrid design incorporating an ordinary RGB camera along with a low-cost TOF (Time-of-Flight) sensor could be adopted. Sensor calibration and depth-image translation techniques may help maintain accuracy while reducing costs. For the multimodal fusion method based on heart rate and PERCLOS, the model architecture should be simplified through techniques such as model pruning or quantization. Additionally, introducing a multi-sensor compensation mechanism—for example, fusing IMU (Inertial Measurement Unit) data for motion robustness—could enhance overall system stability without significantly increasing costs.

The second major challenge is the inadequate extraction of comprehensive modal features: In the CNN-LSTM-based driver fatigue recognition method, the physiological signals only include respiratory data and heart rate data. The method relies on obvious features and fails to incorporate deep-level feature data. The method based on the multimodal feature coupling model mainly relies on three types of modal data: EEG, ECG and facial images. This method lacks other key features, and existing detection methods may not cover some fatigue manifestations. In addition, the multimodal driver fatigue detection method combining EEG and eye movement tracking data does not include other key features related to fatigue. Furthermore, the sensitivity of the existing detection mode varies at different fatigue stages, which makes it difficult to fully cover the dynamic process of fatigue development.

In order to overcome the limitations of existing methods, we propose the following optimization scheme. For the method based on CNN-LSTM, physiological indicators such as EEG, EDA, EMG and skin temperature could be integrated. It is particularly recommended to use dry electrode EEG equipment, which can not only improve the convenience of use, but also maintain signal quality. Furthermore, fine facial movements can be captured by using the OpenFace toolkit. For the method based on multimodal feature coupling model, it is suggested to expand the modal dimension, such as adding environmental awareness data and behavioral signals. In addition, the introduction of structured hierarchical fusion architecture with attention-based dynamic weighting can enhance the inter-modal interaction and feature adaptive selection ability. For the multimodal driver fatigue detection method combining EEG and eye movement tracking, it is necessary to construct a more comprehensive multimodal feature system. For example, the introduction of head pose estimation or blink dynamic information can supplement key auxiliary data. In order to improve temporal adaptability, a dynamic weighted fusion mechanism such as a GRU (Gated Recurrent Unit) which is based adaptive weighting network could be designed to automatically adjust feature importance according to different fatigue phases.

5. Practical Applications and Implementation Prospects

This chapter delves into the specific application prospects of the reviewed multimodal fatigue detection technologies in the automotive and transportation sectors. It will analyze how these technologies can address the pain points of existing systems and discuss the potential challenges in real - world deployment.

5.1. Automotive Industry Applications

Tesla: Currently, Tesla mainly uses cameras and sensors to read the driver's facial posture, steering wheel operation, vehicle speed, and lane changes for fatigue detection. This approach focuses on behavioral and vehicle status data, ignoring the driver's internal physiological state. A CNN-LSTM method that fuses vehicle, visual, and physiological signals can be introduced. By adding physiological data like heart rate and EEG signals to the original data, CNN can extract features from multi-source data, and LSTM can analyze time-series data to more accurately assess fatigue.

Mercedes Benz: The Attention Assist system combines steering wheel input, vehicle driving patterns, and driver facial features. However, it lacks deep fusion and interaction between different modal data. A multimodal feature coupling model can be integrated. This model enables features

from different modalities to influence each other's contributions to the final prediction, improving detection accuracy and providing more reliable warnings.

BYD: BYD uses cameras to capture facial images and AI to grade blink frequency and eye opening/closing amplitude for fatigue and distraction detection. This method mainly relies on facial visual features and is insensitive to physiological state changes. A multimodal fusion method based on heart rate and PERCLOS can be added. By combining heart rate detection with PERCLOS and facial image analysis, a more accurate fatigue assessment can be achieved.

5.2. Transportation Sector Applications

Public Buses: Currently, infrared cameras and sensor control units monitor eye movements, gaze direction, and head tilt for fatigue detection in public buses. This method lacks physiological characteristic utilization. A non-invasive multimodal fusion method based on heart rate, eye, and facial features can be used. By extracting physiological features like heart rate and HRV using RPPG and MTCNN and integrating eye movement and facial feature information, a more comprehensive fatigue state can be reflected.

