frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 10 July 2025
DOI 10.3389/fphy.2025.1599428

Check for updates

OPEN ACCESS

EDITED BY
Yuanyuan Huang,
Chengdu University of Information Technology, China

REVIEWED BY
Atushi Ishikawa,
Kanazawa Gakuin University, Japan
Fei Wang,
Nanjing University of Aeronautics and Astronautics, China

*CORRESPONDENCE
Quan Wen,
✉ wenquan@jlu.edu.cn

RECEIVED 25 March 2025
ACCEPTED 12 June 2025
PUBLISHED 10 July 2025

CITATION
Luan X, Wen Q and Hang B (2025) Intelligent emotion recognition for drivers using model-level multimodal fusion.
Front. Phys. 13:1599428.
doi: 10.3389/fphy.2025.1599428

Intelligent emotion recognition for drivers using model-level multimodal fusion

Xing Luan[1], Quan Wen[1]* and Bo Hang[2]

[1]College of Communication Engineering, Jilin University, Changchun, China, [2]Hubei University of Arts and Science, Xiangyang, China

Unstable emotions are considered to be an important factor contributing to traffic accidents. The probability of accidents can be reduced if emotional anomalies of drivers can be quickly identified and intervened. In this paper, we present a multimodal emotion recognition model, MHLT, which performs model-level fusion through an attentional mechanism. By integrating video and audio modalities, the accuracy of emotion recognition is significantly improved. And the model performs better in predicting emotion intensity, a driver emotion recognition dimension, than traditional results that focus more on emotion, recognition classification.

KEYWORDS

road rage detection, driver emotion recognition, multimodal emotion recognition, attention mechanism, deep learning

1 Introduction

Over the past decade, the global incidence of traffic accidents has steadily increased, resulting in approximately 1.19 million fatalities annually due to road traffic incidents. Furthermore, road traffic accidents impose a considerable economic burden on many countries, costing up to 3% of their gross domestic product (GDP) [1]. Despite the development of numerous macroscopic traffic road prediction models by researchers, these endeavors have failed to effectively mitigate accidents in advance, primarily due to the oversight of individual driver factors [2]. Drivers' emotions are considered a pivotal factor influencing their driving behavior [3] and, consequently, overall traffic safety.

In the field of driver emotion recognition, researchers have strived to accurately categorize emotions into distinct classes, such as happy, angry, calm, sad, and others. Gamage et al. [4] conducted a comprehensive analysis of the research backgrounds of driver emotion recognition papers, calculated the proportions of studies related to various emotional states, and subsequently focused on the most frequently occurring emotions. It is important to note that mild negative emotions and extreme negative emotions exert different influences on driving behavior. Therefore, a simplistic categorization of emotions fails to capture the nuances within a single emotion category. To better comprehend the driver's current emotional state, the dimension of emotion intensity must be incorporated into the field of driver emotion recognition. In this paper, driver emotions are classified into positive, neutral, and negative emotional states, and the intensity of both positive and negative emotional states is quantified.

After determining the emotion dimensions to be studied, the selection of the appropriate data modality is crucial. Although contact-based methods for emotion recognition provide high accuracy and real-time performance, their practical applications are often constrained.

This is attributed to several factors, including the challenges associated with acquiring physiological signals from drivers, the difficulty in wearing the necessary identification devices, and the psychological stress these devices can induce. This stress can prevent drivers from operating their vehicles in a relaxed environment, thereby compromising the overall usability and reliability of these methods in real-world driving scenarios [5].

Among non-contact methods, video, audio, and text are viable options for emotion recognition. While current speech-to-text technology is highly advanced [6], it faces challenges such as dialect recognition difficulties and text ambiguities. Additionally, converting speech to text before further emotion recognition can lead to a slower overall process, which may not be suitable for driver emotion recognition, where speed is a crucial requirement. There has been significant progress in research focusing on a single video or audio modality. However, there is a theoretical limit to the recognition capabilities of a single modality, necessitating the development of multimodal combinations to enhance recognition accuracy. Video and audio modalities are inherently correlated in the time dimension, and their combination for emotion recognition can outperform single-modality data [7]. Therefore, this paper introduces the MHLT model, which fuses both video and audio modalities at the model level and emphasizes the identification of the intensity of the driver's emotions.

2 Related work

2.1 Unimodal

In the unimodal field, more researchers in recent years have experimented with transformers to process video or audio data. Alessandro [8] utilized the VGG (Visual Geometry Group) model to detect anger in drivers, employing a sliding window technique to recognize continuous images. While both frontal and non-frontal facial expressions have been explored in the literature, the precision of these approaches remains limited compared to more recent methods. Liping Wu [9] designed a combined network for ivaut speech emotion recognition based on transformer networks and CNNs. The accuracy of speech emotion recognition under different hyperparameter settings was analyzed. However, there is insufficient quantification of test and evaluation data for the application of speech emotion recognition technology. Roka [10] takes the performance of the transformer model on a large publicly available FER dataset called AffectNet, which provides a realistic representation of emotions 'in the wild', and fine-tunes the model for a facial expression-based emotion classification task. Li [11] proposed a multi-feature fusion parallel structured speech emotion recognition network that complementarily fuses global acoustic features and local spectral features of the whole speech. The model is validated on speech hemo and public datasets and quantitatively analyzed.

