Proceedings of CONF-MLA 2026 Symposium: Explainable Computing, Modeling & Data Science in Complex Systems
DOI: 10.54254/2755-2721/2026.GF35912

Application of Emotion Recognition-Based Neural Networks in Human-Computer Interaction

Zihao Wang

Attached Middle School to Jiangxi Normal University, Nanchang, China
Wangzihao090603@outlook.com

Abstract. This paper primarily investigates the application of neural network-based emotion recognition methods in human-computer interaction, with a focus on convolutional neural networks in facial emotion recognition, vision transformers, and multimodal emotion recognition research approaches. This paper first analyzes the basic process and principle of ResNet-50 in facial image feature extraction, explaining that it can extract local facial features through convolution operations and residual connections, and use Softmax to complete emotion classification. Secondly, this paper also introduces Vision Transformer, which models global features through image patch partitioning and a self-attention mechanism, and analyzes its advantages in capturing the correlation between different facial regions. Subsequently, this paper further discusses the application of multimodal emotion recognition in medical scenarios, demonstrating its ability to integrate various information such as facial images, speech, text, and physiological signals. It is pointed out that different neural network models are suitable for different human-computer interaction scenarios, and a single model is difficult to adapt to complex environments. Future emotion recognition technology should further develop towards model lightweighting and multimodal fusion.

Keywords: Neural Network, Emotion Recognition, Human-Computer Interaction

1. Introduction

Emotion recognition plays a crucial role in understanding human social communication and life behaviors. Automated recognition of emotion has been widely applied in various fields such as education, healthcare, network security, and more.

Facial expression recognition (FER), as a primary method of emotion recognition, mainly uses human facial recognition to identify dynamic changes in emotions and infer emotional states. In mainstream FER research, the core task objective is typically defined as the precise classification and recognition of seven basic facial emotions, namely happy, sad, fear, anger, surprise, disgust, and neutral [1, 2].

In the process of human-computer interaction, emotion recognition is often carried out by the camera to capture the user's image, and then the model is used to analyze the expression, thus changing the interaction logic. This means that model selection will affect the neural network's judgment of emotion. For the use of methods and models, the mainstream method is to use a convolutional neural network (CNN). Because of its computational efficiency and feature extraction

DOI: 10.54254/2755-2721/2026.GF35912

ability, it is widely used in image feature extraction. It can use VGGNet or ResNet-50 and other models for reasoning [3]. Talele et al. tested the performance of ResNet-50 in the facial emotion recognition task, and the results showed that the accuracy of ResNet-50 reached 85.5% [4]. Transformers use the attention mechanism, which can add weights to key parts (such as eyes and lips above) on the basis of local feature extraction, so as to improve the recognition results of the model [5]. The model can use Vit-b/16/s, ViT-B/16/SG, and ViT-B/16/SAM, etc. Research based on Swin Transformer proposed the SwinFer model and conducted experiments on FER2013, CK+, AffectNet, and other datasets. The experimental results show that the recognition accuracy of this model on the FER2013 dataset is 71.11%, while it reaches 100% on the CK+ dataset in a laboratory environment [6].

This paper focuses on the application of neural network emotion recognition methods in human-computer interaction, mainly introducing the ResNet-50 model based on CNN, the ViT model based on Transformer, and a multimodal emotion recognition method. This paper will analyze the basic processes, technical characteristics, and usage scenarios of these methods, and discuss their advantages and disadvantages in human-computer interaction applications.

2. Neural network

2.1. Recognition using ResNet-50 of convolutional neural network (CNN)

2.1.1. Main processes

When using ResNet-50, the main process is the same as that of most CNN models. First, the image obtained by the camera is preprocessed, then the ResNet-50 model is given, and then Softmax is used to classify the score results of the model into a probability distribution. After obtaining the emotion category, the interactive logic is adjusted according to the requirements of different scenes.

The process of image expression is, facial image → preprocessing → ResNet-50 → Softmax → emotion type → adjustment interaction

2.1.2. Principle

ResNet-50 is a deep network model developed on the basis of the revolutionary neural network (CNN). The basic idea of CNN is to gradually extract image features through multi-layer networks. Generally speaking, CNN mainly includes a convolution layer, a pooling layer, and a full connection layer. The convolution layer slides on the image through the convolution kernel to extract local information such as edges, texture, contour, etc. The pooling layer compresses the feature map to reduce the amount of data and retain obvious features; The full connection layer will integrate the previously extracted features for final classification. For facial expression recognition tasks, CNNs can extract regional features such as eyes, lips, eyebrows, and facial muscle changes from face images, so it is more suitable for processing image data with spatial structure.

