Depression Detection Method Based on Multimodal Emotion Recognition

Yihao Huang

Department of University of Leeds, Leeds, LS2 9JT, United Kingdom

* Corresponding Author Email: el23yh2@leeds.ac.uk

Abstract. Depression screening traditionally relies on subjective questionnaires, which face limitations such as low efficiency and high risk of misdiagnosis. To address this, this study proposes a real-time depression detection system based on multimodal emotion recognition, integrating speech and micro-expression features. The system leverages prosodic and acoustic biomarkers in speech, along with micro-expression analysis via the FACS framework, to enhance objectivity and reliability. Two representative models are compared in this study: the MFM-Att model, which adopts multi-level attention to achieve comprehensive multimodal fusion, and the Q-learning model, which applies reinforcement learning to support adaptive and lightweight detection. The experiments demonstrate that multimodal fusion methods outperform single-modality approaches, with the MFM-Att model showing higher accuracy and the Q-learning model exhibiting better real-time adaptability. To further balance accuracy and efficiency, the optimized framework integrates MobileNet-ViT with adaptive weight adjustment. Overall, this work underscores the potential of multimodal emotion recognition in early depression screening and provides technical support for scalable clinical applications and personalized mental health interventions.

Keywords: Depression detection; Multimodal emotion recognition; Speech biomarkers; Micro-expression analysis; Multimodal fusion.

1. Introduction

1.1. Research Background

Right now, the assessment of depression is still carried out mostly with subjective rating scales. Although these scales are common in clinical practice, they are far from perfect-very often they are inefficient, and sometimes they fail to identify real cases or end up giving misleading results [1,2]. In the past few years, however, the rapid progress in multimodal emotion recognition has offered researchers a chance to move past these drawbacks. The use of objective indicators for real-time screening has become a realistic option, and this shift opens up a promising direction for detecting mental health issues at an earlier stage [3,4].

What this study sets out to do is to make depression detection not only more accurate but also quicker. To reach that goal, we bring together speech features with micro-expression analysis, with the intention of offering a more solid technical basis for timely mental health intervention [5].

1.2. Technological Base

The current technical foundation is relatively complete. The new methods for detecting depression mainly integrate and optimize three core technologies: voice biomarkers, micro-expression detection, and multimodal fusion [3].

Voice biomarkers can be divided into prosodic features and acoustic features. Studies have shown that patients with depression often exhibit features such as slower speech rate and lower voice pitch in their verbal expressions [1]. These features can be identified from the prosodic and acoustic dimensions through corresponding algorithms.

In micro-expression detection, existing technologies can accurately capture the micro-expression features of patients with psychological abnormalities by using the FACS system [6]. Due to the subjective nature and unstable effect of traditional face-to-face diagnosis, using micro-expression detection can significantly improve the objectivity and reliability of the diagnosis [2].

In multimodal fusion, it mainly includes two paths: feature-level fusion and decision-level fusion. Feature-level fusion uses the FFN network to achieve molecular-level integration of features such as vocal tremor and resonance peak, to improve diagnostic accuracy [7]. Decision-level fusion uses adaptive allocation of the weight ratio of each modality to achieve more personalized diagnostic strategies, thereby further improving the overall recognition accuracy [3].

1.3. Research Significance

This research aims to build a real-time screening system based on multimodal emotion recognition, optimize the cross-modal feature fusion algorithm, and achieve efficient and non-intrusive preliminary screening for depression, thereby helping more individuals understand and actively cope with mental health issues [4,5].

2. A Comparative Study on Cross-modal Fusion and Adaptive Learning Mechanisms for Intelligent Depression Diagnosis and System Optimization

2.1. Performance Comparison of Depression Detection Models Based on MFM-Att and Q-learning

Currently, there are various advanced methods for detecting depression both domestically and internationally. Among them, the MFM-Att model is relatively advanced in China [8]. This model can extract information from three modalities: audio, vision, and text, and process it through two multi-modal feature fusion methods: intra-modal attention fusion and inter-modal attention fusion [7]. The former is used to extract key sub-features within each modality, such as identifying the most emotionally intense sentences in the text modality; the latter maps the final features of each modality to a unified dimension to enhance the fusion effect. The model ultimately uses the PHQ-8 scale score as a measure of depression severity [1].

One of the more advanced detection methods abroad is Q-learning, which has a long development history [9]. Its advantage lies in its self-learning ability. It extracts features from audio and visual modalities using a CNN network and processes them with MFCC coefficients [1]. The core of Q-learning is the reinforcement learning algorithm, which can continuously optimize the detection strategy through repeated testing [9]. Although the initial recognition accuracy may be low, it will continue to improve with long-term training. Additionally, this method has the potential to formulate personalized treatment plans for individuals and is of great value in the long-term monitoring of a single patient [5,9].