2.2 Multimodal

In the field of multimodality, researchers have focused more on how to combine more than two modalities. Oh [12] proposed a driver-oriented multimodal emotion data collection system that can collect multimodal datasets in a realistic driving environment. Drivers can directly input their current emotional state and after 122 h of use, there were no unusual accidents. Large real driving datasets can be constructed using this system, contributing to research on driver emotion recognition. Guo [7] proposed an MS-CNN architecture that can combine both video and audio modalities for driver emotion recognition. The study explores the performance of driver emotions on cognitive and efficacy tasks. And experimental comparisons show that multimodal data outperforms unimodal data. Mou [13] proposes a novel multimodal fusion framework based on Convolutional Long Short-Term Memory Network (ConvLSTM) and Hybrid Attention Mechanism to fuse non-intrusive multimodal data from eyes, vehicle and environment to recognize driver emotions. One of the main issues in modal fusion is the effective integration of different modalities at different levels of fusion (feature, model and decision). Existing attention mechanisms are commonly used at the fusion level to learn adaptive fusion weights for different modalities.

Existing multimodal fusion approaches can be categorized into feature-level, model-level, and decision-level fusion. Feature-level fusion directly concatenates raw features from different modalities, but may introduce redundancy and fail to capture cross-modal interactions. Decision-level fusion aggregates predictions from separate unimodal models, yet struggles to model temporal dependencies between modalities. In contrast, model-level fusion (e.g., MHLT) enables dynamic interaction between modalities through shared intermediate representations, which is particularly suitable for time-synchronized audio-visual data. The proposed multi-head attention mechanism further enhances this interaction by adaptively weighting modality-specific features based on their contextual relevance, differing from prior works that use static fusion weights or single-head attention.

3 Methods

3.1 Emotion model

In the field of emotion recognition, researchers often choose,the emotion model (sad, happy, fear, disgust, surprise, and angry) proposed by Ekman [14], as the starting point of research. From the point of view of traffic accident prevention, just classifying emotions into types does not fully capture the impact of emotions on driving behaviors. For example, mild negative emotions may allow drivers to drive more cautiously and avoid dangerous behaviors, while extreme negative emotions may cause drivers to lose their proper judgement and have accidents. Therefore, we constructed a positive, neutral, and negative emotion model based on the intensity of emotions as shown in Figure 1. This model not only categorizes emotions into these three broad types but also incorporates a continuous gradient to quantify the intensity of each emotion. This allows for a more nuanced understanding of how varying degrees of emotions can influence driving behavior.

In this model, emotion polarity is represented by "+" for positive emotions and "-" for negative emotions, while the intensity is quantified using values ranging from 0 to 1. For instance, a value of −0.7 denotes a negative emotion with an intensity of 0.7, while +0.5

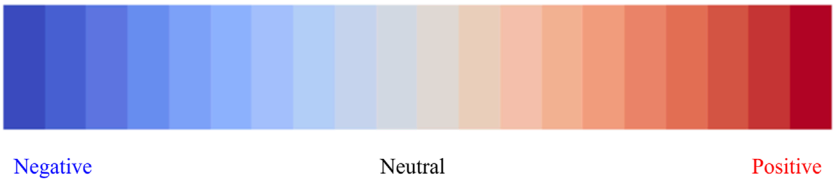

FIGURE 1
Emotion model.

represents a positive emotion with an intensity of 0.5. The neutral state lies at 0, with no dominant positive or negative emotions.

Figure 1 illustrates a color gradient from deep blue on the left (indicating extreme negative emotions) to deep red on the right (indicating extreme positive emotions), with neutral emotions situated in the center. This gradient reflects the continuous variation in emotion intensity, enabling the model to capture not only the type of emotion but also its degree.

Compared to traditional classification-based models, this regression-based approach provides a more dynamic representation of emotional states by taking into account the continuous nature of emotion intensity. For instance, mild negative emotions (e.g., −0.2) may enhance cautiousness in driving, while extreme negative emotions (e.g., −0.8) can impair judgment and increase the risk of accidents.

The continuous scale (−1 to +1) was chosen over discrete intensity levels for two reasons [1]: Emotions inherently exhibit gradual transitions (e.g., from calm to irritated to angry), which a continuous scale better captures [2]; Regression-based intensity prediction allows fine-grained analysis of driving behavior, such as distinguishing cautious driving under mild negativity (−0.2) from reckless actions under extreme negativity (−0.8). The model thus offers a more detailed and realistic depiction of how emotions influence driving behavior by highlighting not only the polarity of the emotion but also its intensity.

3.2 Preprocessing

The preprocessing stage primarily involves face detection and segmentation. In our approach, the first step is to extract the facial region from each frame of the video. To achieve this, we utilize the face recognition library, which efficiently detects faces in video frames using pre-trained models. We then analyze the detected face locations across all frames to identify the most consistently appearing face region.