In an ordinary CNN, with the deepening of the network layers, the model can theoretically learn more complex image features, but the training process may also have problems such as gradient disappearance or model degradation. ResNet-50 is different from an ordinary CNN in that it adds Skip Connections, which enables the network to maintain a more stable training effect with deeper layers.

$$H(x) = F(x) + x \tag{1}$$

DOI: 10.54254/2755-2721/2026.GF35912

Wherein, represents the input characteristics, represents the residual mapping that the network needs to learn, and represents the final output. For emotion recognition, this method helps the model to capture the subtle changes in local areas of the face, such as mouth corners rising, eyebrows tightening, or eye changes. Therefore, ResNet-50 has strong expressive ability in processing complex facial image features [7].

After feature extraction, the model usually uses the Softmax function to convert the scores of different emotion categories into a probability distribution. The model will calculate the possibility that the image belongs to happy, sad, angry, surprised, and other categories, and take the category with the highest probability as the final recognition result. In general, CNN provides the basis for image local feature extraction, and ResNet-50 further improves the training stability and feature expression ability of the deep network through residual connection, so it can be used for facial expression recognition tasks.

2.1.3. Application scenarios

Since models such as CNN and ResNet-50 can extract local feature changes of the face, they can be applied to scenes that need to analyze the state of the face, such as intelligent education [8]. In the classroom or online learning process, students' emotional changes can reflect their learning state to a certain extent, such as confusion, concentration, fatigue, etc. The CNN model has been used for online students' emotion recognition in previous studies. The facial images of students in the learning process are recorded through the camera, and their emotional states are analyzed at regular intervals. The CNN model of this study consists of a convolution layer, a pooling layer, a full connection layer, and a Softmax output layer, and finally outputs the classification of emotions. The recognition accuracy in the report reaches 95%, and the precision is 89% [9].

This result shows that the emotion recognition method based on vision has certain application value in intelligent education, which can adjust the teaching rhythm for teachers or provide learning feedback reference in an online education platform. But at the same time, students' emotions are somewhat subjective, and their expressions are easily affected by the environment, camera perspective, and personal habits. Therefore, the system is more suitable as an auxiliary judgment tool, rather than making accurate judgments about students' class status.

2.2. Use the transformer's ViT (vision transformer) for identification

2.2.1. Main processes

The transformer highly relies on the attention mechanism to model global relationships. It needs to preprocess images or videos and divide them into fixed-size patches for processing. Then, like CNN, it converts softmax into probability, so as to identify emotion types and apply them to human-computer interaction [10]. The expression process is: facial image → preprocessing → Patch → Transformer → Softmax → emotion type → adjustment interaction.

2.2.2. Principle

Transformers process discrete token sequences, so it needs to convert the image into fixed-size patches, and then convert each patch into vectors through calculation. Since all vectors belong to one dimension, they can be put into the transformer model. Since these image blocks only contain

DOI: 10.54254/2755-2721/2026.GF35912

content information, the model also needs to add position coding to enable the model to recognize the position relationship of different patches in the image.

The core of ViT is his Self Attention mechanism. In facial emotion recognition, different facial regions often have connections. For example, changes in the eyes and corners of the mouth often do not occur alone, but jointly reflect an emotional state. ViT can focus on such key parts, thus reducing the impact of non-key parts on emotion recognition. In the ViT proposed by Tian et al., the combination of local features and global context is used, which shows that FER tasks need to pay attention to both local facial details and the overall relationship of the face [11]. The basic expression is,

$$Attention\,(Q,K,V) = \ Softmax\left(\frac{Q\,K^T}{\sqrt{d_k}}\right)V \tag{2}$$

Wherein, represents the word vector that needs to be considered at present, represents the similarity used to match the query vector, represents the actual information of each word, and represents the scaling factor.

