Strategies and Research on Improving Emotional Recognition of Elderly People Based on Human-Computer Interaction Perspective

Weijie Li[a]
Data Science and Engineering, South China Normal University, Shanwei, Guangdong, China

Keywords: Elderly, Emotion Recognition Models, Emotional Needs, Silver Economy.

Abstract: The number of elderly people has increased dramatically, and aging is becoming more and more serious. The children of the elderly cannot accompany them, resulting in more and more elderly people having mental health problems. Since emotional health affects the quality of life and overall health of the elderly, this paper focuses on improving emotional health, and accurate recognition of emotions is the premise of emotional intervention and improvement. This paper analyzes multimodal recognition methods based on voice and face, facial image and heart rate variability, and electroencephalogram (EEG) signal fusion with face image. The data sets on which different emotion recognition methods rely are sorted out and compared. In addition, the existing limitations and future prospects are also studied. The purpose of this study is to summarize the research of previous researchers on emotion recognition models and the design ideas for products that can meet the emotional needs of the elderly, so as to provide relevant ideas for subsequent researchers in improving the emotional value of the elderly and improving emotion recognition models.

1 INTRODUCTION

The aging population is becoming increasingly severe, and due to the long work hours and the need for some time for self adjustment and relaxation, young children have very little time to spend with their parents, resulting in the emergence of many empty nests for elderly people, especially those who are only children, who face greater pressure in retirement (Zhao, 2010), due to a prolonged lack of companionship from their children, these elderly people have developed psychological and emotional problems over time (Wang, 2025), similarly, it can also lead to various physical problems, which make young people today face great pressure in retirement. However, due to the low accuracy of existing emotion recognition models in emotion recognition, research on emotion recognition for the elderly is also lacking (Huang et al., 2024), thus, it can be reflected that the current emotion recognition model is difficult to meet the requirements of accurately recognizing the emotions of the elderly, and there is a lack of products on the market that can effectively meet the emotional needs of the elderly(Guo, 2024), due to the two reasons mentioned above, it has been impossible to open up the market in the silver economy. This article focuses on the research direction of multimodal emotion recognition based on speech and facial images, multimodal fusion emotion recognition algorithm based on facial images and heart rate variability, and deep learning method for multimodal emotion recognition combining EEG signals and facial images, as well as several publicly available datasets. The significance of this article is to provide relevant ideas and suggestions for future researchers in optimizing emotion recognition models and meeting the emotional needs of the elderly.

2 EMOTION RECOGNITION METHODS FOR ELDERLY PEOPLE

2.1 Emotion Recognition based on Speech and Facial Features

Dual modal recognition of speech and face can quickly determine the current emotions and psychological state of the elderly, especially for the

[a] https://orcid.org/0009-0004-1718-2960

Li, W.
Strategies and Research on Improving Emotional Recognition of Elderly People Based on Human-Computer Interaction Perspective.
DOI: 10.5220/0014360300004718

In *Proceedings of the 2nd International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2025)*, pages 399-405
ISBN: 978-989-758-792-4

special group of elderly people who need to address their negative emotions in a short period of time. Therefore, it is particularly important to use dual modal recognition of speech and face to judge the emotional changes of the elderly.

By studying the dual modal emotion recognition of speech and facial images, a feature extraction model of Multi branch Convolutional Neural Networks combined with channel spatial attention mechanism is proposed for the speech modality, and a feature extraction model of Residual Hybrid Convolutional Neural Networks is proposed for the facial image modality. The extracted speech and facial image features are classified and recognized separately through classification layers, and decision fusion is used to perform final fusion classification on the recognition results. The experimental results show that the proposed dual modal fusion model is effective in RAVDESS, eNTERFACE'05, RML The recognition accuracies on the three datasets reached 97.22%, 94.78%, and 96.96%, respectively, which were 11.02%, 4.24%, and 8.83% higher than the recognition accuracies of single modal speech, and 4.60%, 6.74%, and 4.10% higher than the recognition accuracies of single modal facial images, respectively. Compared with relevant methods on the corresponding datasets in recent years, these accuracies have also been improved(Xue et al., 2024).