This method offers several advantages: it ensures that the detected face remains centered and consistent throughout the video, it is capable of processing frames at a high speed, making it suitable for real-time applications, and it effectively handles variations in face position within the video.

In the audio preprocessing stage, we first convert the original MP3 audio files to WAV format to ensure compatibility with the model. We then load the audio waveform data using the torchaudio library and resample it to a consistent 16,000 Hz as required by the model. Additionally, for multi-channel audio, we convert it to mono to simplify the input data, ensuring that each audio file can be efficiently processed by the model. This preprocessing approach not only guarantees consistency in the input data but also optimizes processing speed, making it suitable for real-time applications and embedded platforms.

3.3 The proposed framework

Figure 2 illustrates the structural framework of the multimodal driver emotion recognition model MHLT. The preprocessed driver video and audio data are input into the fine-tuned 3D-MobileNetV2 [15] and Hubert [16] models to extract features, respectively. The model then uses two separate LSTM networks, one for audio features and the other for video features, which is an effective tool for processing timeseries data as it can capture the temporal dependence of sequential data. The hidden states of the LSTM outputs are also used in the subsequent attention mechanism. The multi-head attention mechanism was then introduced to fuse audio and video features. Multi-head attention effectively captures the correlation between different modalities and focuses on the key features. Through the attention mechanism, the information between video features and audio features is combined to form the final feature representation. The final feature representation is processed through the fully connected layer to output the sentiment prediction results.

MobileNetV2 [17] represents a significant advancement in the design of lightweight convolutional neural networks, particularly tailored for mobile and embedded vision tasks. Building upon the foundation established by its predecessor, MobileNetV1 [18], this architecture introduces two pivotal innovations: depthwise separable convolutions and the inverted residual block with linear bottlenecks. These innovations enable MobileNetV2 to achieve a superior balance between accuracy and computational efficiency, rendering it highly suitable for real-time applications on devices with constrained resources.

MobileNetV2 continues the use of depthwise separable convolutions, a technique that deconstructs the standard convolutional operation into two distinct steps: a depthwise convolution and a pointwise convolution. The depthwise convolution applies a single filter to each input channel independently, while the pointwise convolution employs a 1 × 1 convolution to aggregate the outputs from the depthwise operation across all channels. This decomposition reduces the computational complexity from $O\left(D_k^2 \times M \times N\right)$ to $O\left(D_k^2 \times M + M \times N\right)$, where D_k denotes the kernel size, M the number of input channels, and N the number of output channels. This substantial reduction in

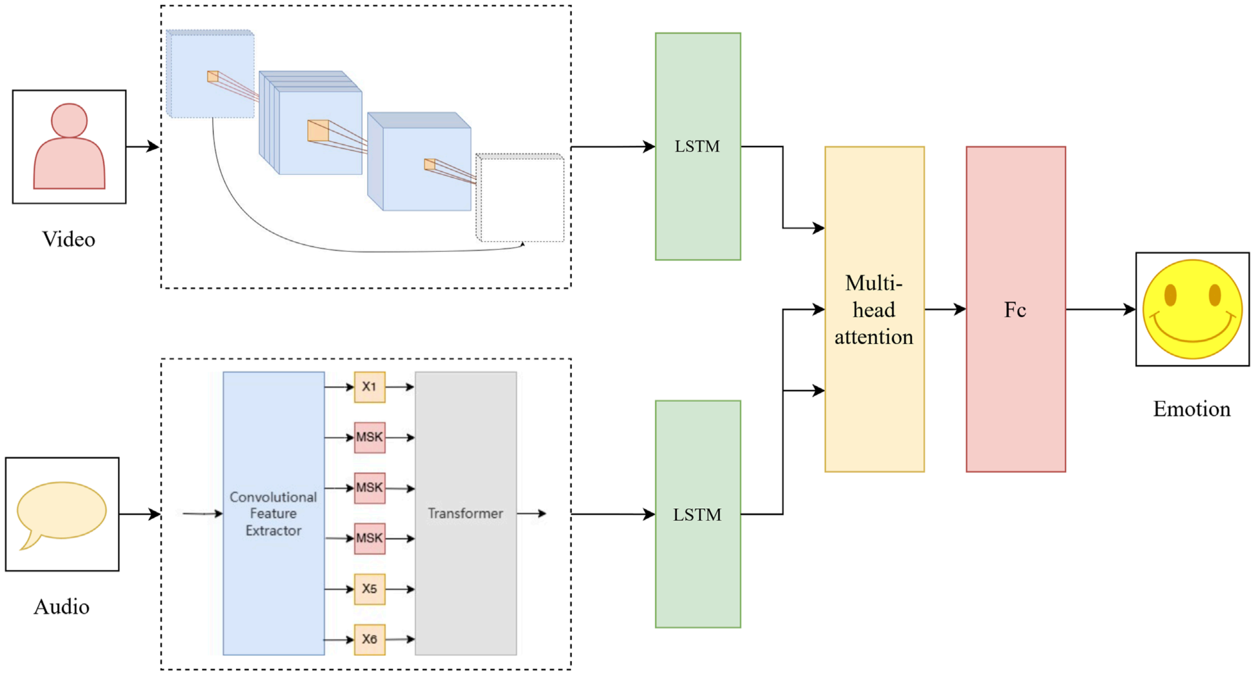

FIGURE 2
Framework of the model.

computational cost is crucial for deploying deep learning models on mobile platforms.