2.2.3. Application scenarios

The ViT model can be used as the technical basis for emotion recognition in virtual character interaction. In such scenarios, the system not only needs to judge the current user's state through obvious body movements, but also needs to accurately identify the current user's real emotional state through analyzing facial expressions. As mentioned earlier, transformer models such as Swin Fer can already be used for facial emotion recognition tasks, and the recognition accuracy on the FER2013 dataset is 71.11%. At the same time, ViT can explain that ViT-type models are also suitable for dealing with the relationship between different facial regions by combining local features and global context information. Therefore, in the interaction of virtual roles, such models can help the system judge whether users are confused, dissatisfied, or happy, thus changing the response mode of virtual roles. However, due to such factors as lighting, facial occlusion, or angle change in the real interactive scene, the actual effect of the Vit model still needs to be further verified with the real dataset.

2.3. Multimodal emotion recognition

2.3.1. Basic modes of multimodal emotion recognition

In a real human-computer interaction environment, it is often impossible to accurately identify the user's real emotion from a single facial expression, and it is also necessary to comprehensively judge the user's behavior, tone, and other factors. At this time, it is necessary to use multimodal emotion recognition. Multimodal emotion recognition is to put several different types of information together to judge the user's state.

In common multimodal emotion recognition models, facial images, speech texts and physiological signals are all important data that can be used. Facial images can feed back regional changes of organs such as eyebrows and mouth, voice text can reflect the current user's tone and speed, and factors such as heart rate can reflect the user's tension. The comprehensive use of these data can make a more comprehensive and accurate analysis of emotion recognition [12].

In specific processing, the model usually extracts the features of different modes respectively. For example, images can be processed by CNN and ResNet first, voice by Transformer, text by BERT

DOI: 10.54254/2755-2721/2026.GF35912

and other models. After that, the system will integrate these features and finally output emotion categories. In this case, when the single information is not clear enough, other modes can play a complementary role. However, multimodal methods will also make data acquisition and model training more complex, especially the alignment between different modes, which needs to be further solved in real human-computer interaction scenarios.

2.3.2. Application and limitation of multimodal emotion recognition

Multimodal emotion recognition has been used in medical service and other fields. Compared with single mode, multimodal method can combine multiple information, so it has stronger stability in complex environment.

In the medical service scenario, users' emotions will directly affect the priority of complaint handling and the way of communication. Some researches have built a voice and text dual-mode emotion recognition system for medical complaint processing, in which voice extracts acoustic features such as intonation, speech speed and volume through WavLM, and text uses BERT+Bi LSTM to analyze users' emotional words and context information. The system can divide users' emotions into three categories: neutral, mild dissatisfaction, and strong anger, and classify them according to different emotions to carry out different appeasement and treatment countermeasures. The experimental results show that the accuracy rate is 78.6% when using voice mode alone and 81.9% when using text mode alone, while the accuracy rate of multimodal model after voice and text fusion reaches 89.4% and F1 score reaches 87.1% [13]. However, multimodal emotion recognition also has some practical limitations. First, multimodal systems need to collect a variety of data, such as voice, text, facial images and physiological signals, which will increase equipment costs and data processing difficulties. And because the data sampling frequency of different modes is different, the time may not be synchronized between voice, video and physiological signals, so additional timing alignment is required [14].

3. Discussion and analysis

From the previous analysis of ResNet-50, Vit and multimodal emotion recognition methods, it can be seen that different models are not better related to each other, but different use scenarios and tasks. ResNet-50 is a model developed on the basis of CNN. Its advantage is that it can extract features of local facial areas, such as changes in the positions of eyes, lips and eyebrows. Therefore, ResNet-50 has certain advantages in tasks with clear expression and need to analyze local details [15]. The ViT model emphasizes the relationship between different image regions. It can establish a global relationship through the self attention mechanism, so it is more suitable for scenes that need to comprehensively judge the overall expression state. The hybrid Vision Transformer model proposed by Li et al. extracts local and global features in facial expressions through multi-scale feature fusion, which also shows that facial expression recognition cannot only focus on a single region, but also needs to analyze the overall relationship between different facial regions [16].

The advantage of multimodal emotion recognition is that it can combine more information. When the facial image is blocked, or the user's expression is not obvious, voice, text and physiological signals can play a complementary role, so it is usually more stable than a single mode in the real scene. However, multimodal methods also have new problems. First, the acquisition cost of multimodal data is higher, requiring cameras, microphones and other equipment, and sometimes even high-precision medical devices. Secondly, the time alignment between different modes is a problem. For example, the changes of voice emotion and facial emotion may not be completely

DOI: 10.54254/2755-2721/2026.GF35912

synchronized. Relevant reviews also point out that modal alignment, missing modes and model complexity are still the main problems in the application of multimodal emotion recognition [17].

