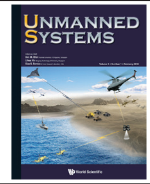
Unmanned Systems, Vol. 14, No. 1 (2026) 143–156

DOI: 10.1142/S2301385025500864

UNMANNED SYSTEMS
https://www.worldscientific.com/worldscinet/us



Equipping with Human Cognition: Driver Intention Recognition with Multimodal Information Fusion

Bo Zhang [*,†], Xiaohui Hou [*,†,‡], Wei Wu [*,†], Minggang Gan [*,†]

[*] *School of Automation, Beijing Institute of Technology, Beijing 100081, P. R. China*

[†] *National Key Lab of Autonomous Intelligent Unmanned Systems, Beijing Institute of Technology, Beijing 100081, P. R. China*

To address the issue of autonomous driving systems' inability to timely detect danger and respond in urban environments, an intention recognition system integrating driver cognitive information was developed. By utilizing the drivers' electroencephalogram (EEG), eye movement and operational data, the system identifies driver intentions in a type of hazardous scenario. Initially, a driver-in-the-loop simulation platform was used for data collection, followed by experiments and data preprocessing to create a multimodal fused dataset through feature-level fusion. Models based on multilayer perceptron (MLP), convolutional neural network (CNN) and transformer were then developed to predict emergency braking and steering evasion intentions. The transformer-based model, with multimodal data fusion, achieved the best performance with an accuracy of 93.02%, significantly surpassing EEG-only (80.80%) and eye movement and operational data-only models (78.46%). This highlights the transformer's superior ability to capture complex spatiotemporal correlations in multimodal data. Additionally, pre-extracted EEG frequency domain features could improve model performance, though less significantly than changes in model architecture. Embedding this system into autonomous driving systems is expected to enhance their ability to quickly and accurately recognize and respond to dangerous scenarios.

Keywords: Multimodal information fusion; automatic driving; intention recognition; human–machine cooperation.

1. Introduction

Autonomous driving systems are currently deployed in relatively simple environments with minimal external disturbances, such as highways and expressways. However, in complex urban traffic scenarios with dynamic spatial conditions and numerous participants, onboard sensors, particularly vision-based cameras, face significant challenges. The interplay of light, shadow and the presence of dense, colorful billboards in urban settings increases the likelihood of sensor misinterpretation. Urban environments also present a higher risk of accidents due to diverse road users and unpredictable behaviors with potentially more severe consequences. These challenges make hazard detection and timely responses more difficult, often requiring human intervention [1, 2]. Moreover, existing autonomous systems lack a "human-like" design, potentially conflicting with human drivers' expectations and real-time reactions, leading to discomfort, tension among passengers and drivers and even traffic congestion or accidents [3]. To address these issues, integrating cognitive information from drivers to assist autonomous driving systems can enhance their perception and decision-making capabilities. This approach also holds promise for achieving a more human-like driving experience in autonomous vehicles [4–6].

Electroencephalogram (EEG) and eye movement are two commonly used physiological signals that capture drivers' cognitive information. EEG reflects driver's operational intentions, providing insights into processes such as perception, judgment and decision-making, and serves as a reliable reference for autonomous driving systems, compensating for the limitations of onboard sensors [7]. Eye movement, on the other hand, offer additional

Received 18 July 2024; Accepted 31 October 2024; Published 24 December 2024. This paper was recommended for publication in its revised form by editorial board member, Shenghai Yuan.
Email Address: ‡*houxh1220@163.com*
‡Corresponding author.

information on drivers' attention and awareness, aiding in the assessment of their vigilance and understanding of the surrounding environment [8].

Several studies on driver intention recognition based on EEG and eye movement information have been published in recent years [9–21]. Haufe *et al.* designed an EEG-based emergency braking intention detection experiment on a simulation platform and achieved prediction 130 ms prior to the actual braking event using a linear discriminant analysis (LDA) classifier [10]. Hernández *et al.* explored the feasibility of detecting emergency braking intentions using EEG under different cognitive states in a simulated driving environment. Their results indicated that the average classification accuracy of support vector machine (SVM) and convolutional neural network (CNN) models exceeded 70% [13]. Teng *et al.* proposed an LDA model based on spatial-frequency features of EEG for detecting emergency braking intentions, achieving an accuracy of up to 94% [14]. As for eye movement, Wu *et al.* proposed a framework based on probabilistic dynamic time warping and hidden Markov models that efficiently predicts driver intentions in semi-autonomous vehicles following scenarios by analyzing the driver's eye movements and gaze patterns [17]. Deng *et al.* demonstrated that fusing eye movement data with environmental information can accurately predict a driver's lane-changing intention [20]. Guo *et al.* analyzed eye movement, head movement, vehicle and driving operation data during simulated driving and developed a lane-changing intention recognition model, achieving a recognition accuracy of 93.33% within 3 s before the lane change occurred [21].

In the above studies, EEG-based models have primarily focused on identifying braking intentions, while eye movement-based models have centered on steering intentions. Integrating these modalities could enable simultaneous prediction of both braking and steering intentions, allowing each signal type to complement the other and offset potential errors or interference from a single modality, thereby enhancing data reliability [22].

Some research has already explored combining EEG and eye movement signals for driving applications [23–29]. Guo *et al.* developed a logistic regression model based on multi-source data from EEG, eye movements and driving performance to estimate driver alertness, using the driver's reaction time to the lead vehicle's braking signals [25]. Yang *et al.* proposed a multimodal attention recognition network with a decision-level fusion of EEG, eye movements and vehicle state data to assess driver load, achieving notable model performance [26]. Hu *et al.* introduced data and decision-level fusion methods for EEG and eye movement data to predict driver decision intentions at intersections, validated through simulation experiments [27]. Zhu *et al.* integrated EEG, eye movements and head movements to develop a model using risk potential fields and XGBoost, aiming to predict driver takeover quality under varying levels of nondriving-related tasks [28].

However, most studies that integrate multiple physiological signals have focused on assessing driver load, stress, or alertness, with limited exploration in the area of driver intention recognition. Furthermore, simulation environments of existing studies on intention recognition using physiological signals are generally designed around car-following scenarios or short, fixed-route segments, which restricts the real-world applicability of these works.

Based on previous research and addressing its limitations, the main contributions of this paper are as follows:

(i) **Novel system architecture**
A driver intention recognition system based on cognitive information was designed, which integrates EEG, eye movement and operational data into a driver intention recognition model for the first time. This system can simultaneously predict braking and steering intentions, achieving the best performance with an accuracy of 93.02%, providing a new perspective for addressing current limitations in perception and decision-making in autonomous driving systems.

(ii) **Realistic experimental scenarios**
Unlike previous studies that typically use simple, specific road segments, the simulation scenarios in this study encompass a variety of common urban road environments, with greater randomness in the occurrence of hazardous situations, offering stronger real-world applicability.

(iii) **Data enhancement and deep learning**
This study employs multiple approaches to enhance EEG data, including scientific preprocessing methods and data segmentation using overlapping sampling windows, ensuring the system remains effective at any moment between hazard occurrence and intervention. Additionally, deep learning-based intention recognition models developed in this work could reduce the need for specialized expertise and time costs for researchers.

(iv) **Comprehensive discussion**
This paper discusses the impact of various factors on the driver intention recognition model, including the effect of EEG frequency-domain feature extraction, model architecture and input modes, providing a valuable reference for future research.

This paper is organized as follows. In Sec. 2, the development of the driver-in-the-loop simulation platform and the experimental framework is described. The preprocessing of raw experimental data to create a multimodal dataset is explained in Sec. 3. Section 4 addresses the construction of deep learning-based driver intention recognition models. Section 5 presents and analyzes the experimental results, and the conclusion is provided in Sec. 6.