By studying the dual modal emotion recognition of speech and facial images, it is found that most existing emotion recognition algorithms rely on a single perceptual modality for construction, and there are still some shortcomings in the research of multimodal recognition. Therefore, parallel convolution module (Pconv), attention mechanism based bidirectional long short-term memory network (BiLSTM Attention), and cross attention fusion module are proposed. A multi-scale convolution kernel combining 3D convolution and 2D convolution is used to improve the Inception Res Net V2 to construct an expression emotion recognition model for continuous video frames, reducing computational difficulty. CH-SIMS and CMU-MOSI are utilized to improve the expression feature extraction of continuous video frames Two publicly available emotion datasets were used to validate the effectiveness of the emotion recognition model proposed in this paper. The experimental results showed that the proposed models achieved higher recognition accuracy than existing audio and video baseline models, and each component of the model contributed to the improvement of model performance. The proposed multimodal model achieved higher recognition accuracy than the baseline model, reaching 97.82% and 98.18% respectively, demonstrating the effectiveness of the proposed cross attention based multimodal emotion model(Wu, 2024).

2.2 Emotion Recognition based on Facial Images and Heart Rate Variability

Compared to speech and facial images, physiological signals often have higher stability and objectivity in emotional changes, especially for elderly people with unclear facial expressions or decreased language communication abilities. Emotion recognition methods based on heart rate variability (HRV) and facial images show stronger adaptability and recognition accuracy.

Propose a multimodal fusion emotion recognition algorithm based on facial images and heart rate variability (MFER-FIHRV). By analyzing and capturing data information from both facial images and heart rate variability modalities, MFER-FIHRV can sensitively perceive users' emotional changes and create personalized human-computer interaction experiences that fit their current state. Firstly, a multimodal fusion Transformer is designed to perform multimodal complementary learning on facial images and heart rate variability. Then, multimodal feature fusion is used to concatenate the fused features with the original features. Additionally, a lightweight self attention mechanism is employed to learn advanced representations within the multimodal domain. A large number of experiments were conducted on two publicly available datasets, and the results showed that the proposed method had better performance. This experiment demonstrates that the proposed method is effective and can guide the design and development of user experience systems (Zhang et al., 2025).

The research mainly adopts a fusion algorithm based on attention mechanism and residual thinking for multimodal emotion recognition of EEG, peripheral physiology, and facial expression signals. Relevant methods of deep learning are used to extract emotional features from EEG modality, peripheral physiology modality, and facial expression modality, targeting peripheral physiology modality, The DEAP database uses traditional methods to extract shallow features and then uses neural networks for emotion feature extraction. The MAHNOB-HCI database uses convolutional neural networks for emotion feature extraction, fully considering the differences in different types of peripheral physiological signals. For facial expression patterns, the CNN-LSTM network model is used for emotion feature extraction of expression videos. Experimental results show that compared with multimodal emotion recognition, the multimodal recognition results of EEG, peripheral physiological, and facial expression signals have also been greatly improved. In the MAHNOB-HCI database, the dimension sentiment labels were divided into three categories for experimentation, and the above results also exist. Both methods have good generalization ability and robustness (Zhu, 2021).

2.3 Emotion Recognition based on EEG Signals and Facial Images

The emotion recognition method based on EEG signals and facial images can simultaneously combine internal brain activity and external facial expressions, with higher accuracy and real-time response capability in emotion recognition. In contrast, heart rate variability is influenced by multiple physiological factors, with lower emotional specificity and weaker information complementarity with facial images. Therefore, the fusion of EEG and facial images is more robust in complex environments and can comprehensively reflect an individual's true emotional state. The overall recognition effect is better than the combination of facial images and heart rate variability.