Domestic models have discovered through continuous experiments and ablation studies that the method of integrating the three modalities achieves the highest detection accuracy. Single-modal detection results are generally poor, while multi-modal fusion (especially three-modal) shows significant advantages [8]. As shown in the comparison chart, regardless of the average absolute error or root mean square error indicators, the MFM-Att model performs the best among existing methods [8].

The Q-learning model abroad also has excellent performance. It automatically optimizes the recognition strategy through reinforcement learning and introduces a Target Q-Network to achieve stable training, thereby further improving accuracy [9]. Compared to the large structure of MFM-Att, the Q-learning model has a smaller computational volume and is more suitable for lightweight real-time applications [9,10]. Although its model size is small, the detection accuracy is as high as 90.2%, and its accuracy and personalized recommendation capabilities can be further enhanced in the long-term monitoring of a single patient [9].

Overall, the MFM-Att model is more accurate and suitable for clinical and research environments [8]; the Q-learning model is more lightweight and suitable for real-time detection in practical scenarios [9].

The main advantage of the MFM-Att model lies in integrating the three modalities, providing more comprehensive information, and automatically extracting key features, thereby improving detection accuracy [7,8]. The main drawback of the MFM-Att model lies in its complexity, high computational demand, slow processing speed, and substantial hardware requirements, which limit its suitability for long-term monitoring [8].

In contrast, the Q-learning model does not depend on environmental modeling, features a lightweight structure, and demonstrates strong adaptability with continuous learning ability, making it more practical for long-term tracking [9]. However, its training phase requires significant computational resources, its initial accuracy is lower, and its final detection results are slightly weaker than those of the MFM-Att model [8,9].

From the perspective of future application, the MFM-Att model is better suited for standardized screening, while the Q-learning model shows greater potential for personalized tracking and monitoring through mobile devices [9,10]. The innovation of Q-learning lies in introducing reinforcement learning into emotion recognition [9]. By continuously interacting with the environment and optimizing recognition strategies, it improves adaptability and robustness [9]. Its online learning capability allows dynamic adjustment based on user feedback, enabling personalized, real-time monitoring and screening [5,9].

2.2. System Optimization Framework

Regarding system optimization, both models have achieved notable advancements [3,8].

Firstly, both adopt the MobileNet-ViT hybrid architecture to enhance computational efficiency. The ViT model focuses on extracting detailed three-dimensional facial features, while the MobileNet can quickly analyze the speech spectrum. The combination of the two significantly improves the processing performance of multimodal fusion [10].

Secondly, both have implemented intelligent weight adjustment algorithms, that is, the adaptive adjustment of modal weights, to enhance the adaptability for clinical applications [7]. However, there are differences in the implementation methods: MFM-Att dynamically weights different modalities through the attention mechanism to optimize the fusion effect [7]; Q-learning uses the policy network in reinforcement learning to control and select modalities and weights, continuously optimizing the modal combination through learning historical feedback, demonstrating a more proactive strategy learning ability [9].

3. The Technical Bottlenecks and Clinical Transformation Paths of The Multimodal Depression Detection System

3.1. Technical Challenges

Although these two models perform well both in theory and experiment, they still encounter many technical challenges in practical applications. The main issues include the following three points:

Firstly, the problem of small sample learning. In the case of emotion recognition for depression screening, one of the core problems is the scarcity of sample data [1]. Compared to the fields of natural language processing or image recognition, the collection of multimodal data for patients with depression is more difficult [3]. To address this issue, pre-trained models and lightweight network architectures can be adopted to extract effective emotional features from speech and visual sequences, ensuring recognition performance even in the case of small samples [10].

Secondly, the complexity of multimodal feature fusion. The data structures and feature expressions of different modalities are significantly different, which poses challenges to the fusion process [7]. To address this problem, a "fusion feature function" as shown in the figure can be used to concatenate and cross-fuse multimodal emotional features and then input them into the attention mechanism and fully connected network (FFN) for processing. This formula expresses the feature concatenation and cross-product between different modalities, thereby enhancing the model's ability to understand cross-modal emotional expressions [7].

Thirdly, the requirement for real-time performance. The emotion screening system needs to balance accuracy and response speed during actual deployment [10]. To improve real-time performance, modular design and parameter optimization can be adopted to achieve fast reasoning without sacrificing recognition accuracy [10].

Due to the large structure of the MFM-Att model and high computational complexity, it requires a large amount of computing resources. To alleviate this problem, a modality-adaptive mechanism can be introduced into the model, enabling it to automatically adjust the importance of allocation of speech and micro-expressions based on the input content, thereby improving the inference speed while ensuring recognition accuracy [7].

3.2. Practical Application Challenges

At the practical application level, the emotion recognition system needs to handle highly sensitive user data including voice, video, and facial expressions, which brings ethical and legal risks during data collection, transmission, and storage [5]. Therefore, in system design, mechanisms such as local processing, edge computing, data encryption, and federated learning should be introduced to enhance the privacy protection level of user data [5].