2. Platform Construction and Data Collection

2.1. *Driver-in-the-loop simulation platform*

The experimental platform structure, as shown in Fig. 1, involved participants wearing both eye-tracker and EEG acquisition devices. These devices collected two types of physiological data — EEG and eye movement — which are indicative of the driver's cognitive state. The participants operated vehicles within simulated scenarios on Computer 1 through a driving simulator. The data collected by the EEG acquisition device was transferred to Computer 2 via Bluetooth for recording. The eye-tracker and driving simulator were connected to Computer 1 through wired cables, with Computer 1 transmitting the eye movement data and driving operation data to Computer 2. These physiological signals, which reflect the driver's cognitive information, would be subsequently fused with the vehicle operation data for integrated processing, and the necessary data preprocessing steps were conducted on Computer 2.

The driving simulation in the experimental platform is developed based on the town maps in CARLA, incorporating various road types and common urban structures. The designed hazardous driving scenario in this study involves a pedestrian suddenly crossing the road from the right-front of the vehicle while the driver is driving normally. As shown in Fig. 2, the solid dot in Fig. 2(a) represents the generated position of the pedestrian, and the dashed arrowed line represents the pedestrian's trajectory. Figure 2(b) presents the simulation result of this scenario in CARLA from a third-person perspective, while the actual experimental view in the simulation is from the driver's first-person perspective.

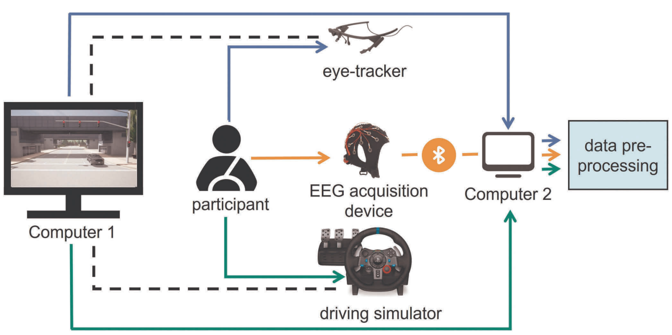


Fig. 1. Experimental platform structure.

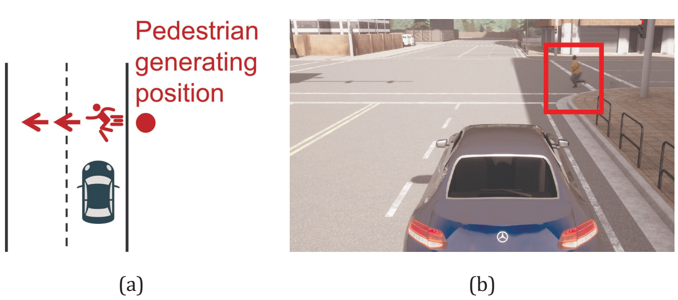


Fig. 2. Design of dangerous driving scenarios.

Figure 3 illustrates the calculation method for generating pedestrian position coordinates, where the solid dot represents the pedestrian's generated position. The xy-coordinate system corresponds to the world coordinate system. Assuming the position of the vehicle driven by the driver in the world coordinate system is (x_1, y_1), and the complementary angle of the vehicle body coordinate system relative to the world coordinate system along the z-axis (yaw angle) is denoted as α, the pedestrian's generated position is (x_2, y_2). The calculation method for (x_2, y_2) is as follows:

$$\begin{cases} x_2 = x_1 + d_1 \cos\alpha - d_2 \sin\alpha, \\ y_2 = y_1 + d_1 \sin\alpha + d_2 \cos\alpha, \end{cases} \tag{1}$$

The definition of d_1 and d_2 is shown in Fig. 3.

The generating frequency, walking speed and relative distance to the vehicle of pedestrians in this scenario can be adjusted by modifying the parameters. Specifically, the setting of the pedestrian's appearance location is modified based on the time to collision (TTC). At the moment, when the pedestrian appears, assuming the vehicle's acceleration relative to the pedestrian is zero, the calculation for TTC is as follows:

$$\mathrm{TTC} = d_1 \div |v|, \tag{2}$$

the definition of d_1 is shown in Fig. 3, v is the speed of the vehicle relative to the pedestrian.

By modifying the TTC value, the level of danger in this scenario can be adjusted. Conducting multiple repeated tests and collecting data for scenarios with different levels of danger can enhance the generalization capability of the intention recognition model. In this study, TTC values of 1.6, 2.0 and 2.4 were selected to correspond to the low, medium and high levels of danger, respectively.

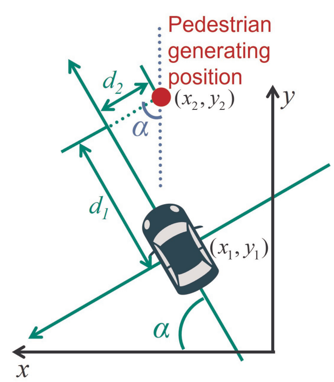
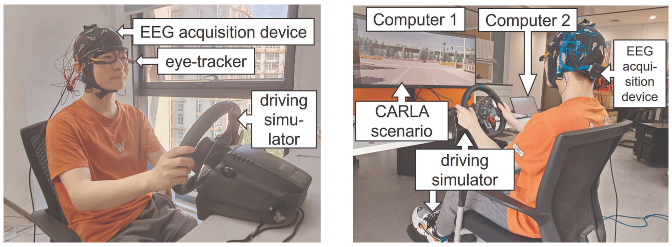


Fig. 3. Generate pedestrian coordinate calculation.



(a) (b)

Fig. 4. Working effect of the driver-in-the-loop simulation platform.

The cooperative effect of the software and hardware in the driver-in-the-loop simulation platform is depicted in Fig. 4.

2.2. *Experimental process*

This study recruited five participants with varying driving styles for testing, each holding a driving license of at least Class C issued by the People's Republic of China. They were all informed and agreed to the experiment in advance, and none of the participants fell sick during or after the experiment.

The experimental process consisted of three parts: Equipment calibration and verification, pre-experiment and formal experiment. During the pre-experiment phase, participants familiarized themselves with the experimental setup and operational procedures to ensure a clear understanding of the tasks and to prevent any physiological discomfort in the simulated environment. In the formal experiment, each test session lasted 5 min, during which participants completed designated driving tasks on specified routes and avoided collisions with randomly appearing pedestrians. A 2 min inter-group break followed the completion of each test session to prevent fatigue driving. During the formal experiment phase, each participant underwent six test sessions across three different levels of danger.

Despite the number of participants may limit the generalizability, each test session involved multiple random triggers of hazardous scenarios to increase the data volume. Additionally, overlapping sampling windows for data segmentation and cross-validation methods were employed to data augmentation and enhance the reliability, respectively.

3. Data Preprocessing and Multimodal Data Fusion

The raw data collected through experiments contain significant noise and interference, necessitating preprocessing to enhance data quality. Additionally, some parameters, such as the start times and sampling frequencies, differ across different data collection devices. Therefore, multimodal data fusion is required to facilitate further analysis.

3.1. *EEG data preprocessing*

Preprocessing of EEG data primarily involved data selection and electrode positioning, band-pass filtering for noise removal, re-referencing and independent component analysis (ICA) for artifact removal [30, 31].

(i) **Data selection and electrode positioning**
Redundant information was removed from EEG data, retaining voltage data from 32-channel electrodes positioned according to the international 10-20 system.

