Application of Multimodal Emotion Recognition in Intelligent Vehicles

Yongrui Wu *

School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China

* Corresponding Author Email: wuyongrui@bupt.edu.cn

Abstract. Road traffic safety is a critical global issue, as approximately 1.2 million people die from road traffic accidents annually. Driver behavior is influenced by negative emotions like anger, anxiety, and fatigue, which is a key contributing factor of road accidents. Multimodal emotion recognition, which integrates diverse data types such as images, audio, physiological signals, and vehicle data, significantly enhances the accuracy and robustness of emotion recognition systems, playing a important role in preventing potential accidents. This paper systematically analyzes representative research on multimodal emotion recognition used in automobile scenarios from 2020 to 2025, focusing on core technologies including data preprocessing, feature extraction, and fusion strategies. It critically analyzes algorithm frameworks while examining typical cases, and discusses limitations in privacy and public datasets offering potential ways of addressing them. It aims to provide comprehensive references for advancing driver emotion recognition research in intelligent vehicles. Ultimately facilitating accident prevention and optimizing vehicle interaction experiences.

Keywords: multimodal; emotion recognition; traffic accident.

1. Introduction

1.1. Research Background

Road traffic accident is an important topic of global public safety. According to the World Health Organization (WHO), approximately 1.2 million people die from road traffic accidents each year, and driver behavior is an important factor leading to accidents [1]. Fig. 1 shows the percentage distribution of various accident causes

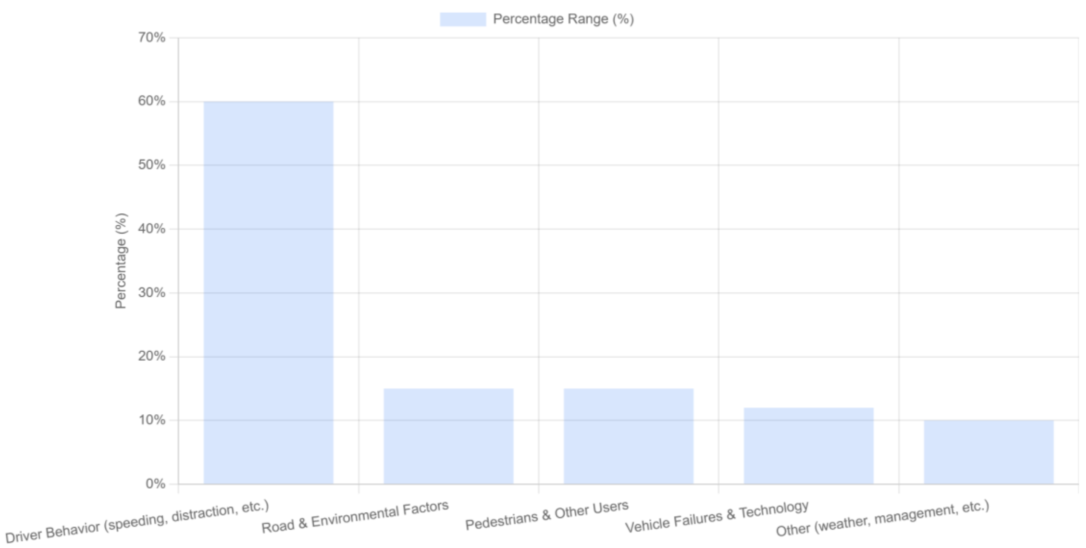

Fig. 1 Analysis of Road Traffic Accident Causes

Negative emotions such as fatigue, irritability, anger, and anxiety are key factors affecting driving behavior [2]. Studies have shown that anger may lead to aggressive driving behavior, increasing the risk of an accident by nearly 10 times, while anxiety can distract the drivers' attention and increase their reaction time. Therefore, real-time recognition of driver emotions can be useful in preventing potential accidents and ensuring road traffic safety.

The idea of using computers for emotion recognition emerged as early as 1997. However, early emotion recognition was mostly implemented through artificially predefined rules, based on single-modal data, and not sufficiently integrated with driving scenarios, making it difficult to handle the complexity and diversity of emotions in real driving scenarios. With the rapid development of technologies such as convolutional neural networks (CNN) and recurrent neural networks (RNN), and the appearance of computing algorithms such as support vector machines, decision trees, and random forests, the accuracy and anti-interference ability of emotion recognition system combining multimodality have been significantly improved [3].