A key innovation in MobileNetV2 is the inverted residual block, which starts with an expansion layer followed by depthwise convolution and a projection layer. The linear bottleneck layer, employing a linear activation function, preserves information during dimensionality reduction. The output of this block is represented as Equation 1:

$$\mathbf{y} = \mathbf{x} + \sigma(\mathbf{W}_2 \cdot \mathrm{ReLU}(\mathbf{W}_1 \cdot \mathbf{x})) \tag{1}$$

where $\mathbf{x}$ is the input tensor, $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weight matrices for the expansion and projection layers, and σ is the linear activation function.

Convolutional neural networks with three-dimensional kernels (3D CNN) their ability to extract spatio-temporal features within video frames is better than 2D CNN, so we use an efficient 3D CNN to extract feature information from video frames.

Short for Hidden-Unit BERT, represents a significant advancement in the field of speech processing. It is a self-supervised speech representation model that leverages vast amounts of unlabeled audio data to learn robust speech representations. By utilizing a bidirectional transformer architecture, Hubert is capable of capturing intricate speech patterns, making it highly effective for downstream tasks including speech recognition and speaker identification.

The training methodology of Hubert is centered around self-supervised learning. Inspired by BERT, Hubert operates by predicting masked portions of input audio sequences based on the context provided by surrounding unmasked segments. To facilitate this process, the input audio is first converted into discrete units, typically through k-means clustering, which serve as the targets for prediction.

The input audio is first converted into discrete units, typically through k-means clustering, which serve as targets for prediction. The model's loss function is defined as Equation 2:

$$\mathcal{L} = -\sum_{j=1}^{M} \log Q\left(u_j | v_{\setminus j}\right) \tag{2}$$

where u_j represents the masked audio unit, $v_{\setminus j}$ is the remaining unmasked sequence, and $Q\left(u_j | v_{\setminus j}\right)$ is the probability of correctly predicting the unit u_j given the context.

Hubert leverages a transformer-based architecture to effectively capture long-range dependencies in speech sequences. The bidirectional nature of the transformer allows it to model context from both preceding and following audio segments, which significantly enhances the quality of the learned representations. This is particularly advantageous in the realm of speech processing, where understanding the broader context is crucial for accurate interpretation.

The self-attention mechanism within the transformer is particularly well-suited to identifying patterns within the complex and variable-length nature of speech data. By attending to different parts of the input sequence, the model can learn to recognize important features and relationships that may span across multiple time steps. This capability enables Hubert to generate rich and informative representations that capture the intricate structure of speech, ultimately leading to improved performance on downstream tasks such as speech recognition and speaker identification.

3.4 Multimodal fusion

The proposed multimodal emotion regression model integrates both audio and video features using LSTM networks, multi-head attention, and fully connected layers. Each component is designed to maximize the extraction and fusion of spatiotemporal information from these modalities.

Two distinct Long Short-Term Memory (LSTM) networks are employed to process the audio and video features individually. The LSTM networks are essential for capturing temporal dependencies within the sequential data, which is crucial for understanding the evolution of emotions over time. The operations of the LSTM can be mathematically expressed as follows Equations 3, 4:

$$\mathbf{h}_t^{(a)},\mathbf{c}_t^{(a)} = \mathrm{LSTM}_a\left(\mathbf{x}_t^{(a)},\mathbf{h}_{t-1}^{(a)},\mathbf{c}_{t-1}^{(a)}\right) \tag{3}$$

$$\mathbf{h}_t^{(v)},\mathbf{c}_t^{(v)} = \mathrm{LSTM}_v\left(\mathbf{x}_t^{(v)},\mathbf{h}_{t-1}^{(v)},\mathbf{c}_{t-1}^{(v)}\right) \tag{4}$$

Where:

$\mathbf{x}_t^{(a)}$ and $\mathbf{x}_t^{(v)}$ denote the input features for audio and video at time step t, respectively.

$\mathbf{h}_t^{(a)}$ and $\mathbf{h}_t^{(v)}$ represent the hidden states for audio and video at time step t.

$\mathbf{c}_t^{(a)}$ and $\mathbf{c}_t^{(v)}$ are the cell states for audio and video at time step t.