A multi-level spatiotemporal feature adaptive integration and unique shared feature fusion model, as well as a multi granularity attention and feature distribution calibration model, are proposed to address the problems in the field of multimodal emotion recognition combining EEG signals and facial images. The loss function is used to constrain the similarity or difference between each feature, ensuring the model's ability to capture the unique emotional semantic information of each modality and the shared emotional semantic information between modalities. The multi-level spatiotemporal feature adaptive integration and unique shared feature fusion model and method were cross validated on the DEAP and MAHNOB-HCI datasets, with values of 82.60%/79.99%, 83.09%/78.60%, and 67.50%/62.42% on the Valence, Arousal, and Emotion indicators, respectively. The 5-fold cross validation showed values of 98.21%/97.02%, 98.59%/97.36%, and 90.56%/88.77% on the Valence, Arousal, and Emotion indicators, respectively, achieving competitive results and demonstrating the feasibility and effectiveness of the proposed model. The multi granularity attention and feature distribution calibration model was validated across experiments on the DEAP and MAHNOB-HCI datasets, with values of 82.56%/81.63%, 82.44%/88.81%, and 66.51%/65.28% for the Valence, Arousal, and Emotion metrics, respectively. The results of 5-fold cross validation were 97.48%/98.83%, 97.96%/99.26%, and 90.04%/91.89% for the Valence, Arousal, and Emotion metrics, respectively, demonstrating the feasibility and effectiveness of the proposed model over other existing methods (Chen, 2024).

In response to the significant achievements in current research on single modal emotion recognition, it is difficult to improve the accuracy of single modal emotion recognition. However, multimodal signal emotion recognition has gradually attracted the attention of researchers. In order to recognize emotions based on EEG signals more quickly and accurately, an emotion recognition algorithm combining discrete wavelet transform (DWT) and empirical mode decomposition (EMD) is proposed. To solve the problem of the inability to improve the accuracy of single modal emotion recognition, this paper establishes a dual-mode database combining EEG signals and facial microexpression signals, and increases the emotional dimension to five dimensions (excitement, happiness, neutrality, fear, sadness). Two experimental paradigms are used to collect signals from subjects, and the database contains 24 samples. Subject. The above algorithm was deployed on a self built database, and the results showed that the algorithm proposed in this paper achieved an accuracy of 46.43% in the emotion recognition five classification task. Simultaneously extracting facial micro expressions and facial feature tracking characteristics as features, combined with differential entropy features of EEG signals for feature fusion, achieved an accuracy of 52.26% in the five classification tasks. Compared to single modal features, the accuracy of five classifications corresponding to multimodal features has increased by more than 6%. In order to address the issues of high computational complexity and long testing time in neural networks, this paper uses FPGA for hardware acceleration of neural networks. Similarly, a convolutional neural network consisting of two convolutional layers was trained on the publicly available database SEED. Using 10% of the data in the database as the test set, the trained network achieved an accuracy of 78.88% in the test set, with an EEG signal testing time of 0.35 seconds per minute. At the same time, this article adopts an 8-bit quantization method to quantify the parameters of the trained network, and achieves an accuracy of 73.34% on the test set through software simulation, with an EEG testing time of 0.29 seconds per minute. In addition, the IP core of the quantized convolution kernel was constructed using C language in HLS, and the quantized neural network acceleration was implemented on the Xilinx development board ZCU104. We achieved an accuracy of 71.95% on the test set, with a testing time of 0.032 seconds per minute for EEG signals. While ensuring accuracy, we increased the testing time by 10 times, demonstrating the feasibility and effectiveness of this method (Zhang, 2022).

3 DATASET INTRODUCTION

3.1 Dataset based on Speech and Facial Expression Recognition

The datasets based on speech and facial expression recognition are concentrated on OpenDataLab, GitCode, Select Dataset, and GitHub's official website. Common datasets include the RAVDESS dataset, RML dataset, eNTERFACE'05 dataset, CH-SIMS dataset, and CMU-MOSI dataset. The above datasets are plotted in a chart as shown in Table 1.