1.2. Main Contents

The application of multimodal emotion recognition in the driving environment not only helps solve problems such as fatigue driving and road rage, ensuring traffic safety, but also contributes to improve car driving experience and promotes the humanized and intelligent development of automobiles. This article focuses on the specific scenario of vehicle driving, and systematically analyses the representing research results of multimodal emotion recognition in the past five years (2020-2025). Comparing important technologies in multiple stages such as data preprocessing, feature extraction, multimodal fusion, and evaluating performance on various data modal including images and audio, discusses the limitations of new technologies, and ultimately aims to provide references and guidance for driver emotion recognition research.

2. Related Research and Technologies on Emotion Recognition used in Driving

2.1. Background Knowledge

Fig.2 shows the data types used for driver emotion recognition, which mainly includes image data, audio data, physiological signals, and vehicle data. Among them, image data specifically, includes videos and pictures; physiological signals include electrocardiogram (ECG), electrodermal activity (EDA), electroencephalogram (EEG), etc.; vehicle data mainly includes vehicle control operations and vehicle dynamic parameters [4].

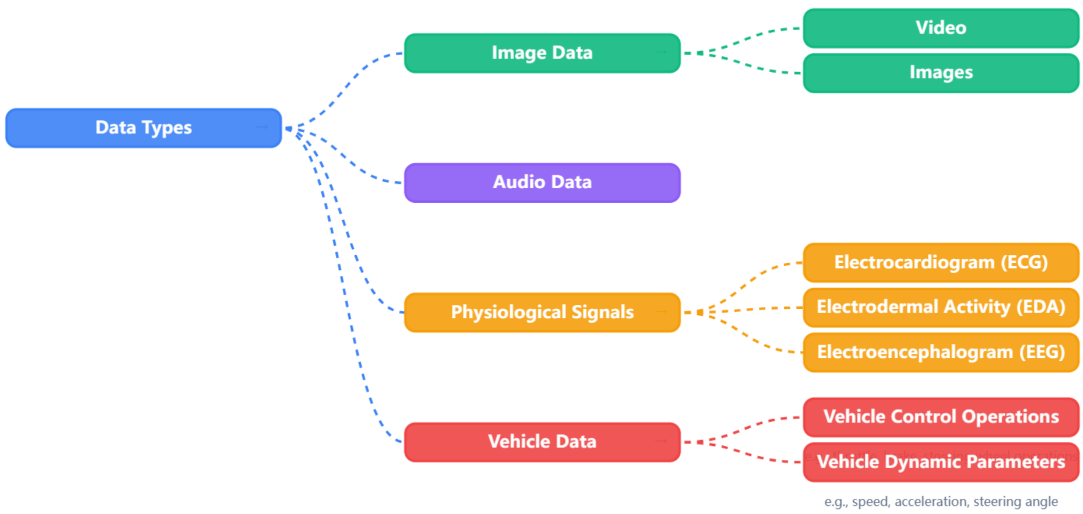

Fig. 2 Data Types used for Driver Emotion Recognition

The theoretical models describing emotions are mainly divided into two categories: discrete emotion models and dimensional emotion models [4,5]. The discrete emotion model divides emotions into independent types. For example, Ekman's theory proposes six basic emotions: joy, sadness, anger, fear, disgust, and surprise [6]. The continuous dimensional emotion model, on the other hand, holds that emotions are continuous, which can be regarded as points in a multi-dimensional space. Common dimensional emotion models include the Pleasure-Arousal-Dominance space theory (PAD space), Valence-Arousal-Dominance theory (VAD space), etc.

The goal of multimodal emotion recognition is to achieve real-time emotion recognition through the process of "data preprocessing, feature extraction, multimodal fusion, final result". Preprocessing raw data before extracting data features can effectively improve the efficiency and generalization ability of the model [7]; Generally speaking, the feature extraction process uses strategies in traditional machine learning and deep learning, to process image data, audio data, physiological signals, and vehicle data; modal fusion analyses the results obtained from different data modalities, synthesizes various features to determine the final emotion. Multimodal fusion methods can be divided into data-level fusion, feature-level fusion, decision-level fusion, etc. depending on the stage the fusion happens.