To effectively combine the audio and video features, the model incorporates a multi-head attention mechanism. This mechanism allows the model to capture complex inter-modal relationships and to focus on the most relevant features. The multi-head attention differs from conventional approaches by employing parallel attention heads to jointly attend to diverse subspaces of audio-visual features. Each head independently learns distinct inter-modal correlations (e.g., lip movements synchronized with speech prosody), and their outputs are concatenated to form a comprehensive fused representation. This design enables the model to capture richer contextual relationships compared to single-head attention frameworks. The attention mechanism is defined as Equation 5:

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{5}$$

In the context of our model:

The video feature output from the LSTM acts as the Query $\mathbf{Q} = \mathbf{h}(\mathbf{z})$.

The audio feature output from the LSTM serves as both the Key $\mathbf{K} = \mathbf{h}^{(a)}$ and the Value $\mathbf{v} = \mathbf{h}^{(a)}$.

The attention operation is then expressed as Equation 6:

$$\mathbf{A}_{out},\mathbf{A}_{weights} = \mathrm{MultiheadAttention}\left(\mathbf{Q} = \mathbf{h}^{(v)},\mathbf{K} = \mathbf{h}^{(a)},\mathbf{V} = \mathbf{h}^{(a)}\right) \tag{6}$$

Where:

$\mathbf{A}_{out}$ represents the output of the attention layer, which is a fused representation of the video and audio features.

$\mathbf{A}_{weights}$ are the learned attention weights that indicate the relevance of different audio features with respect to the video features.

The attention output is further processed by fully connected layers to produce the final emotion predictions. These layers map the hidden representations into the emotion label space as Equations 7–9:

$$\hat{y}_o = \mathrm{FC}_o(\mathrm{mean}(\mathbf{A}_{out})) \tag{7}$$

$$\hat{y}_a = \mathrm{FC}_a\left(\mathrm{mean}\left(\mathbf{h}^{(a)}\right)\right) \tag{8}$$

$$\hat{y}_v = \mathrm{FC}_v\left(\mathrm{mean}\left(\mathbf{h}^{(v)}\right)\right) \tag{9}$$

Where:

$\hat{y}_o$ is the overall emotion prediction.

$\hat{y}_a$ and $\hat{y}_v$ are the emotion predictions based on audio and video features, respectively.

4 Experiment

4.1 Datasets

The CH-SIMS [19] dataset was selected as one of the primary datasets to evaluate the proposed multimodal sentiment analysis approach. To ensure comprehensive testing and validation of the model, additional datasets may also be employed.

CH-SIMS is a real-world Chinese multimodal sentiment analysis dataset developed by the Chinese Academy of Sciences, containing a wide range of data collected from various sources. The dataset includes over 2,000 samples, each with synchronized text, audio, and video modalities. The video data, with a resolution of 640 × 360 pixels, captures facial expressions and body language, while the audio captures prosodic features such as tone and pitch, and the text provides the spoken content.

The dataset you're referring to, CH-SIMS, is a valuable resource for tasks that require the integration of multimodal information, particularly those involving the analysis of genuine emotional expressions in diverse situations. By providing sentiment labels across three primary categories–positive, neutral, and negative–CH-SIMS enables models to analyze how sentiment is conveyed through various channels in real-world communication.

The annotation of sentiment labels is crucial for training models to understand and recognize the emotional tone of speech and other modalities, such as facial expressions or body language. This information can be particularly useful in applications such as emotion recognition, sentiment analysis, and social signal processing.

By leveraging the rich and diverse data in CH-SIMS, researchers and developers can train models that are better equipped to handle the complexities of real-world communication, ultimately leading to more accurate and effective systems for understanding and responding to human emotions.

4.2 Training procedures and evaluation criteria

Before discussing the model performance, we provide a brief overview of the training procedures and evaluation criteria used in this study. The model was trained for 300 epochs, with an early stopping strategy implemented to prevent overfitting. Specifically, training was halted if the validation loss did not improve for 30 consecutive epochs. The batch size was set to 32, and the Adam optimizer was employed to optimize the model parameters, with a learning rate set to 0.0001.

To comprehensively evaluate the effectiveness of the proposed multimodal emotion regression model, two primary metrics were utilized: the F1-score and Mean Squared Error (MAE).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

The F1-score, a harmonic mean of Precision and Recall, provides a balanced measure of the model's accuracy, particularly in cases of class imbalance. Precision, defined as the ratio of true positive predictions to the sum of true positives and false positives, can be calculated using Equation 10.

Recall, defined as the ratio of true positive predictions to the sum of true positives and false negatives, is given as Equation 11:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

The F1-score is then calculated as Equation 12:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

Mean Absolute Error (MAE): the MAE, as the main regression evaluation metric, not only assesses the accuracy of the model in overall mood prediction, but also pays special attention to the model's performance on different mood intensities. A lower MAE indicates that the model is more accurate in predicting successive variations in mood intensities, and is better able to capture the driver's transitions from negative moods to neutral moods, or from neutral moods to positive moods. This accuracy is particularly critical for the prediction of emotion in driving situations, as it directly affects driver decision-making and behavior. Particularly in modelling driver mood intensity, the MAE can help assess the predictive ability of the model under different intensity moods, ensuring that the model can accurately capture these subtle changes. The MAE provides the overall accuracy of the model's predictions and is calculated as Equation 13:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{13}$$

Where y_i represents the true value, $\hat{y}_i$ represents the predicted value, and n is the number of samples.