Commercial Vehicle Fleets: A single RGB camera and traditional computer vision methods are used to detect eye features for fatigue detection in commercial vehicle fleets. This method performs poorly under varying lighting conditions. A multimodal Multiview fusion method based on an attention mechanism can be employed. Multiple cameras collect facial expression data from different perspectives, and the attention mechanism helps the model focus on important fatigue-related feature regions, improving adaptability in complex environments.

Aviation Field: Non-contact monitoring combining eye movements, posture, and physiological signals is used for flight fatigue prevention in aviation, but it ignores EEG signals. A driving fatigue detection method based on hybrid EEG and eye-tracking can be adopted. By combining EEG and eye-tracking data, a more comprehensive and accurate assessment of the pilot's fatigue level can be made, addressing the shortcomings of the existing scheme.

In conclusion, this article has analyzed and compared six multimodal fatigue detection methods, all of which demonstrate accuracies exceeding 85%. It also discussed their hardware cost, deployment, and modal extraction shortcomings and proposed corresponding solutions. Additionally, it presented specific applications of these technologies in the automotive and transportation industries.

6. Conclusion

This article has conducted a comprehensive review of six prominent multimodal fatigue detection methods. Furthermore, this paper contrasts these methods, finding the accuracy of all methods exceeded 85%. Moreover, the paper also discusses the shortcomings of these technologies in terms of hardware cost and deployment and lack of comprehensive modal extraction. For the future, corresponding solutions are put forward according to the shortcomings of these technologies. Meanwhile, this paper also comes up with the specific applications of these six technologies in the automotive industry and transportation industry.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Y. Peng et al. A multi-source fusion approach for driver fatigue detection using physiological signals and facial image. IEEE Transactions on Intelligent Transportation Systems, 2024, 25 (11): 16614 - 16624.

[2] L. Mou et al. Driver emotion recognition with a hybrid attentional multimodal fusion framework. IEEE Transactions on Affective Computing, 2023, 14 (4): 2970 - 2981.

[3] P. Zhao, C. Lian, B. Xu, Z. Zeng. Multiscale global prompt transformer for EEG-Based driver fatigue recognition. IEEE Transactions on Intelligent Transportation Systems, 2025, 22: 2700 - 2711.

[4] Shahbakhti M, Beiramvand M, Nasiri E, Far SM, Chen W, Sole-Casals J, Wierzchon M, Broniec-Wojcik A, Augustyniak P, Marozas V. Fusion of EEG and eye blink analysis for detection of driver fatigue. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023, 31: 2037 - 2046.

[5] Kong L, Xie K, Niu K, He J, Zhang W. Remote photoplethysmography and motion tracking convolutional neural network with bidirectional long short-term memory: non-invasive fatigue detection method based on multi-modal fusion. Sensors, 2024, 24 (2): 455.

[6] Li J, Li Y, Wang H. A multimodal fusion fatigue driving detection method based on heart rate and PERCLOS. International Journal of Transportation Science and Technology, 2022, 11 (4): 846 - 857.

[7] Lian Z, Xu T, Yuan Z, Li J, Thakor N, Wang H. Driving fatigue detection based on hybrid electroencephalography and eye tracking. IEEE Journal of Biomedical and Health Informatics, 2024, 28 (11): 6568 - 6576.

[8] Chen J, Dey S, Wang L, Bi N, Liu P. Attention-based multimodal multi-view fusion approach for driver facial expression recognition. IEEE Access, 2024, 12: 137203 - 137215.

[9] Cao S, Feng P, Kang W, Chen Z, Wang B. Optimized driver fatigue detection method using multimodal neural networks. Scientific Reports, 2025, 15 (12240): 1 - 15.

[10] Priyanka S, Shanthi S, Kumar AS, Praveen V. Data fusion for driver drowsiness recognition: A multimodal perspective. Egyptian Informatics Journal, 2024, 27: 100529.

[11] Kong L, Xie K, Niu K, He J, Zhang W. Photoplethysmography and motion tracking convolutional neural network with bidirectional long short-term memory: Non-invasive fatigue detection method based on multimodal fusion. Sensors, 2024, 24: 455.