2.2. Examples of Typical Studies

Mou et al. proposed a multimodal fusion framework of Convolutional Long Short-Term Memory Network (ConvLSTM) and hybrid attention mechanism to fuse driver emotion features, which for the first-time combined eye, vehicle, and environmental data to recognize driver emotions. The study collected data from 19 participants through an advanced driving simulator, and used the levels of Valence-Arousal-Dominance (VAD space) to represent driver emotions [8]. The accuracy rates of the three dimensions in the reserved scene cross-validation all reached more than 95%; in the reserved subject cross-validation, they all reached more than 80%. This study focuses on emotion recognition in the context of driving and considers the connection between environment pressure and emotions. They found that drivers' pleasure and dominance levels are higher under low pressure, and emotional distribution is more scattered under high pressure. It provides a feasible method for emotion recognition in intelligent driving.

Sonali focuses on the application of AI-driven emotion recognition technology in vehicle safety [7]. Deep learning methods such as Convolutional Neural Network (CNN) and Long Short-Term Memory Network (LSTM) are used to process images, audio, and text. They classify driver emotions are into: happiness, sadness, neutrality, fear, disgust, surprise, and anger. The system achieves high accuracy rates of 89.99%, 85.11%, and 96.92% for expressions, voices, and texts respectively. A specific algorithm is defined to fuse the results of the three modalities, before audio, text, and visual prompts are generated and personalized music is recommended to adjust the driver's emotions. Studies have shown that emotion recognition based on multimodality can effectively improve recognition accuracy, but there is still some distance from practical applications.

Ying et al. proposed a non-contact multimodal driver emotion recognition algorithm based on audio and video signals [9]. The raw data is processed and fused at the feature layer to retain original details and improve feature discriminability. Combined with the Internet of Vehicles (IoV) platform, the recognition accuracy and robustness are improved.

2.3. Data Preprocess and Feature Extraction Technologies

Data preprocessing is a key step in driver multimodal emotion recognition. This step aims to improve data quality through systematic processing and make sufficient preparations for the subsequent feature extraction. Raw data is easily affected by factors such as environment or individual differences during data collection. At the same time, data formats may be inconsistent. To reduce such interference, it is necessary to preprocess the data. It is worth mentioning that Ying's research extracts key frames of video data through LBPH features in the video preprocessing stage [9].

Specifically, in a frame of a video, the LBP value of a certain point c is given by the following formula, where I(c) am the grayscale value of the point, and p are the 8 positions adjacent to the center c.

$$LBP(x_c, y_c) = \sum_{p=1}^{8} 2^p s(I(p) - I(c)) \tag{1}$$

$$s(x) = \begin{cases} 1, & x > 0 \\ 0, & x \le 0 \end{cases} \tag{2}$$

On this basis, the image is divided into blocks to calculate LBP features and generate histograms. Finally, the histograms of all blocks are merged to obtain a global feature vector for face recognition tasks.

As for audio data, Short-Time Fourier Transform (STFT) is used to extract the audio spectrum and it is mapped to Mel spectrogram for further analysis. Here, x is the original signal, and w is the window function, which is used to intercept local segments of the signal.

$$STFT_x(t, \omega) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-j\omega\tau} d\tau \tag{3}$$

Since Convolutional Neural Network (CNN) is suitable for capturing spatial features, and Long Short-Term Memory Network (LSTM) is good at capturing dependencies in time-series data, it is a common to use CNN model when dealing with image data and ConvLSTM model when processing audio data. Mou et al. used z-score preprocessing for audio normalization, ConvLSTM to extract audio spatiotemporal features, and hybrid attention to fuse features. Ying et al. used a fine-tuned ResNet18, temporal attention, and bidirectional GRU to extract emotional features of key video frames, and expanded multi-scale residual connections based on residual networks to extract multi-dimensional emotional features of audio spectrograms [9]. Sonali also used a hybrid model of LSTM and CNN to process text data [7].