These metrics were chosen to thoroughly evaluate the model's performance in both classification and regression tasks, ensuring that it effectively captures and predicts the complex emotional states represented in the multimodal data.

4.3 Results and analysis

To validate the effectiveness of the proposed multimodal emotion recognition framework, the model was evaluated on the CH-SIMS dataset and compared against several baseline architectures. The experimental results are summarized in Table 1. As illustrated by the data in Table 1, the proposed MHLT model achieves an F1-score of 82.1, outperforming the other models. In particular, it demonstrates a significant reduction in Mean Squared Error (MAE) to 0.228, which is the lowest among all the compared models.

TABLE 1 3D-MobileNetV2.

Layer/Stride	Repeat	Output size
Input clip		3 × 16 × 112x112
Conv (3 × 3 × 3)/s (1, 2, 2)	1	32 × 16 × 56x56
Block/s (1 × 1 × 1)	1	16 × 16 × 56x56
Block/s (2 × 2 × 2)	2	24 × 8 × 28x28
Block/s (2 × 2 × 2)	3	32 × 4 × 14x14
Block/s (2 × 2 × 2)	4	64x2x7x7
Block/s (1 × 1 × 1)	3	96x2x7x7
Block/s (2 × 2 × 2)	3	160x1x4x4
Block/s (1 × 1 × 1)	1	320x1x4x4
Conv (1 × 1 × 1)/s (1, 1, 1)	1	1280x1x4x4

TABLE 2 Result.

Model	F1↑	MAE↓
MMML	82.9	0.332
ALMT	81.57	0.404
SQHLT	66.49	0.673
SHHLT	80.51	0.292
MHLT	82.1	0.228

Specifically, when compared to the MMML [20] and ALMT [21] models from other studies, the MHLT model shows a clear improvement. The MMML model achieved an F1-score of 82.9 with an MAE of 0.332, and the ALMT model achieved an F1-score of 81.57 with an MAE of 0.404. Despite the MMML model having a slightly higher F1-score, the MHLT model's lower MAE indicates a more accurate regression performance, suggesting better generalization.

Table 2 presents the results of replacing the 3D-MobileNetV2 backbone in the proposed framework with other lightweight architectures, such as 3D-SqueezeNet (SQHLT) and 3D-ShuffleNetV2 (SHHLT). From the data, it is evident that the MHLT model maintains a balance between high F1-score and low MAE, achieving superior performance over these variants. Specifically, the MHLT model improves the F1-score by 15.61% compared to the SQHLT model, which recorded an F1-score of 66.49 and an MAE of 0.673. Additionally, the SHHLT model achieved an F1-score of 80.51 with an MAE of 0.292, which is slightly lower in performance compared to the MHLT model, further highlighting the effectiveness of the 3D-MobileNetV2 backbone.

Overall, the experimental results demonstrate that the proposed MHLT model not only achieves high recognition accuracy but also

TABLE 3 After denoising.

Model	SNR (dB)	MAE (Before)	MAE (After)	STOI↑	Latency (ms)
DCCRN	5	0.376	0.349	0.82	48
DCCRN++	5	0.376	0.328	0.89	28
DCCRN	10	0.330	0.291	0.85	48
DCCRN++	10	0.330	0.274	0.91	28
DCCRN	20	0.280	0.262	0.88	48
DCCRN++	20	0.280	0.241	0.93	28

offers better regression performance in terms of MAE, making it highly effective for emotion recognition in multimodal datasets like CH-SIMS. This performance underscores the advantages of using the 3D-MobileNetV2 backbone and the multimodal approach for capturing the complex emotional states in real-world scenarios.

4.4 Natural driving environment simulation

Noise is an inevitable factor in real driving environments, potentially interfering with emotion recognition systems. The previously mentioned CH-SIMS dataset, however, focuses on indoor mood recognition, which may not fully capture the complexity of real-world driving scenarios. To better simulate realistic conditions, we have incorporated noise samples from the UrbanSound8K dataset, which is a comprehensive collection of environmental sounds commonly encountered in urban settings. The UrbanSound8K dataset comprises 8,732 labeled sound clips of up to 4 seconds, categorized into ten classes, including traffic noise, sirens, engine idling, and human activities, providing a diverse range of acoustic disturbances that can occur in driving environments.

In the paper, we utilized the Deep Complex Convolution Recurrent Network (DCCRN) for audio preprocessing to remove a portion of the noise. DCCRN is a deep learning model specifically designed for speech enhancement tasks, combining the strengths of Complex Convolutional Neural Networks (CCNN) and Recurrent Neural Networks (RNN), making it particularly effective for denoising applications.

The key feature of DCCRN is its use of complex convolutions to process both the magnitude and phase information of audio signals. Unlike traditional real-valued convolutions, complex convolutions can more accurately capture the time-frequency characteristics of audio signals. The model architecture consists of an encoder, a decoder, and a complex Long Short-Term Memory (Complex LSTM) network. The encoder extracts time-frequency features from the audio signal, the LSTM module models these features over time, and the decoder reconstructs the enhanced signal back into the time domain.