(ii) **Band-pass filtering for noise removal**
A finite impulse response filter was applied for band-pass filtering from 0.5 Hz to 32 Hz to reduce noise. For an input time series $x(n)$, the output $y(n)$ of an Nth order FIR filter is computed as follows:

$$y(n) = \sum_{k=0}^{N-1} h(k)x(n-k), \tag{3}$$

where $h(k)$ represents the filter coefficients.

(iii) **Re-referencing**
All electrodes in this study were symmetrically distributed. Thus, the EEG data were re-referenced using the average of all electrodes to minimize the impact of reference electrodes on experimental data.

(iv) **ICA for artifact removal**
ICA, based on the statistical properties of signals, effectively separates overlapping EEG signals, removing noise and artifacts from the original EEG data to enhance data quality. The main concept is that for an observed signal vector $\mathbf{X}$, assuming the source signal vector is $\mathbf{S}$, the relationship can be expressed as follows:

$$\mathbf{X} = \mathbf{AS}, \tag{4}$$

where $\mathbf{A}$ is the mixing matrix. The goal of ICA is to find the unmixing matrix $\mathbf{W}$ that provides the best possible approximation $\mathbf{Y}$ of the source signals $\mathbf{S}$, such that

$$\mathbf{Y} = \mathbf{WX}. \tag{5}$$

In this study, the Infomax method is used to separate independent components by maximizing the entropy of the output signals, i.e. maximizing $H(\mathbf{Y})$.
The matrix $\mathbf{W}$ is updated using gradient ascent:

$$\Delta\mathbf{W} = \eta[(\mathbf{I} + (1 - 2g(\mathbf{Y}))\mathbf{Y}^T)]\mathbf{W}, \tag{6}$$

where η is the learning rate, $\mathbf{I}$ is the identity matrix and g is the tanh function. After separation, artifact components in the EEG signals are identified and removed based on the EEG topographic maps, refer to typical artifact paradigms such as muscle activity, eye blinks and channel noise. Examples of scalp topographic maps and power spectra of these three typical artifact components are shown in Fig. 5.

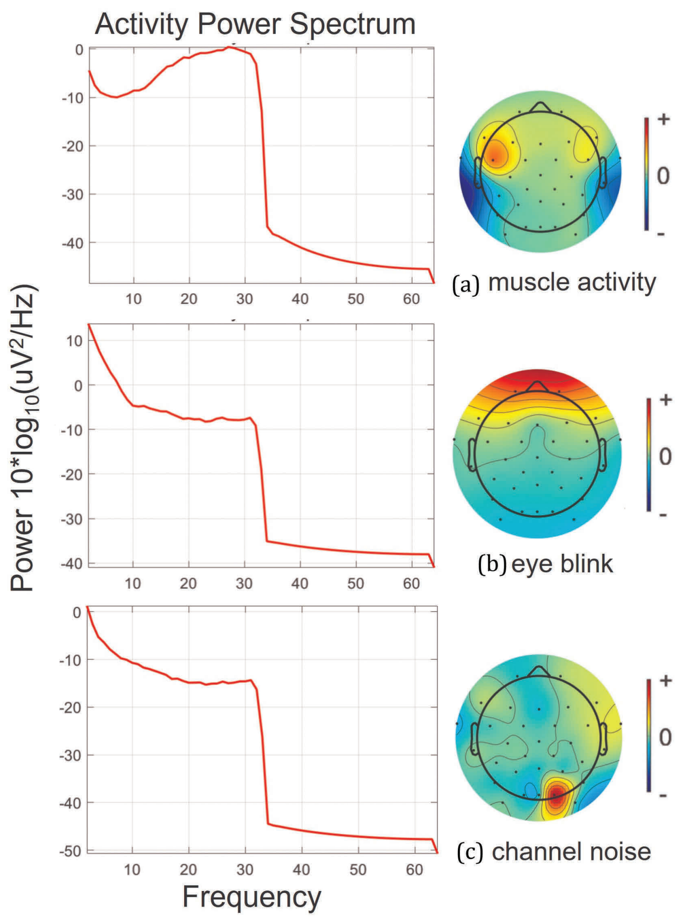


Fig. 5. Examples of scalp topographic maps and power spectra of three typical artifact components.

The 32-channel waveform curves of a segment of EEG data before and after preprocessing are shown in Fig. 6. It can be observed that the EEG data preprocessing reduced noise and artifacts interference present in the original data.

3.2. *EEG domain feature extraction*

To facilitate comparison with the automatic feature extraction capabilities of deep learning-based intention recognition models, this paper manually extracted frequency domain features from EEG after data preprocessing [32].

The EEG data were sampled at a frequency of 128 Hz. Each 40 EEG sampling instances formed a sampling window with a step size of one sampling instance. Within each window, the power in different frequency bands of the EEG was computed, segmented as shown in Table 1. The power was obtained by first calculating the power spectral density (PSD).

For a discrete-time signal $x[n]$ of length N, the method for calculating its PSD is as follows:

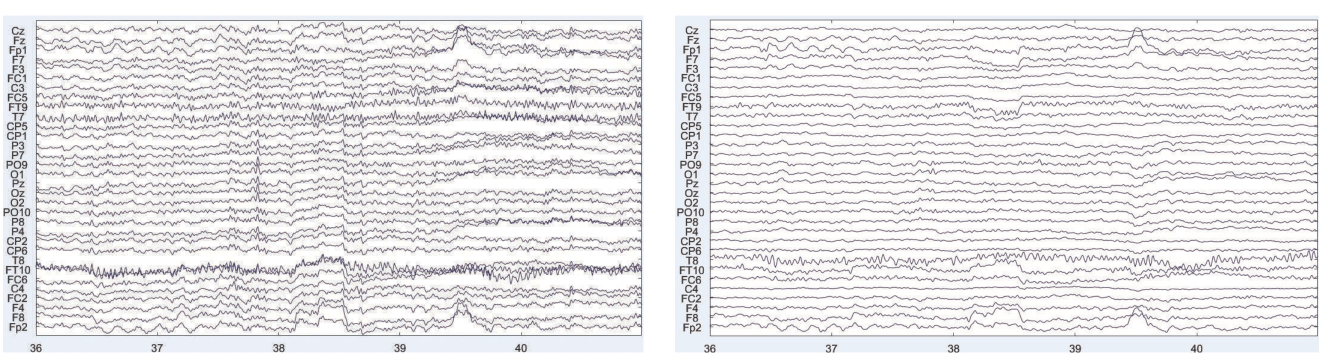
$$\hat{S}_x(f) = \frac{1}{N}|X_N(f)|^2, \tag{7}$$

where $X_N(f)$ is the discrete Fourier transform (DFT) of $x[n]$.

In this study, to reduce the impact of spectral leakage, each segment of the signal is processed using a Hann window before computing the PSD. The Hann window function is defined as follows:

$$w(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right), \quad n = 0, 1, \ldots, N-1. \tag{8}$$

The total power in a specific frequency band is obtained by summing the PSD values over the desired frequency



(a) A segment of EEG before EEG data preprocessing.

(b) A segment of EEG after EEG data preprocessing.

Fig. 6. Effect of EEG data preprocessing.

Table 1. EEG frequency segmentation.

Name	Frequency range (Hz)
theta	4–8
alpha	8–12
betaL	12–16
betaH	16–25

range:

$$P_{\text{band}} = \sum_{f_1}^{f_2} \hat{S}_x(f)\Delta f, \tag{9}$$

where f_1 and f_2 represent the lower and upper frequency limits, and Δf is the frequency resolution.

Due to the computational intensity involved in calculating frequency domain power features across all channels, and the similarity in waveform changes across some certain channels, the 32-channel EEG data was partitioned, as illustrated in Fig. 7.