2.4. Modal Fusion Methods

In a driver multimodal emotion recognition system, modal fusion is the step to integrate the information from different data modalities [5]. Fusion can be roughly divided into three types depending on the stage it occurs: Raw data fusion, intermediate stage fusion, and post-decision fusion [4].

Raw data fusion refers to directly integrating the data of different modalities after the raw data is preprocessed and before feature extraction. Fusion at an early stage converts multimodal data into a unified data format, aiming to allow the model to independently learn emotional features without predefined rules, but it faces problems such as high preprocessing difficulty and the risk of overfitting.

Intermediate stage fusion is to fuse the extracted feature vectors after each modal data has completed feature extraction, which belongs to feature-level fusion. It first extracts effective features through modal-specific models, and then integrates high-level features of different modalities, which is a relatively common fusion method.

Post-independent decision fusion is the latest stage of fusion. Each data modality first independently makes a decision, and then the system integrates multiple decision results to output the final emotion conclusion. Sonali uses a specific algorithm to make the final judgment with the results from the three modalities. When a certain modality fails, the system can still work normally, with high robustness.

3. Summary

At the technical level, with the continuous development of technologies such as convolutional neural networks and recurrent neural networks, algorithms such as support vector machines, decision trees, and random forests have been applied to emotion recognition, and the accuracy and efficiency of emotion recognition have been significantly improved, enabling more effective processing of various modal data such as images, audio, and text. Driver emotion recognition has developed from single-modal to multimodal (visual, audio, environment), which has greatly improved the accuracy and efficiency of emotion recognition.

Existing emotion recognition research has made great progress, but there are still limitations in some aspects. In terms of privacy protection, although Ying et al. have considered encryption mechanisms, which to a certain extent protect the security of sensitive information including facial images, voices, and physiological signals, the raw data still has the risk of leakage [9]. In the future, advanced anonymization technologies can be developed or appropriate noise can be added to the data to prevent personal privacy leakage on the premise of not affecting the model's extraction and recognition of emotional features. At the same time, the public data sets relied on by the research, such as RAVDESS, SAVEE, and FER-2013, have limited sample sizes and narrow coverage of participants. For example, SAVEE has only 4 males, and parts of Mou et al.'s research has only included 19 participants, which is difficult to cover different types of drivers [8]. This could result in limited generalization ability of the system.

In conclusion, the application of multimodal emotion recognition in intelligent driving will develop in a more accurate, robust, and humanized direction. In the future, this technology will be deeply integrated into intelligent cockpits, will help to achieve early warning of negative driver emotion in personalized cockpits, making driving safer and more humanized.

References

[1] World Health Organization. Global status report on road safety 2023. Geneva: World Health Organization, 2023. DOI:https://iris.who.int/bitstream/handle/10665/375016/9789240086517-eng.pdf

[2] Shams Zhaleh; Naderi Hossein; Nassiri Habibollah. Assessing the effect of inattention-related error and anger in driving on road accidents among Iranian heavy vehicle drivers. IATSS Research,2020,45(2).

[3] Abdullah, Sharmeen M. Saleem Abdullah, et al. Multimodal emotion recognition using deep learning. Journal of Applied Science and Technology Trends 2.01 (2021): 73-79.

[4] Zepf, Sebastian, et al. Driver emotion recognition for intelligent vehicles: A survey. ACM Computing Surveys (CSUR) 53.3 (2020): 1-30.

[5] Zhou, Xinying, et al. Review of Multi-Modal Driver Emotion Recognition. Computer Engineering and Applications,2025,61(10):1.

[6] Ekman, Paul. Emotions revealed: Recognizing faces and feelings to improve communication and emotional life. Times Books, 2003.

[7] Jadhav, Sonali Subhash. Advancing Safety in Vehicles with AI-Driven Emotion Recognition. Diss. Dublin, National College of Ireland, 2024.

[8] Mou, Luntian, et al. Driver emotion recognition with a hybrid attentional multimodal fusion framework. IEEE Transactions on Affective Computing 14.4 (2023): 2970-2981.

[9] Ying, Na, et al. A multimodal driver emotion recognition algorithm based on the audio and video signals in internet of vehicles platform. IEEE Internet of Things Journal 11.22 (2024): 35812-35824.