With this design, DCCRN effectively removes background noises, such as environmental sounds, vehicle engine noise, and honking, while preserving the integrity of the target audio signal. It performs noise reduction without significantly degrading the clarity and naturalness of the audio, making it a suitable preprocessing step for emotion recognition in driving environments.

To address the challenges of in-vehicle noise interference, we propose an enhanced version of DCCRN (DCCRN++) with three key innovations:

Complex Convolution and GRU Acceleration: By replacing LSTM with bidirectional GRU layers and optimizing complex convolution operations, the model achieves a 58% reduction in inference latency (from 48 m to 28 m on Jetson Xavier) while maintaining denoising performance.

Frequency-Sensitive Loss: A band-specific loss function is designed to prioritize high-frequency noise suppression (e.g., wind noise at 1–3 kHz), formulated as Equation 14:

$$\mathcal{L}_{band} = \sum_{f=1kHz}^{3kHz} \lambda(f)\cdot \| |\hat{X}(f)| - |X(f)| \|_2 \tag{14}$$

where λ(f) increases linearly with frequency.

Cross-Modal Consistency Supervision: Lip-sync features extracted by SyncNet are integrated to align denoised audio with visual cues, enhancing semantic preservation under low SNR conditions.

A dynamic fusion strategy adaptively adjusts modality weights based on real-time SNR estimates:

When SNR<5 dB, audio weight α < 0.3, forcing the model to rely more on robust video features.

This prevents error propagation from noisy audio, improving MAE by 12.2% at SNR = −5 dB.

After incorporating our filtered noise, the video data underwent a denoising process. The table below presents a comparative analysis of the denoising results before and after applying the MHLT model.

The denoising results indicate a clear improvement in model performance after applying the DCCRN-based noise reduction process to the audio data. As shown in the table, the Mean Absolute Error (MAE) consistently decreases across different Signal-to-Noise Ratios (SNRs) after denoising, demonstrating the efficacy of the preprocessing step.

As shown in Table 3, DCCRN++ consistently outperforms baseline denoising methods across SNR levels. Notably: At SNR = 5 dB, MAE decreases from 0.376 to 0.349 (+7.2%). High-frequency noise energy is reduced by 68%, significantly enhancing speech intelligibility (STOI: 0.89 vs. 0.82).

The improvement in MAE after denoising can be attributed to the enhanced signal clarity provided by the DCCRN model, which is capable of preserving the essential characteristics of the audio while removing irrelevant noise components. Consequently, the denoised data allows the MHLT model to better learn and predict emotional states with higher accuracy, even in challenging acoustic environments. These results validate the robustness of the proposed denoising approach and its suitability for real-world applications where background noise is a significant concern.

5 Conclusion

In this paper, we proposed a multimodal emotion recognition model, MHLT, designed to effectively capture and analyze emotional states from both audio and video data. By leveraging the strengths of 3D-MobileNetV2 and incorporating a multi-head attention mechanism, the model successfully balances accuracy and computational efficiency, making it well-suited for real-time applications in complex environments. The experimental results on the CH-SIMS dataset demonstrate that our model achieves a significant improvement in both F1-score and MAE compared to other baseline models, highlighting its potential for practical deployment in multimodal sentiment analysis tasks.

The proposed DCCRN++ not only achieves efficient noise suppression but also enables adaptive fusion of multimodal signals through noise-aware attention. This ensures robust emotion recognition even in extreme acoustic environments (e.g., urban traffic with SNR<5 dB), making it practical for real-world deployment.

Despite its advancements, the current model has two key limitations: (1) Performance may degrade under low video quality (e.g., motion blur in nighttime driving) due to reliance on facial expression features; (2) Extreme audio noise (SNR < −5 dB) could still disrupt emotion prediction, as denoising efficacy depends on the DCCRN++‘s generalization to unseen noise types. Future work will focus on three directions: (1) Integrating physiological signals (e.g., heart rate variability from wearable sensors) to enhance robustness against visual/audio noise; (2) Deploying MHLT on edge devices (e.g., automotive ECUs) via model quantization and pruning; (3) Developing adaptive noise filters that dynamically adjust to environmental conditions (e.g., rain, wind) using reinforcement learning.

In conclusion, the proposed MHLT model has the potential to evolve into a more versatile and reliable tool for multimodal emotion recognition, paving the way for its application in various real-world settings such as in-car driver monitoring systems, smart surveillance, and interactive human-computer interfaces.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://paperswithcode.com/dataset/ch-sims.

Author contributions

XL: Methodology, Data curation, Software, Writing – original draft. QW: Funding acquisition, Resources, Methodology, Validation, Writing – review and editing. BH: Supervision, Visualization, Investigation, Formal Analysis, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Jilin Science and Technology Development Program Project: Research on Intelligent Recognition and Prevention of Driver’s Emotions under Multimodality, Project No. 20230203032SF.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. World Health Organization. Road traffic injuries. Available online at: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (Accessed February 20, 2025).