Among EEG frequency bands, specific waves are strongly correlated with various cognitive and physiological processes that are critical to driving. For example, alpha waves are closely associated with attentional processes. Research has shown that decreased alpha activity is linked to heightened attention and mental effort, making it a reliable indicator for monitoring a driver's focus. Theta waves, on the other hand, are involved in working memory and somatosensory processing, which play essential roles in reacting to driving stimuli. Beta waves, particularly, are tied to driver alertness. Elevated levels of beta activity have been found to promote stimulation and arousal. However, excessive beta wave activity may lead to anxiety and stress.

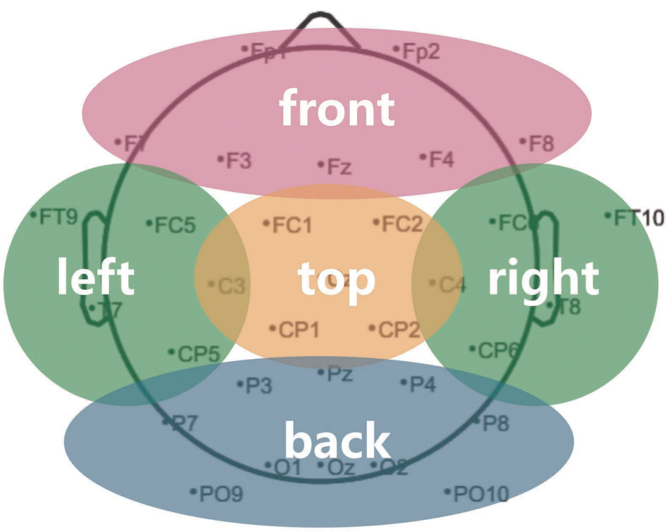


Fig. 7. EEG electrode partitioning.

Further, the beta band is divided into betaH and betaL in this paper for more refinement. For each frequency band, the average power across electrodes within each partition was considered as the power feature for that frequency band in that partition [32–34].

Based on preliminary analysis of experimental data and to reduce computational load, five "region_band" powers — (1) front_alpha, (2) front_betaL, (3) top_betaL, (4) back_betaH, (5) and left_theta — were selected as the EEG frequency domain features. The computed result for each sampling window corresponded to the EEG frequency domain features at the last sampling instance of that window.

3.3. *Eye movement and operation features extraction*

For the eye movement data, gaze position and pupil size have been demonstrated to hold significant value in understanding driving intention. Gaze position is a direct indicator of the driver's focus. Prior research has shown that gaze position is highly predictive of driver behavior and reflects the planning of future actions such as lane changes or obstacle avoidance. Pupil size, on the other hand, is a recognized proxy for cognitive load and mental effort. Changes in pupil diameter have been linked to fluctuations in a driver's psychological workload and alertness levels, which directly affect reaction potential and emergency handling capabilities. This makes pupil size a valuable feature for predicting imminent responses in hazardous driving situations [8, 35]. Concurrently, real-time vehicle speed along with steering, acceleration and braking intensities indirectly reflect a driver's anticipatory judgment and decision-making processes regarding surrounding traffic conditions [36, 37]. Therefore, this paper utilizes driver gaze position coordinates and pupil diameter as eye movement features, and vehicle three-dimensional speed, acceleration, braking and steering intensities as driving operation features.

3.4. *Multimodal fusion*

In this study, the feature-level multimodal fusion method was used to integrate the EEG, eye movement and operational data features obtained during the preprocessing stage, combining them into a new multimodal fusion feature matrix:

$$X_{\text{fusion}} = [X_{\text{EEG}} \vdots X_{\text{eye}} \vdots X_{\text{operational}}], \tag{10}$$

where X_{fusion} represents the multimodal fusion feature matrix, and X_{EEG}, X_{eye}, $X_{\text{operational}}$ represent the EEG feature matrix, the eye movement feature matrix and the operational feature matrix, respectively. Each type of feature was assigned equal weight in the fusion process.

For models using EEG raw voltage signals, the 32-channel voltage value at each sampling moment was directly taken as the initial EEG feature at that moment in this step.

Due to the varying sampling times of data collected from different devices and the significant discrepancies in their values, it was necessary to carry out timestamp alignment and data normalization. Finally, the dataset was labeled with three classes.

(i) **Timestamp alignment**
Timestamp alignment is a critical step in constructing a multimodal fusion dataset. Initially, timestamps from EEG, eye movement and driving operation data were unified into system epoch time. For eye movement data, the original timestamp T started from a random number T_{random} and incremented, while the system epoch time T_{system} at the start moment was recorded. Assuming their difference was

$$d = T_{\text{system}} - T_{\text{random}}, \tag{11}$$

the epoch timestamp T_{epoch} for eye movement data was

$$T_{\text{epoch}} = T + d. \tag{12}$$

After converting to system epoch time, the latest start time and earliest end time among the three signals were selected as the new start and end times. Subsequently, all data was uniformly resampled to the same timestamp using linear interpolation. There were two methods of interpolation. For EEG data processed with frequency domain feature extraction, all data were resampled to timestamps matching the driving operation data. For EEG data without frequency domain feature extraction, all data was resampled to timestamps matching the EEG voltage data.

(ii) **Data normalization**
The original data were scaled to the range [0, 1] using min-max normalization:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}, \tag{13}$$

where X_{norm} represents the normalized data, X represents the original data, X_{min} and X_{max} represent the minimum and maximum values in the original data, respectively.

(iii) **Label tagging**
Based on pedestrian presence flags from the data files and manual annotations of driver behaviors during EEG data collection, combined with driving operation data, the test data was tagged with labels. Driving intention labels in this paper were tagged into three classes: Normal driving, emergency braking intention and steering evasion (including steering with braking) intention, corresponding to classes 0, 1 and 2 in the classification model.

4. Intention Recognition Models

Traditional EEG classification algorithms, such as SVMs, LDA and *k*-nearest neighbors, have several limitations, particularly the inability to automatically extract features, requiring a high level of expertise from researchers [7, 30].

In recent years, deep learning models capable of automatically extract features from high-dimensional data have been proven effective in classifying physiological signals [38]. Common deep learning algorithms applied to EEG processing include multilayer perceptron (MLP) and CNN. These models could effectively manage the nonlinear and nonstationary characteristics of EEG with small sample sizes, while providing rapid training and inference speeds. Additionally, Transformer models, with powerful self-attention mechanisms, have shown outstanding performance across various classification tasks. This makes them particularly well-suited for integrating and analyzing multimodal data, especially those containing EEG [30, 39–42].

Based on the above, this paper constructed multiple driver intention recognition models using MLP, CNN and Transformer methods. The MLP and CNN models also serve as baselines for comparative analysis with the Transformer model.

4.1. *MLP-based model*

The MLP model is one of the most common and fundamental network architectures in deep learning [43–45]. This paper constructed two driver intention recognition models based on MLP, differing primarily in whether the EEG data had undergone frequency domain feature extraction in the input datasets. Both models shared a similar basic structure, consisting of a fully connected network with four hidden layers, regularized using dropout method. Assuming the input $x \in \mathbb{R}^m$, the output $a \in \mathbb{R}^n$ of a fully connected layer is calculated as follows:

$$a = \sigma(Wx + b), \tag{14}$$

where $W \in \mathbb{R}^{m \times n}$ is the weight matrix, b is the bias and σ is the nonlinear activation function. Assuming p is the proportion of neurons that are randomly deactivated, the output a' after the dropout regularization is as follows:

$$a' = \frac{a}{1 - p}. \tag{15}$$

For datasets without EEG frequency domain feature extraction, each sample window consisted of 40 sampling instances with a step size of five sampling instances. The original dataset was transformed into labeled sample windows comprising 32-channel EEG voltages, driver gaze positions, pupil sizes, vehicle three-dimensional speeds and acceleration, steering and braking intensities. Input data were flattened into one-dimensional vectors before being fed into the neural network.

