Deep Learning Based on Facial Expression Recognition from Images to Videos

Rui Deng*

School of Software, Henan University, Henan, China

Abstract. Facial expressions, as a vital conduit for human emotional expression, are among the most observable features of machines in the field of computer vision. Consequently, facial expression recognition holds broad potential for applications in artificial intelligence and health monitoring, among others. Given the diversity and complexity of expressions, the development of efficient and accurate models for expression recognition is of significant importance. This paper systematically reviews the foundational knowledge and related research in facial expression recognition, analyzing the application of current primary models in expression recognition. Employing a combination of literature review and experimental analysis, this study evaluates existing facial expression recognition algorithms. Special attention is given to advanced models based on Convolutional Neural Networks (CNNs), with a detailed comparison of their architectures and characteristics, analyzing their performance under various conditions. The paper concludes with a summary of the latest advancements in the field of facial expression recognition and proposes potential directions for future research.

1 Introduction

Facial expression recognition has seen rapid development in recent years. In 1971, psychologist Paul Ekman classified expressions into six basic categories: happiness, sadness, anger, disgust, surprise, and fear [1]. Building on this foundation, Ekman and his colleagues introduced the concept of micro-expressions [2].

In recent years, numerous deep learning-based models for facial expression recognition have been proposed and utilized. In 1989, Yann LeCun first introduced Convolutional Neural Networks (CNNs) for recognizing facial expressions, sparking a wave of innovation in network structures [3]. In 2015, Christian Szegedy and his team proposed the Inception network, significantly enhancing the recognition rate and computational efficiency of networks [4]. In 2016, Kaiming He and his colleagues introduced the ResNet network, which notably improved training outcomes and accuracy, accelerating the development of convolutional neural networks [5].

Expressions can be categorized into macro-expressions and micro-expressions based on the intensity and duration of facial expressions [2]. This paper divides the recognition methods into two parts: traditional macro-expression recognition and micro-expression

* Corresponding author: naoh@henu.edu.cn

recognition, detailing recent research progress in both areas. It systematically summarizes and analyzes existing research in the field of facial expression recognition and compares different facial expression recognition algorithms, analyzing the strengths and applicable scenarios of algorithmic models.

2 Deep expression recognition

2.1 Neural networks

Neural Networks (NNs) are machine learning techniques inspired by biological neurons, simulating the human brain's learning process. Neural Networks create neuron models based on the principles of neurons [6]. Computers achieve learning and classification by increasing the number of neurons, commonly used for image recognition tasks. In 1986, Rumelhart and his team of researchers developed the Back-Propagation Algorithm (BPA), an algorithm for gradually adjusting the weights of neural networks [7]. Unlike forward propagation, the back-propagation phase starts from the output layer and retraces to the input layer, calculating the error gradients of each layer, thus adjusting weights and biases to reduce the overall network's prediction error.

2.2 Convolutional neural networks

Convolutional Neural Networks (CNNs) are deep learning models inspired by the human visual processing system. Since deep learning cannot discern the positional changes of the same object, CNNs employ convolution operations in convolutional layers to capture local features of objects unaffected by their positional changes [3]. CNNs typically consist of multiple convolutional and pooling layers overlapped to capture more detailed features and represent more complex patterns. Convolutional layers are composed of multiple filters (or kernels), where the principle involves a kernel (filter) moving across the selected values, multiplying and summing them to extract corresponding features [8]. Multiple kernels form a convolutional layer, repeatedly extracting features from feature images. Pooling is similar to convolution, but it differs in that it outputs the maximum value (Max Pooling) or sometimes the average value (Mean Pooling), serving to reduce computational complexity without losing features [9].

3 Datasets and preprocessing

3.1 Dataset selection

Datasets are crucial for model training, and facial expression recognition datasets, they need to include as many people and complex environments as possible. This significantly impacts the optimization and innovation of deep expression recognition models and can even influence the direction of model improvements. This paper will introduce several publicly available and outstanding datasets, with Table 1 being a summary of the datasets.

Table 1. Overview of databases

No.	Database	Sample	Condition
1	RAF-DB	1608 images	Lab
2	AffectNet	450,000 images	Internet
3	CK+	593 images sequences	Lab
4	FER2013	35,887 images	Lab

Continue Table 1.