Table 1: Speech and facial expression recognition Dataset.

Dataset name	Classification of content	The emotions contained within	Platform location
RAVDESS dataset	Audio and video	Neutral, calm, happy, sad, angry, fearful, disgusted, and surprised	OpenDataLab
RML dataset	Audio and video	Happiness, sadness, anger, fear, disgust, and surprise	GitCode
eNTERFACE’05 dataset	Audio and video	Happiness, sadness, anger, fear, disgust, and surprise	Select Dataset
CH-SIMS dataset	Audio, Text, and Emoji Images	Negative, neutral, and positive	GitHub
CMU-MOSI dataset	Video, audio, and text	Strong negativity, negativity, weak negativity, weak positivity, positivity, and strong positivity	GitHub

The datasets involved in bimodal emotion recognition based on speech and facial expressions are divided into five categories: RAVDESS dataset, RML dataset, eNTERFACE'05 dataset, CH-SIMS dataset, and CMU-MOSI dataset.

The RAVDESS dataset has a moderate total data volume, including 1440 video clips and 1440 audio clips. The video clips include synchronized facial expressions and sounds, as well as high-definition facial video recordings. The audio clips include both normal intonation and song intonation, as well as sentences with different emotions. Each speech segment lasts about 3-4 seconds, with a total of 12 male participants and 12 female participants. RAVDESS is a multimodal dataset that includes two modalities: audio and video. The emotions in the RAVDESS dataset are divided into 8 discrete emotion categories, namely: neutral, calm, happy, sad, fearful, disgusted, surprised, and angry.

The RML dataset has a moderate overall size, including 720 speech segments, each lasting about 2-4 seconds. There are 3 male participants and 3 female participants in total. The RML dataset is an unimodal emotional speech database that simulates experimental participants consciously expressing target emotions, rather than natural conversational speech. The emotions in the RML dataset are divided into six categories: happiness, anger, sadness, fear, disgust, and surprise.

The eNTERFACE’ 05 dataset is a medium-sized emotion database consisting of 1270 video clips, each approximately 5 seconds long, with a total of 43 participants from 14 different countries, predominantly male. It is a multimodal emotion recognition dataset that covers two sensory modalities: video and audio. The emotions in the eNTERFACE'05 dataset are divided into six discrete emotion categories, namely: happiness, sadness, anger, surprise, disgust, and fear.

CH-SIMS dataset, the total data volume of CH-SIMS is moderate, including 2281 video clips from 60 Chinese movies, TV dramas and variety shows. The CH-SIMS data set is a three mode data set, including text mode, audio mode and video mode. Three categories of sentiment tags are used, and the values of emotion tags are negative, neutral and positive.

The CMU-MOSI dataset, released by the Multicomp Lab at Carnegie Mellon University in the United States in 2016, is one of the most representative English benchmark datasets in the field of multimodal sentiment analysis. It includes 2199 video clips, with a total of 93 videos and 48 male participants and 41 female participants. CMU-MOSI is a standard multimodal dataset that includes three modalities: text modality, speech modality, and video modality, divided into six types of emotions: strong negative, negative, weak negative, weak positive, positive, and strong positive.

3.2 Emotion Recognition Dataset Based on Facial Images and Heart Rate Variability

The emotion recognition datasets based on facial images and heart rate variability are concentrated on Zenodo, GitCode, DEAP dataset, and the GitHub official website. Common datasets include the DEAP dataset, DECAF dataset, and MAHNOB-HCI dataset. The above datasets are plotted in a chart as shown in Table 2.

Table 2: Facial images and heart rate variability Dataset.