For datasets preprocessed with EEG frequency domain feature extraction, each EEG frequency domain feature already encapsulated information from the past 40 sampling instances. These features were directly combined with contemporaneous eye movement and operation data features as inputs to the neural network. Eye movement and operation data features selected identically to the dataset without EEG frequency domain feature extraction.

Segmenting data using overlapping sampling windows serves as an effective data augmentation technique, enhancing data diversity and robustness. This method helps to mitigate the limitations posed by fewer participants. Additionally, it ensures that the system remains effective throughout the critical period between the onset of a hazard and the initiation of countermeasures, as it continuously assesses relevant data for accurate prediction and timely response.

To mitigate the impact of class imbalance on model performance, dataset balancing operations were performed prior to inputting samples into the model.

4.2. *CNN-based model*

CNN exhibits significant advantages in handling data with spatial and temporal locality, such as images and sequence data. Currently, CNN has been widely applied in EEG classification tasks [31, 46].

For a given input $X \in \mathbb{R}^{c\times m\times n}$, the computation of a feature map $Y \in \mathbb{R}^{d\times m'\times n'}$ in a convolutional layer is defined as follows:

$$Y_{d,i,j} = \sum_{p=-\delta_1}^{\delta_1} \sum_{q=-\delta_2}^{\delta_2} \sum_{r=1}^{c} K_{d,r,p,q} X_{r,i+p,j+q}, \tag{16}$$

where c and d represent the number of input and output channels, $K \in \mathbb{R}^{d\times c\times(2\delta_1+1)\times(2\delta_2+1)}$ denotes the convolutional kernel of the layer determined by its input and output. The shape of the convolutional kernel K is typically simplified as $(2\delta_1+1)\times(2\delta_2+1)$.

The CNN-based driver intention recognition model constructed in this paper comprised three convolutional layers and three pooling layers. The output from the final pooling layer underwent flattening and was subsequently processed through two fully connected layers for classification output. Dropout regularization is applied within the fully connected layers to mitigate overfitting. Similar to the approach used in MLP, the model input utilized datasets without EEG frequency domain feature extraction. Each 40 sampling instances served as a sampling window with a step size of five sampling points, transforming the original dataset into a series of labeled sample windows. Each sampling window was treated as an “image” input into the CNN, with sample “images” filtered before model input to ensure data balance.

To extract temporal features more effectively and capture channel-specific characteristics, a rectangular convolutional kernel was used in the first convolutional layer, emphasizing temporal dimensions over channel dimensions, whereas subsequent two layers employed conventional square kernels. Max-pooling was applied in the pooling layers for feature extraction.

4.3. *Transformer-based model*

The driver intention recognition model based on Transformer architecture designed in this paper was structured as depicted in Fig. 8, featuring layer normalization, multi-head attention mechanism and residual connections.

Similar to the CNN-based driver intention recognition model, the input to this model treats 40 sampling points as a “image”, with each sampling instance as a patch. These patches, along with the learnable embedding vector “Classification Token” (CLS), are fed into the Transformer Encoder.

The input data X first pass through a normalization layer, resulting in X_1. X_1 then goes through the multi-head attention module, the main functions of this module are as follows:

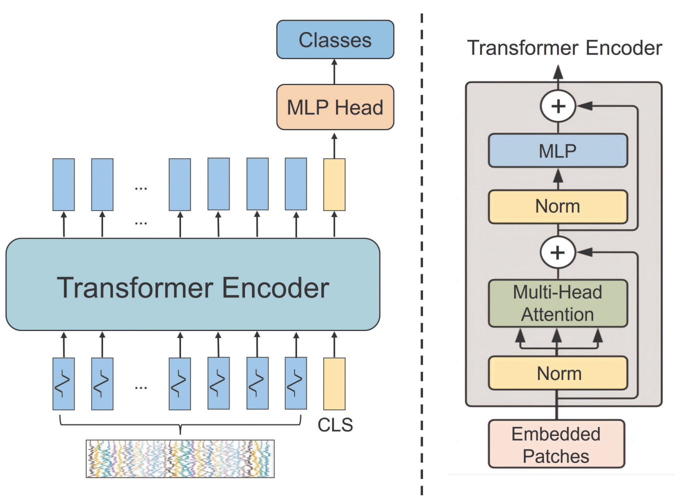


Fig. 8. Driver intention recognition model structure based on the Transformer.

Table 2. Parameter settings.

Parameters	Value
Window length of EEG frequency domain feature extraction	40
Window step size of EEG frequency domain feature extraction	1
Input dimension	42 ∗ 40 (EEG+ET)
(no EEG frequency domain features)	32 ∗ 40 (EEG) 10 ∗ 40 (ET)
Input dimension	15 ∗ 1 (EEG+ET)
(use EEG frequency domain features)	5 ∗ 1 (EEG)
Hidden layers size of both MLP	128, 256, 256, 64
Dropout rate of both MLP	0.5
Batch size of both MLP	1024
Convolution kernel shape of CNN	5 ∗ 3 (1st layer), 3 ∗ 3
Pooling kernel shape of CNN	2 ∗ 2
Batch size of CNN	32
Heads number of Transformer	2
Dropout rate of Transformer	0.2
Batch size of Transformer	8

First, X_1 is linearly transformed to obtain the Query (Q), Key (K) and Value (V) matrices:

$$Q = X_1 W^Q, K = X_1 W^K, V = X_1 W^V, \tag{17}$$

where W^Q, W^K and W^V are learnable weight matrices for the Query, Key, and Value projections, respectively. After that, the attention matrix is computed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{18}$$

where QK^T measures the similarity between the Query and Key matrices, d_k is the dimensionality of the Key matrix, the division by $\sqrt{d_k}$ is applied to prevent the Softmax function from saturating when working with high-dimensional matrices.

To capture information from different representational subspaces, the model employs a multi-head attention mechanism. In this approach, the Query, Key and Value are projected into different subspaces, and multi-head attention is computed in parallel. The final output is obtained by concatenating the results:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{19}$$

where h is the number of head and each attention head is computed as follows:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{20}$$

where W_i^Q, W_i^K, W_i^V are learned projection matrices for the Query, Key and Value for each head, and W^O represents the output projection matrix.

At the end of the model, the CLS token is extracted from the processed sequence within the Transformer Encoder, and subsequently processed through MLP Head, culminating in the classification output.

4.4. *Parameter settings*

Our model is built on the PyTorch framework and trained on an NVIDIA GeForce RTX 4060 GPU. The network parameters for the recognition models and feature extraction processes were determined through trial and error, guided by their functional principles. To evaluate the model's performance and avoid overfitting, we employed the Hold-Out cross-validation method during training, where the dataset was split into distinct training and validation sets. Table 2 presents a summary of the model parameters and their corresponding values.