Based on the improvement of the privacy protection mechanism, it is still necessary to ensure the legality and compliance of the model. After the model enters the practical application stage, it is necessary to ensure that the use of all data is based on the informed consent of users, and to effectively protect the legitimate rights and interests of users and those being detected.

In addition, in the real application environment, voice collection is prone to be interfered with by background noise, affecting the recognition quality. To compensate for the loss of voice features, text modalities can be introduced as auxiliary information sources to enhance the robustness of the system. Any emotion recognition system that intends to be applied in clinical depression screening must undergo strict clinical-level verification [2]. The following figure shows the comparison results of multiple research models (including MFM-Att) in terms of clinical applicability, indicating that this system has a good application prospect. Of course, further related control experiments need to be carried out to ensure the stability and reliability of the system in the clinical environment.

In different cultural backgrounds, individuals have significant differences in emotional expression. For example, patients in Europe and America are better at verbal expression, while patients in Asia have more obvious micro-expressions [6,11]. Regarding this issue, an adaptive normalization mechanism can be considered, and the sample collection range can be expanded to improve the adaptability and universality of the model for multi-cultural scenarios [12].

3.3. Applied Analysis

In response to practical application requirements, this study proposes two main functional scenarios, namely "Identification of Resistance to Medical Consultation Behavior" and "Clinical Auxiliary Diagnosis":

On one hand, considering that some patients with depression refuse to seek medical treatment due to feelings of shame or psychological burden, in the future, when the model and technology are mature, the behavior recognition module can identify the patients' resistant behaviors, and use non-contact data collection methods for emotional monitoring. Patients do not need to enter a closed space, and the system can automatically capture their multimodal emotional data, improving the efficiency and convenience of screening [5].

On the other hand, in clinical applications, mature models can achieve remote mental health detection and personalized feedback. By embedding the patients commonly used electronic devices, continuous tracking of their daily emotional states can be realized. Even if the doctor is not present on site, they can provide timely and appropriate treatment suggestions for the patients, enhancing the feasibility and effectiveness of remote intervention [5].

4. Conclusion

In conclusion, the multimodal depression screening system proposed in this study fully integrates voice and micro-expression features. Based on the adaptive modal fusion mechanism, it demonstrates stronger robustness and flexibility in heterogeneous information fusion. At the technical level, the designed architecture not only improves the recognition performance but also significantly optimizes the real-time performance and scalability of the system.

In terms of clinical value, the multimodal emotion recognition technology provides a new approach for the early screening of depression and is expected to achieve the goals of early identification, early intervention, and early treatment.

However, emotion recognition still faces challenges such as subjective data annotation and complex temporal synchronization between modalities. To promote this technology from the laboratory to practical applications, the key lies in the deep integration of interdisciplinary fields and long-term clinical validation.

In the future, we look forward to continuous exploration in the intersection of artificial intelligence and mental health, and to promoting the emotional computing technology to better serve the protection and improvement of human mental health.

References

[1] Kroenke Kurt, Strine Terrence W., Andrews Gavin, et al. The PHQ-8 as a measure of current depression in the general population. Journal of Affective Disorders, 2009, 114(1–3): 163–173.

[2] Cabanac M. What is emotion? Behavioural Processes, 2002, 60(2): 69–83.

[3] Fang Ming, Peng Siyu, Liang Yujia, Hung Chih-Cheng, Liu Shuhua. A multimodal fusion model with multi-level attention mechanism for depression detection. Biomedical Signal Processing and Control, 2023, 82: 104561.

[4] John E. Laird. Emotion. In: The Soar Cognitive Architecture. The MIT Press, 2012.

[5] David Matsumoto (ed.), Hyisung C. Hwang (ed.). Culture and Emotion: Integrating Biological Universality with Cultural Specificity. Oxford University Press, July 2019.

[6] Rejaibi E, Komaty A, Meriaudeau F, Agrebi S, Othmani A A. MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech. Biomedical Signal Processing and Control, 2022, 71(A): 103107.

[7] Dongdong Li, Jinlin Liu, Zhuo Yang, Linyu Sun, Zhe Wang. Speech emotion recognition using recurrent neural networks with directional self-attention. Expert Systems with Applications, 2021, 173: 114683.

[8] Sardari S, et al. Audio based depression detection using Convolutional Autoencoder. Expert Systems with Applications, 2022, 189: 116076.

[9] Dai Zhijun, Zhou Heng, Ba Qingfang, et al. Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. Journal of Affective Disorders, 2021, 295: 1040–1048.

[10] Kalateh S, Estrada-Jimenez L A, Nikghadam-Hojjati S, Barata J. A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. IEEE Access, 2024, 12: 103976 - 104019.

[11] Cacioppo J T, Gardner W L. Emotion. Annual Review of Psychology, 1999, 50: 191–214.

[12] Roseman I. J., Dhawan N., Rettek S. I., Naidu R. K., & Thapa K. Cultural differences and cross-cultural similarities in appraisals and emotional responses. Journal of Cross-Cultural Psychology, 1995, 26(1).