5	ExpW	91,793 images	Internet
6	BU-3DFE	2,500 3Dimages	Lab
7	BU-4DFE	606 3D sequences	Lab
8	CAMSE II	60,600 frame models	Lab
9	SMIC	164 videos	N/A
10	JAFFE	213 images	Lab
11	4DFAB	1.8 million 3D	Lab

3.2 Data preprocessing

Features unrelated to facial expressions, such as image backgrounds and occlusions, can affect the accuracy of expression recognition. Therefore, preprocessing images for expression recognition is a key step in improving recognition accuracy and algorithm robustness. Before training neural networks, it is generally necessary to perform face alignment, normalization, and data augmentation on the transmitted facial features. This not only removes redundant information and reduces its interference with recognition accuracy but also enhances model robustness, increases data diversity, prevents model overfitting, and allows the model to more accurately recognize facial features.

4 Advanced methods

4.1 Macro-expression recognition

4.1.1 Improvement of ResNet50 algorithm with attention module integration

To address the issue of numerical stability degradation in deep convolutional networks and the need for simultaneous recognition of multiple regions, an Improved Residual Network with attention (IRNet) is proposed [10].

This model is based on ResNet and replaces a part of the convolutional layers in ResNet50 with a sub-attention module. By paralleling multiple attention heads, it obtains multiple sets of features with attention weights. The resulting features are then integrated and classified, using batch normalization to improve network training speed and stability, as shown in Figure 1:

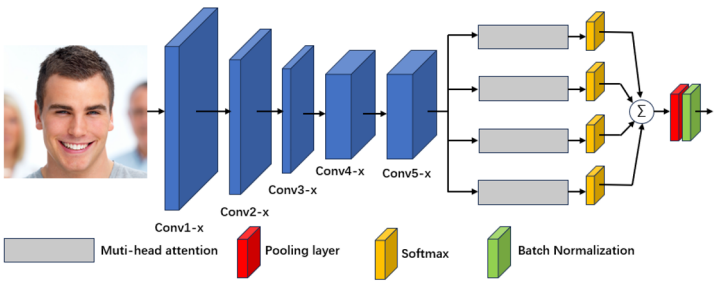

Fig.1. IRNet Model [10]

Self-attention mechanisms adaptively adjust the global input items, determining the weight of each input item to focus on different features and ignore others. Due to its characteristic of not changing the size and dimension of input features, it can be flexibly integrated into other network structures. Therefore, this module can replace some

convolutional layers in the residual network. Here, conv2 to conv5 replace the 3x3 convolutional layers in the second to fifth parts.

The multi-head cross-attention module is composed of multiple parallel cross-attention structures, each of which includes a channel attention module and a spatial attention module. The cross-attention structure is shown in Figure 2:

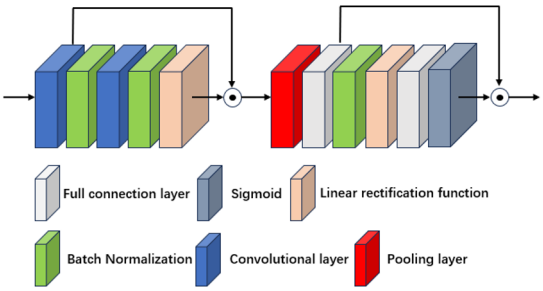

Fig. 2. Cross-Attention Structure [10]

The input feature map first passes through the spatial attention module to reallocate weights for spatial features, then through a channel attention module to reallocate weights for different channels, ultimately outputting a feature map with the highest relevance to the task. The role of the BN layer is to accelerate network convergence and avoid divergence caused by training with high learning rates [11].

Experimental results show that it achieves a recognition rate of 90.40% in the RAF-DB dataset, significantly improving the accuracy compared to other models.

4.1.2 Deformable convolution combined with attention mechanism

Traditional CNNs with a fixed framework use a uniform-size receptive field, making it difficult to recognize multiple facial regions simultaneously. Therefore, the concept of deformable convolution is proposed [12]. By allowing the sampling points of the convolution kernel to shift spatially, the sampling area is expanded, and the range of capturing spatial features is improved.

In deformable convolution, an additional convolutional kernel is added to learn the offset quantity, which determines the size of the offset field. Based on traditional convolution, it includes an offset quantity, as shown in formula (1):

$$y(A_0) = \sum_{A_n} w(A_n) \cdot x(A_0 + A_n + \Delta A_n) \quad (1)$$

A_0 indicates the center position of the convolution kernel, and A_n represents all pixel values in the convolution kernel. However, during the calculation, it is found that the sum of these three values is not an integer, meaning the sampling point position is not an integer and cannot correspond to the pixel points in the image, so bilinear interpolation is required to process the offset points.

In this way, the values of the existing sampling points can be used to find integer sampling point positions with the same linear relationship.