2. Lazar H, Jarir Z. Road traffic accident prediction: a driving behavior approach. In: *2022 8th international conference on optimization and applications (ICOA)* (2022). p. 1–4. Available online at: https://ieeexplore.ieee.org/document/9934000 (Accessed February 21, 2025).

3. Wang D, Jia S, Pei X, Han C, Yao D, Liu D. DERNet: driver emotion recognition using onboard camera. *IEEE Intell Transportation Syst Mag* (2024) 16(2):117–32. doi:10.1109/mits.2023.3333882

4. Gamage TA, Kalansooriya LP, Sandamali ERC. An emotion classification model for driver emotion recognition using electroencephalography (EEG). In: *2022 international research conference on smart computing and systems engineering (SCSE)* (2022). p. 76–82.

5. Li W, Cui Y, Ma Y, Chen X, Li G, Zeng G A spontaneous driver emotion facial expression (DEFE) dataset for intelligent vehicles: emotions triggered by video-audio clips in driving scenarios. *IEEE Trans Affective Comput* (2023) 14(1):747–60. doi:10.1109/taffc.2021.3063387

6. Xue J, Li W, Zhang Y, Xiao H, Tan R, Xing Y, et al. Driver's speech emotion recognition for smart cockpit based on a self-attention deep learning framework. In: *2021 5th CAA international conference on vehicular control and intelligence (CVCI)*. Tianjin, China: IEEE (2021). p. 1–5. Available online at: https://ieeexplore.ieee.org/document/9661268/ (Accessed February 17, 2025).

7. Guo L, Shen Y, Ding P. An algorithm of emotion recognition and valence of drivers on multimodal data. In: *2022 IEEE international symposium on broadband multimedia systems and broadcasting (BMSB)* (2022). p. 1–5.

8. Leone A, Caroppo A, Manni A, Siciliano P. Vision-based road rage detection framework in automotive safety applications. *Sensors* (2021) 21(9):2942. doi:10.3390/s21092942

9. Wu L, Liu M, Li J, Zhang Y. An intelligent vehicle alarm user terminal system based on emotional identification technology. *Sci Program* (2022) 2022:1–11. doi:10.1155/2022/6315063

10. Roka S, Rawat DB. Fine tuning vision transformer model for facial emotion recognition: performance analysis for human-machine teaming. In: *2023 IEEE 24th international conference on information reuse and integration for data science (IRI)* (2023). p. 134–9.

11. Li W, Xue J, Tan R, Wang C, Deng Z, Li S, et al. Global-local-feature-fused driver speech emotion detection for intelligent cockpit in automated driving. *IEEE Trans Intell Vehicles* (2023) 8(4):2684–97. doi:10.1109/tiv.2023.3259988

12. Oh G, Jeong E, Kim RC, Yang JH, Hwang S, Lee S, et al. Multimodal data collection system for driver emotion recognition based on self-reporting in real-world driving. *Sensors* (2022) 22(12):4402. doi:10.3390/s22124402

13. Mou L, Zhao Y, Zhou C, Nakisa B, Rastgoo MN, Ma L, et al. Driver emotion recognition with a Hybrid attentional multimodal fusion framework. *IEEE Trans Affective Comput* (2023) 14:2970–81. doi:10.1109/taffc.2023.3250460

14. Ekman P. Facial expression and emotion. *Am Psychol* (1993) 48:384–92. doi:10.1037//0003-066x.48.4.384

15. Köpüklü O, Kose N, Gunduz A, Rigoll G. Resource efficient 3D convolutional neural networks. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (2019). p. 1910–1919. doi:10.1109/ICCVW.2019.00240

16. Hsu WN, Bolte B, Tsai YHH, Lakhotia K, Salakhutdinov R, Mohamed A. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. *arXiv* (2021) 29:3451–60. doi:10.1109/taslp.2021.3122291

17. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018). p. 4510–4520. doi:10.1109/CVPR.2018.00474

18. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. *arXiv* (2017). doi:10.48550/arXiv.1704.04861

19. Yu W, Xu H, Meng F, Zhu Y, Ma Y, Wu J, et al. CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: D Jurafsky, J Chai, N Schluter, J Tetreault, editors. *Proceedings of the 58th annual meeting of the association for computational linguistics*. Association for Computational Linguistics (2020). p. 3718–27. Available online at: https://aclanthology.org/2020.acl-main.343 (Accessed February 16, 2025).

20. Wu Z, Gong Z, Koo J, Hirschberg J. Multimodal multi-loss fusion network for sentiment analysis. *arXiv* (2024) 3588–602. doi:10.18653/v1/2024.naacl-long.197

21. Zhang H, Wang Y, Yin G, Liu K, Liu Y, Yu T. Learning Language-Guided adaptive hyper-modality representation for multimodal sentiment analysis (2023). Available online at: http://arxiv.org/abs/2310.05804 (Accessed February 23, 2025).