5. Result Analysis

This paper analyzed the results mainly based on accuracy, precision, recall and F1 score of different driver intention recognition models with different input styles. For a multi-class problem with classes C_k ($k = 1, 2, \ldots, n$), where TP_k is the number of true positives for class C_k, FP_k is the number of false positives for class C_k, FN_k is the number of false negatives for class C_k, and TN_k is the number of true negatives for class C_k, the Precision_k, Recall_k and F1 Score_k of class C_k is calculated as follows:

$$\text{Precision}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k}, \tag{21}$$

$$\text{Recall}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}, \tag{22}$$

$$\text{F1 Score}_k = 2 \times \frac{\text{Precision}_k \times \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \tag{23}$$

After computing these metrics separately for each class, the macro-average method is used to calculate the average value for all classes as the overall metric for the entire test set, as follows:

$$\text{Precision} = \frac{1}{n}\sum_{k=1}^{n} \text{Precision}_k, \tag{24}$$

$$\text{Recall} = \frac{1}{n}\sum_{k=1}^{n} \text{Recall}_k, \tag{25}$$

$$\text{F1 Score} = \frac{1}{n}\sum_{k=1}^{n} \text{F1 Score}_k. \tag{26}$$

Accuracy is a function of the total number of classes and is calculated by dividing the number of correctly predicted samples by the total number of samples:

$$\text{Accuracy} = \frac{\sum_{k=1}^{n}(\text{TP}_k + \text{TN}_k)}{\sum_{k=1}^{n}(\text{TP}_k + \text{TN}_k + \text{FP}_k + \text{FN}_k)}. \tag{27}$$

Experimental results are summarized in Table 3, where EEG denotes the use of EEG data, ET denotes the use of eye movement and operation data, and “feature” indicates EEG data processed with frequency domain feature extraction. It can be observed that the Transformer model incorporating multimodal data fusion inputs consistently outperformed other models across all metrics, achieving the highest recognition effectiveness.

5.1. *The influence of EEG domain feature extraction*

Comparing the two MLP-based driver intention recognition models constructed in this paper enables us to assess the impact of EEG frequency domain feature extraction on the results. In contrast to solely inputting EEG voltage information into the models, using EEG data preprocessed with frequency domain feature extraction failed to improve the model performance and exhibited virtually no classification effect. This phenomenon may stem from limitations in our neuroscientific practices and physiological theory understanding, resulting in the selected frequency domain features providing insufficient accuracy contributions to the classification task, potentially exacerbated by the insufficient number of selected features. However, when eye movement and operational data were incorporated, the use of EEG data preprocessed with frequency domain features demonstrated superior recognition performance compared to fusion using only EEG voltage signals. This suggests that EEG frequency domain features encompass information that exhibits higher levels of complementarity with eye movement and operation data, thereby facilitating an overall enhancement of the classification performance in multimodal data.

5.2. *The influence of multimodal data fusion*

In this study, all driver intention recognition models, except the MLP model using raw EEG voltage data, demonstrated improved performance when applied to the multimodal dataset, which includes EEG, eye movement and operational data. Specifically, the transformer-based model, when leveraging multimodal fusion, exhibited substantial improvements in classification performance on the test set compared to models using only EEG data or eye movement and operational data. This finding underscores the critical role of multimodal data fusion in enhancing recognition accuracy and validating the effectiveness of combining multiple data modalities for this task. The results also strongly demonstrate that EEG and eye movement signals each offer distinct advantages in driver intention

Table 3. Summary of experimental results.

Model	Input	Accuracy	Precision	Recall	F1 Score
MLP	EEG + ET	50.88%	62.10%	53.93%	47.60%
	EEG (feature) + ET	73.99%	74.01%	74.33%	73.74%
	EEG	44.32%	29.46%	37.68%	32.60%
	EEG (feature)	41.91%	40.88%	38.22%	37.31%
	ET	64.85%	62.94%	63.66%	63.20%
CNN	EEG + ET	85.98%	86.76%	87.85%	86.41%
	EEG	71.78%	71.45%	73.98%	71.61%
	ET	78.46%	80.58%	80.38%	78.80%
Transformer	**EEG + ET**	**93.02%**	**92.61%**	**93.89%**	**93.04%**
	EEG	80.80%	80.94%	81.01%	80.84%
	ET	67.65%	68.20%	66.52%	67.07%

recognition and could complement one another to enhance the performance of the intention recognition model.

The lower performance of the MLP model using raw EEG voltage data, compared to the model using only eye movement and operational data, may be due to the limited capacity of the MLP to capture the complex spatiotemporal dynamics present in EEG signals. The MLP struggles to effectively model these intricate relationships, which impairs its ability to fully integrate EEG features with eye movement and operational data [47].

5.3. *The influence of model structure*

Further analysis of Table 3 reveals that the Transformer-based driver intention recognition model outperformed the other models in both scenarios: When using only EEG voltage data and when using multimodal datasets containing EEG, eye movement and operation data. This highlights the strength of the Transformer's multi-head attention mechanism in effectively capturing the complex spatio-temporal relationships present in the EEG data. The multi-head attention mechanism enables the Transformer to attend to multiple temporal patterns simultaneously, which is particularly advantageous when dealing with highly dynamic and nonstationary signals like EEG. These signals exhibit intricate temporal dependencies, making the Transformer well-suited for learning long-range correlations and subtle variations over time.

However, when trained and tested solely on eye movement and operation datasets, the CNN-based model achieved superior performance, surpassing even the Transformer model. This result suggests that for certain types of data, the simpler convolutional structure of CNN can be more effective. A CNN-based model excels at recognizing local patterns and features, making them particularly well-suited for processing eye movement and operational data, which often involve more localized patterns with limited long-term dependencies. The focus of the CNN-based model on local feature extraction is sufficient for this task and can be more efficient than the complex global attention mechanisms employed by the Transformer-based model.

Figure 9 shows the performance metrics of various model structures using multimodal inputs, including EEG, eye movement, and driving operational data, where "(feature)" indicating the use of EEG frequency domain features. A comparison between the MLP-based models, CNN-based model and Transformer-based model reveals that while manually extracting EEG frequency domain features improves the MLP model's performance to some extent, the performance gains achieved through advanced network structures are greater than those from manual feature extraction.

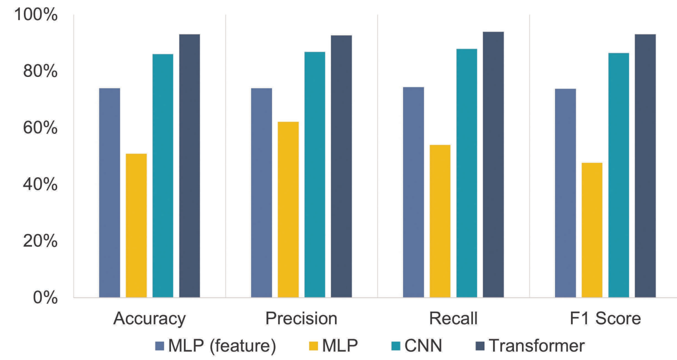


Fig. 9. The performance metrics of different model structures using multimodal inputs.

5.4. *Normalized confusion matrix analysis*

Figure 10 presents the normalized confusion matrices of the driver intention recognition models constructed in this paper for different inputs. In the matrix, each color block along with its numerical value represents the ratio of the predicted result class to the corresponding actual label class.