Dataset name	Classification of Content	The emotions contained within	Platform location
DEAP dataset	Video, audio, physiological signals, and self-evaluation feedback	Pleasure level (1-9 points) Awakening degree (1-9 points) Sense of dominance (1-9 points) Preference level (1-9 points)	DEAP dataset
DECAF dataset	Video, audio, EEG signals, and text	Anger, disgust, fear, happiness, sadness, surprise, and calmness	CMU Multimodal Lab
MAHNOB-HCI dataset	Video, audio, physiological signals, and feedback questionnaire	Pleasure level (1-9 points) Awakening degree (1-9 points) Sense of dominance (1-9 points)	Mahnob-db.eu

The DEAP dataset consists of 40 videos, each with a duration of 60 seconds and a total of 32 participants. DEAP is a multimodal physiological and emotional dataset, with main types including video stimuli, EEG signals, other physiological signals, and self-evaluation labels. The emotional presentation is scored on the following four emotional dimensions (1-9 points): pleasure, arousal, dominance, and preference.

The DECAF dataset consists of approximately hundreds of video clips, each lasting from a few seconds to a dozen seconds, with a total of 30 participants. DECAF is a typical multimodal emotion recognition dataset that includes five types of data: video, audio, EEG signals, text, and label data. DECAF typically uses seven basic emotions, consistent with Ekman's basic emotion theory: happiness, anger, sadness, fear, disgust, surprise, and calmness.

The MAHNOB-HCI dataset consists of 20 videos, each lasting approximately 35-117 seconds, with a total of 27 participants. It is a multimodal dataset, with each record containing the following types: video, audio, physiological signals, and label data. The emotional presentation is scored on three emotional dimensions (1-9 points): pleasure, arousal, and dominance.

3.3 Emotion recognition dataset based on EEG signals and facial images

The emotion recognition datasets based on EEG signals and facial images are available on the DEAP dataset, Casme, and Mahnob-db.eu official websites. Common datasets include the DEAP dataset, CASME II dataset, and MAHNOB-HCI dataset. The above datasets are plotted in a chart as shown in Table 3.

The datasets involved in emotion recognition based on EEG signals and facial images are divided into three categories: DEAP dataset, CASME II dataset, and MAHNOB-HCI dataset.

The DEAP dataset consists of 40 videos, each with a duration of 60 seconds and a total of 32 participants. DEAP is a multimodal physiological and emotional dataset, with main types including video stimuli, EEG signals, other physiological signals, and self-evaluation labels. The emotional presentation is scored on the following four emotional dimensions (1-9 points): pleasure, arousal, dominance, and preference.

Table 3: EEG signals and facial images Dataset.

Dataset name	Classification of Content	The emotions contained within	Platform location
DEAP dataset	Video, audio, physiological signals, and self-evaluation feedback	Pleasure level (1-9 points) Awakening degree (1-9 points) Sense of dominance (1-9 points) Preference level (1-9 points)	DEAP dataset
CASME II dataset	video	Happiness, disgust, repression, surprise, sadness, fear, and others	Casme
MAHNOB-HCI dataset	Video, audio, physiological signals, and feedback questionnaire	Pleasure level (1-9 points) Awakening degree (1-9 points) Sense of dominance (1-9 points)	Mahnob-db.eu

The CASME II dataset includes 255 micro expression video clips, each with a duration of 0.1~0.5 seconds and a total of 26 participants. CASME II is an unimodal dataset presented in video form, with each video recording the process of facial expression changes, which can be decoded into image sequences or video formats. CASME II uses seven basic emotions, namely happiness, disgust, inhibition, surprise, sadness, fear, and others.

The MAHNOB-HCI dataset consists of 20 videos, each lasting approximately 35-117 seconds, with a total of 27 participants. It is a multimodal dataset, with each record containing the following types: video, audio, physiological signals, and label data. The emotional presentation is scored on three emotional dimensions (1-9 points): pleasure, arousal, and dominance.

4 CURRENT LIMITATIONS AND FUTURE PROSPECTS

4.1 Current Limitations

The representativeness of the dataset is limited, and most of the participants in the datasets mentioned above are young people, with almost no participation from the elderly. This can lead to errors and inaccuracies in identifying and judging the emotions of the elderly. Emotion recognition models generally rely on deep learning, but have poor adaptability to edge devices. Previous researchers have mentioned various deep learning models, such as Transformer based on multimodal fusion, CNN-LSTM, attention mechanism, etc. However, these models have large parameter quantities, high computational resource requirements, and are not suitable for deployment in age appropriate products. Modal fusion is complex, and model robustness and practical application environment adaptability are insufficient.