4. Conclusion

This paper focuses on the application of emotion recognition in human-computer interaction, and mainly analyzes ResNet-50 model based on CNN, Vit model based on Vision Transformer, and multimodal emotion recognition methods. In the ResNet-50 model based on CNN, this paper analyzes the process of extracting facial local features through convolution operation, and using residual connection to alleviate the problem of deep network training. This method is suitable for recognizing emotional changes in eyes, lips and other areas, so it has certain value in intelligent education and other scenes that need to analyze the facial state. In the Vit model based on Transformer, this paper introduces the steps of image patch division, self attention mechanism and Softmax classification. Unlike CNN, which pays more attention to local features, Vit is able to establish an overall relationship between different facial regions, so it has more advantages in the task of understanding global expressions. In the part of multimodal emotion recognition, this paper analyzes the fusion methods of facial image, voice, text and physiological signals. It is found that multimodal method can make up for the shortcomings of single face recognition in complex environments, but it will also bring problems such as data alignment, model training cost and actual deployment difficulty.

Therefore, the model selection of emotion recognition should be based on the specific human-computer interaction scenarios. Future related research can further focus on the lightweight of the model, the efficiency of multimodal fusion and the stability in the real environment, so that emotion recognition technology can be better applied to education, medical and other practical scenarios.

References

[1] Chaudhari, A., Bhatt, C., Krishna, A., & Mazzeo, P. L. (2022). ViTFER: Facial emotion recognition with vision transformers. *Applied System Innovation, 5*(4), Article 80.

[2] Cohn, J. F., & Zlochower, A. (1995). A computerized analysis of facial expression: Feasibility of automated discrimination. *American Psychological Society, 2*(6).

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84–90.

[4] Talele, M., & Jain, R. (2025). A comparative analysis of CNNs and ResNet50 for facial emotion recognition. *Engineering, Technology & Applied Science Research, 15*(2), 20693–20701.

[5] Li, J., Jin, K., Zhou, D., Kubota, N., & Ju, Z. (2020). Attention mechanism-based CNN for facial expression recognition. *Neurocomputing, 411*, 340–350.

[6] Bie, M., Xu, H., Gao, Y., Song, K., & Che, X. (2024). Swin-FER: Swin transformer for facial expression recognition. *Applied Sciences, 14*(14), Article 6125.

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

[8] Xiao, X. (2025). Research and design of learning emotional engagement recognition based on improved ResNet algorithm. *Software, 46*(6), 31–36.

[9] Nair, R. R., & Babu, T. (2023). Enhancing student engagement in online learning through facial expression analysis and complex emotion recognition using deep learning (arXiv Preprint arXiv:2311.10343). *arXiv*.

[10] Liu, Y., Sun, G., Qiu, Y., Zhang, L., Chhatkuli, A., & Van Gool, L. (2021). Transformer in convolutional neural networks (arXiv Preprint arXiv:2106.03180). *arXiv*.

[11] Tian, Y., Zhu, J., Yao, H., & Chen, D. (2024). Facial expression recognition based on vision transformer with hybrid local attention. *Applied Sciences, 14*(15), Article 6471.

[12] Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M. A., & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends, 2*(1), 73–79.

DOI: 10.54254/2755-2721/2026.GF35912

[13] Zhang, H. (2026). Research on multimodal emotion recognition and coping in medical scenarios. *Fujian Computer, 42*(2), 16–21.
[14] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443.
[15] Elsheikh, R. A., Mohamed, M. A., Abou-Taleb, A. M., & Ata, M. M. (2024). Improved facial emotion recognition model based on a novel deep convolutional structure. *Scientific Reports, 14*(1), Article 29050.
[16] Li, N., Huang, Y., Wang, Z., Fan, Z., Li, X., & Xiao, Z. (2024). Enhanced hybrid vision transformer with multi-scale feature integration and patch dropping for facial expression recognition. *Sensors, 24*(13), Article 4153.
[17] Wu, Y., Mi, Q., & Gao, T. (2025). A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions. *Biomimetics, 10*(7), Article 418.