It can be observed that using manually extracted EEG frequency domain features significantly improves the recognition accuracy of the steering evasion class and normal driving class compared to using EEG voltage data when

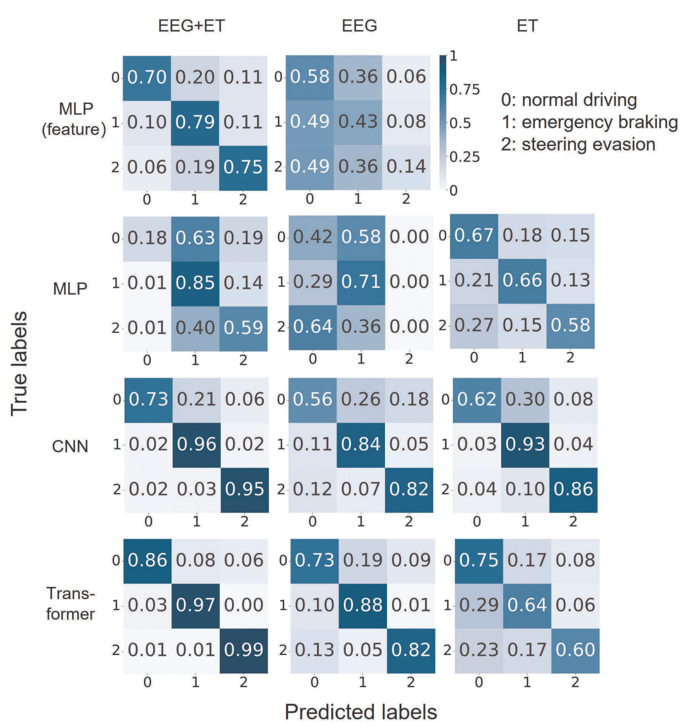


Fig. 10. Normalized confusion matrix of different recognition models with different inputs.

multimodal fusion data inputs. The CNN-based model, after incorporating multimodal fusion data inputs, demonstrates enhanced recognition performance for both emergency driving intentions, especially compared with using EEG data alone. Compared with solely using eye movement and operation data, the CNN-based model with multimodal inputs has mainly improved the recognition ability of the steering evasion class.

For multimodal datasets containing EEG, eye movement and operation information, the Transformer-based recognition model achieves high-recognition accuracy across all classes, showcasing the powerful ability of the multi-head attention mechanism to capture complex spatiotemporal relationship features from multimodal information.

5.5. *Application analysis and limitations*

Embedding the system designed in this study into autonomous driving systems provides a reliable supplementary information source, especially in urban environments where sensor sensitivity may decrease in certain conditions. In these cases, the system can utilize real-time physiological data from the driver to swiftly detect potential danger and take appropriate actions to prevent accidents. Integrating driver cognitive information with autonomous systems offers complementary benefits, potentially enhancing safety in complex and extreme driving scenarios. However, despite promising experimental results, several challenges and limitations must be addressed before real-world implementation.

The first challenge involves computational cost and real-time processing. Multimodal data fusion and deep learning models demand significant computational resources. Current autonomous driving systems rely on specialized hardware to achieve millisecond-level response times, and the added complexity of our system could place additional strain on these resources. Future research should explore more efficient algorithms and hardware acceleration techniques to reduce computational load and improve response times, ensuring the system meets real-time performance requirements.

Second, privacy and ethical concerns regarding the use of EEG data are critical issues. While EEG data enhances accuracy, collecting and processing such sensitive information introduces privacy risks, brain signals may contain highly personal information, raising concerns about data security and user consent. Future work should focus on designing privacy-preserving techniques that comply with regulations such as the general data protection regulation (GDPR). Methods like anonymization, encryption and secure computation can help balance the system's accuracy with the need to protect user privacy.

The third challenge concerns the integration of this system with existing vehicle perception technologies, such as cameras, radar and lidar. Future studies should investigate optimized multimodal data fusion strategies, exploring how this system can complement and enhance existing perception systems effectively. The focus could be on dynamic coordination between different data sources to improve decision-making accuracy in complex scenarios.

Besides, a more detailed study is needed regarding the potential risks associated with false positives, particularly in emergency braking situations. Unintended activations of emergency braking due to false positives could lead to rear-end collisions. Therefore, it is essential to conduct a thorough risk assessment and develop robust decision-making algorithms to mitigate the negative effects of false positives. Future research could explore techniques such as signal post-processing and error correction to improve the system's ability to manage false positives, ensuring safer and more reliable performance. A mutual verification strategy with existing autonomous driving systems can also be used as a means to reduce the risk of false positives.

Finally, the system's adaptability and robustness in real-world driving environments require further validation. While experimental results are promising, real-world scenarios are more complex and unpredictable, with varying traffic, weather conditions and driver states. In addition, the amount of data in this study is not particularly large. Although cross-validation and data enhancement methods are used, the limitation of the amount of data cannot be completely eliminated. Future studies should focus on expanding the dataset and testing the system's scalability, adaptability and long-term performance in diverse and dynamic real-world conditions, ensuring the system is capable of handling the variability and unpredictability of everyday driving.

6. Conclusion

In response to the current limitations of autonomous driving systems in urban environments and the future-oriented development needs for human-like autonomous driving systems, this paper proposes a driver intention recognition system based on driver cognitive information. The system aims to enhance the perception and decision-making capabilities of autonomous driving systems in urban environments, particularly in hazardous scenarios. Integrating multimodal data including driver EEG signals, eye movement information and driving operation data, this system accurately identifies driver intentions in dangerous situations. Compared to single-modal physiological signal-based intention recognition systems, the use of multimodal data fusion provides a more comprehensive understanding of driver cognitive information. In the future, embedding this system into the automatic driving system is expected to improve the automatic driving system's ability to quickly and accurately recognize dangerous scenarios and take timely measures.

Acknowledgments

The work was supported by the China Postdoctoral Science Foundation under Grant No. 2023M730252 and the National Natural Science Foundation Youth Foundation of China under Grant No. 62303058.

ORCID

Bo Zhang https://orcid.org/0009-0004-8916-3419
Xiaohui Hou https://orcid.org/0000-0002-0990-4757
Wei Wu https://orcid.org/0009-0004-2913-0308
Minggang Gan https://orcid.org/0000-0002-2163-2475