4.2 Future Prospects

In the future, research on emotion recognition and emotional value enhancement for the elderly needs to construct an exclusive emotion dataset for the elderly population. Therefore, in the future, multimodal emotion data under natural interaction should be collected based on the actual living situation of the elderly, and a tagging system that conforms to the emotional characteristics of the elderly should be established, with special attention to emotional states such as loneliness, anxiety, depression, and sense of security. Then, it is necessary to explore lightweight emotion recognition models that are more suitable for intelligent terminals for the elderly. Currently, mainstream models have large computational loads and are difficult to deploy on low-power devices. In the future, by introducing lightweight neural network structures, model pruning and quantization technologies, emotion recognition systems can be quickly run on platforms such as wearable devices and companion robots, improving their practicality and portability.

5 CONCLUSIONS

This paper explores how to enhance the emotional value of the elderly through human-computer interaction. Due to the current social reality of aging population and the unmet emotional needs of the elderly, combined with various emotion recognition methods and dataset resources, targeted recognition methods and product design suggestions are proposed, and the limitations and future development directions of current research are deeply analyzed. This paper points out that the current mainstream emotion recognition models have insufficient adaptability in the elderly population, so accurately capturing the true emotional state of the elderly has become the key to the development of emotional interaction systems. This paper reflects on the shortcomings of existing research and points out that the current problems of high computational resource consumption and difficulty in modal fusion in deep models limit the application of emotion recognition systems. At the same time, there is a lack of actual user experience surveys for the elderly, making the research results challenging in practical applications. Finally, it is proposed to establish a more representative emotional dataset for the elderly, develop lightweight recognition models to adapt to low-power terminal devices, and enhance the environmental adaptability and interpretability of the models to meet the growing emotional companionship needs of the elderly.

REFERENCES

Chen, Yuan. (2024). Research on deep learning method of dual-modal emotion recognition combining EEG signals and facial images (Master's thesis). Xi'an University of Technology.

Guo, Jiaqi. (2024). Research on the design of elderly companion products based on emotional needs (Master's thesis). North China University of Technology.

Huang, Yiqi, Mao, Yuntian, Luo, Jingxin, et al. (2024). ChatEase: An AI emotional support and intelligent communication platform for elderly people living alone. Computers and Communications, (11), 38 – 42.

Wang, Bin. (2025). Healthy aging from the perspective of high-quality population development: New national conditions, new mechanisms and new paths. Journal of Yunnan University for Nationalities (Philosophy and Social Sciences), 42(2), 56 – 64.

Wu, Xiao. (2024). Research on multimodal emotion recognition based on speech, text and expression (Master's thesis). Qingdao University.

Xue, Peiyun, Dai, Shutao, Bai, Jing, et al. (2024). Dual-modal emotion recognition of speech and facial images. Journal of Electronics and Information Technology, 46(12), 4542 – 4552.

Zhang, Di. (2022). Research and implementation of multimodal emotion recognition based on EEG signals and facial micro-expressions (Master's thesis). Shandong University.

Zhang, Yuxuan, Lin, Xianxuan, Wang, Shuang, et al. (2025, May 25). Multimodal fusion emotion recognition algorithm based on facial image and heart rate variability. Journal of Information Science and Technology of Nanjing University of Science and Technology, 1 – 10.

Zhao, Zhongjie. (2010). Research on elderly care issues of single-child families in urban areas of Beijing (Doctoral dissertation). Minzu University of China.

Zhu, Qingyang. (2021). Multimodal emotion recognition based on EEG, peripheral physiological signals and facial expression (Doctoral thesis). Nanjing University of Posts and Telecommunications.