References

[1] Y. Li, J. Moreau and J. Ibanez-Guzman, Emergent visual sensors for autonomous vehicles, *IEEE Trans. Intell. Transp. Syst.* **24**(5) (2023) 4716–4737.
[2] Y. Fu, C. Li, F. R. Yu, T. H. Luan and Y. Zhang, A survey of driving safety with sensing, vehicular communications, and artificial intelligence-based collision avoidance, *IEEE Trans. Intell. Transp. Syst.* **23**(7) (2022) 6142–6163.
[3] T. Zhang *et al.*, Human-like decision-making of autonomous vehicles in dynamic traffic scenarios, *IEEE/CAA J. Autom. Sin.* **10**(10) (2023) 1905–1917.
[4] M. Malik *et al.*, An efficient driver behavioral pattern analysis based on fuzzy logical feature selection and classification in big data analysis, *J. Intell. Fuzzy Syst.* **43**(3) (2022) 3283–3292.
[5] M. Malik *et al.*, Deriving driver behavioral pattern analysis and performance using neural network approaches, *Intell. Autom. Soft Comput.* **32**(1) (2022) 87–99.
[6] R. S. Chhillar *et al.*, A dynamic and optimized routing approach for VANET communication in smart cities to secure intelligent transportation system via a chaotic multi-verse optimization algorithm, *Clust. Comput.* **27**(5) (2024) 7023–7048.
[7] J. Ju and H. Li, A survey of EEG-based driver state and behavior detection for intelligent vehicles, *IEEE Trans. Biom. Behav. Identity Sci.* **6**(3) (2024) 420–434.
[8] A. Rosner *et al.*, Eye movements in vehicle control, in *Eye Movement Research. Studies in Neuroscience, Psychology and Behavioral Economics*, eds. C. Klein and U. Ettinger (Springer, Switzerland, 2019), pp. 929–969.
[9] S. Haufe *et al.*, Electrophysiology-based detection of emergency braking intention in real-world driving, *J. Neural Eng.* **11**(5) (2014) 056011.
[10] S. Haufe *et al.*, EEG potentials predict upcoming emergency brakings during simulated driving, *J. Neural Eng.* **8**(5) (2011) 056001.
[11] S. M. Lee *et al.*, Detecting driver's braking intention using recurrent convolutional neural networks based EEG analysis, in *Proc. 4th IAPR Asian Conf. Pattern Recognit. (ACPR)* (IEEE, 2017), pp. 840–845.
[12] L. Z. Bi *et al.*, A novel method of emergency situation detection for a brain-controlled vehicle by combining EEG signals with surrounding information, *IEEE Trans. Neural Syst. Rehabil. Eng.* **26**(10) (2018) 1926–1934.
[13] L. G. Hernández *et al.*, EEG-based detection of braking intention under different car driving conditions, *Front. Neuroinform.* **12** (2018) 29.
[14] T. Teng, L. Z. Bi and Y. L. Liu, EEG-based detection of driver emergency braking intention for brain-controlled vehicles, *IEEE Trans. Intell. Transp. Syst.* **19**(6) (2018) 1766–1773.
[15] H. Hou *et al.*, Effects of driving style on driver behavior, *China J. Highway Transport* **31**(4) (2018) 18–27.
[16] Y. L. Sun and S. M. Feng, Effects of lane-change scenarios on lane-change decision and eye movement, *Ergonomics* **67**(1) (2024) 69–80.
[17] M. Wu *et al.*, Gaze-based intention anticipation over driving maneuvers in semi-autonomous vehicles, in *Proc. 2019 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)* (IEEE, 2019), pp. 6210–6216.
[18] F. Lethaus *et al.*, A comparison of selected simple supervised learning algorithms to predict driver intent based on gaze data, *Neurocomputing* **121** (2013) 108–130.
[19] K. Griesbach, K. H. Hoffmann and M. Beggiato, Prediction of lane change by echo state networks, *Transp. Res. Part C-Emerg. Technol.* **121** (2020) 102841.
[20] Q. Deng *et al.*, Prediction performance of lane changing behaviors: A study of combining environmental and eye-tracking data in a driving simulator, *IEEE Trans. Intell. Transp. Syst.* **21**(8) (2020) 3561–3570.
[21] Y. S. Guo *et al.*, Driver lane change intention recognition in the connected environment, *Physica A Stat. Mech. Appl.* **575** (2021) 126057.
[22] X. M. Tao *et al.*, A multimodal physiological dataset for driving behaviour analysis, *Sci. Data* **11**(1) (2024) 378.
[23] Y. X. Liu *et al.*, Fusion of spatial, temporal, and spectral EEG signatures improves multilevel cognitive load prediction, *IEEE Trans. Human-Machine Syst.* **53**(2) (2023) 357–366.
[24] Y. Liu *et al.*, Cognitive load prediction from multimodal physiological signals using multiview learning, *IEEE J. Biomed. Health Inform.* (2023).
[25] Z. Guo, M. Zhou and G. Li, Driver vigilance state estimation based on multisource data, *IEEE Sens. J.* **23**(3) (2023) 2592–2604.
[26] H. H. Yang *et al.*, Real-time driver cognitive workload recognition: Attention-enabled learning with multimodal information fusion, *IEEE Trans. Ind. Electron.* **71**(5) (2024) 4999–5009.
[27] J. Hu *et al.*, Driver steering intention prediction based on fusion of EEG and eye movement signals, in *Proc. 2023 China Autom. Congr. (CAC)* (IEEE, 2023), pp. 8306–8310.
[28] J. Y. Zhu *et al.*, Takeover quality prediction based on driver physiological state of different cognitive tasks in conditionally automated driving, *Adv. Eng. Inform.* **57** (2023) 102100.
[29] J. Huang, Y. Liu and X. Y. Peng, Recognition of driver's mental workload based on physiological signals, a comparative study, *Biomed. Signal Process. Control* **71** (2022) 103094.
[30] A. Chaddad *et al.*, Electroencephalography signal processing: A comprehensive review and analysis of methods and techniques, *Sensors* **23**(14) (2023) 6434.
[31] C. Z. Sun and C. Z. Mou, Survey on the research direction of EEG-based signal processing, *Front. Neurosci.* **17** (2023) 1203059.
[32] A. K. Singh and S. Krishnan, Trends in EEG signal feature extraction applications, *Front. Artif. Intell.* **5** (2022) 1072801.
[33] H. Díaz *et al.*, EEG beta band frequency domain evaluation for assessing stress and anxiety in resting, eyes closed, basal conditions, *Procedia Comput. Sci.* **162** (2019) 974–981.

[34] T. Aminosharieh Najafi *et al.*, Driver attention assessment using physiological measures from EEG, ECG, and EDA signals, *Sensors* **23**(4) (2023) 2039.
[35] H. Liu *et al.*, Lane-change intention recognition considering oncoming traffic: novel insights revealed by advances in deep learning, *Accid. Anal. Prev.* **198** (2024) 107476.
[36] A. Girma *et al.*, Deep learning with attention mechanism for predicting driver intention at intersection, in *Proc. 2020 IEEE Intell. Vehicles Symp. (IV)* (IEEE, 2020), pp. 1183–1188.
[37] X. Zhao *et al.*, Identification of driver's braking intention based on a hybrid model of GHMM and GGAP-RBFNN, *Neural Comput. Appl.* **31** (2019) 161–174.
[38] A. Shrestha and A. Mahmood, Review of deep learning algorithms and architectures, *IEEE Access* **7** (2019) 53040–53065.
[39] B. Jin, X. Xu and Y. Zhang, Thermal coal futures trading volume predictions through the neural network, *J. Model. Manag.* (2024).
[40] B. Jin and X. Xu, Machine learning predictions of regional steel price indices for east China, *Ironmaking Steelmaking* (2024) 03019233241254891.
[41] A. Dosovitskiy *et al.*, An image is worth 16x16 words: Transformers for image recognition at scale (2020), arXiv:2010.11929.
[42] A. Vaswani *et al.*, Attention is all you need, *Adv. Neural Inf. Process. Syst.* **30** (2017).
[43] B. Jin and X. Xu, Wholesale price forecasts of green grams using the neural network, *Asian J. Econ. Bank.* (2024).
[44] B. Jin and X. Xu, Predictions of steel price indices through machine learning for the regional northeast Chinese market, *Neural Comput. Appl.* **36** (2024) 1–20.
[45] B. Jin and X. Xu, Price forecasting through neural networks for crude oil, heating oil, and natural gas, *Meas. Energy* **1**(1) (2024) 100001.
[46] S. Rajwal and S. Aggarwal, Convolutional neural network-based EEG signal analysis: A systematic review, *Arch. Comput. Methods Eng.* **30**(6) (2023) 3585–3615.
[47] B. Jin and X. Xu, Carbon emission allowance price forecasting for China Guangdong carbon emission exchange via the neural network, *Global Finance Rev.* **6**(1) (2024) 3491.

Bo Zhang received his B.S. degree from Beijing Institute of Technology, Beijing, China, in 2024. His current research interests include intelligent driving and human–machine cooperation systems.

Xiaohui Hou received her B.S. and Ph.D. degrees in the School of Vehicle and Mobility from Tsinghua University in 2017 and 2022. She is currently postdoctoral researcher at the School of Automation, Beijing Institute of Technology. Her research interests include human-like decision, interactive motion planning, reinforcement learning and safety-critical control for autonomous driving.

Wei Wu received her B.S. degree in Automation from Beijing Institute of Technology, Beijing, China, in 2020. She is currently pursuing her Ph.D. degree in Beijing Institute of Technology, Beijing, China. Her current research interests include human–machine systems and sequential decision-making.

Minggang Gan received his B.E. and Ph.D. degrees in Control Science and Engineering from Beijing Institute of Technology, Beijing, China, in 2001 and 2007, respectively. During 2015–2016, he was a visiting scholar in New York University. He is currently a Professor with the National Key Laboratory of Autonomous Intelligent Unmanned Systems, School of Automation, Beijing Institute of Technology, Member of IEEE. His main research interest covers intelligent information processing and intelligent control.