Using CBAM mixed attention to adjust weights allows the model to focus on specific areas. The feature vector first passes through CAM to obtain channel attention weights, and then through SAM to obtain spatial attention weights. The CAM structure is shown in Figure 3:

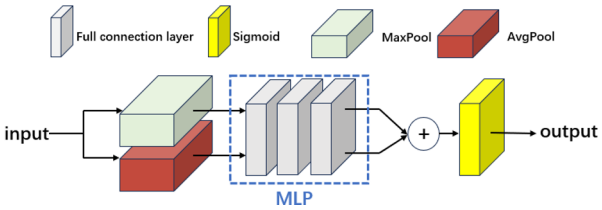

Fig. 3. CAM Structure [11]

The input features are separately subjected to average pooling and global max pooling to generate two feature vectors, which are then processed by a multi-layer perceptron (MLP), followed by the sigmoid activation function to obtain the weighted output feature vector after multiplying with the original data input to CAM.

SAM is used to obtain spatial attention weights, performing weighted processing on the feature vectors weighted by channels in the spatial dimension. Its structure is shown in Figure 4:

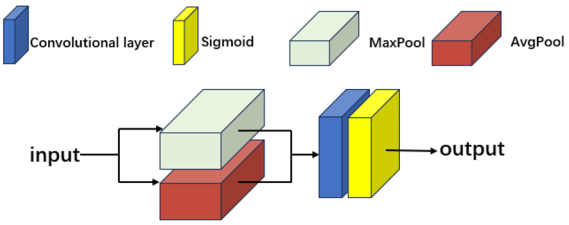

Fig. 4. SAM Structure [11]

By combining deformable convolution and CBAM attention mechanisms, not only is the range of capturing spatial features improved, but there is also a breakthrough in identifying important areas in the input image.

Experimental results show that it achieves a recognition rate of 73.22% on the FER2013 dataset, with a high recognition accuracy.

4.2 Micro-expression recognition

4.2.1 Large kernel convolutional neural networks

Traditional CNNs can only learn from complete images and struggle to accurately capture the details of facial micro-expressions. Traditional convolutions using small kernels overlook global image features during feature extraction, thus affecting the accuracy of micro-expression recognition.

To overcome the shortcomings of small kernels, a large kernel convolutional neural network for micro-expression recognition is proposed here. LKCNN expands the model's receptive field, allowing it to capture data within a larger range. It uses Inception depthwise separable convolution to balance computational costs; it decomposes the expensive deep convolution into three branches with small kernel sizes and an identity mapping branch [13].

In terms of data preprocessing, three types of optical flow are used as input data, aiming to detect facial smile movements by extracting features between vertex frames and starting frames. The input images are divided into three graphics, each passing through a framework as shown in Figure 5.

The features of the three graphs are all obtained from OF (optical flow vectors), with each path calculated by the block segmentation module LK for each block's image features. During the LK process, each image is evenly divided into square blocks of size S×S.

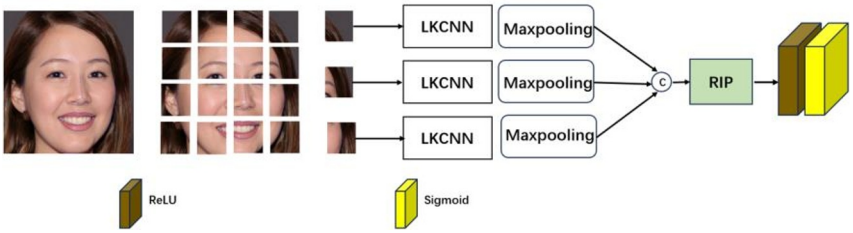

Fig. 5. LKCNN [13]

Then, each block undergoes four Inception depth convolution operations and max pooling operations. After this process, the output of each image feature block is concatenated according to the channel dimension, resulting in the concatenated block features Fk. The concatenated features are then passed through the Relative Position Attention Module (RIP), activated by the ReLU function, and finally activated by the Sigmoid function to obtain attention weights. The original values are then multiplied by these weights to obtain the block feature values. The three-block features obtained from each image through the HC module are finally concatenated to form a feature set. Its data for micro-expression three-category comprehensive indicators Acc, F1-score are 86.01%, and 84.62% respectively, showing a high recognition rate overall.

4.2.2 Multi-region features and feature fusion for micro-expression recognition

In micro-expression recognition, the commonly used Long Short-Term Memory (LSTM) network cannot effectively extract features from multiple related facial local regions in micro-expression video sequences, leading to information loss [14].

To address the above issue, based on the cascaded network model of ResNet34 and LSTM, a Multi-Region Feature Extraction Module (MFEM) is introduced, and a residual network is used to extract spatial information from micro-expression video sequences, paired with LSTM to model and extract micro-expression features to represent the temporal sequence relationship of micro-expression video sequences [15]. Its structure is shown in Figure 6:

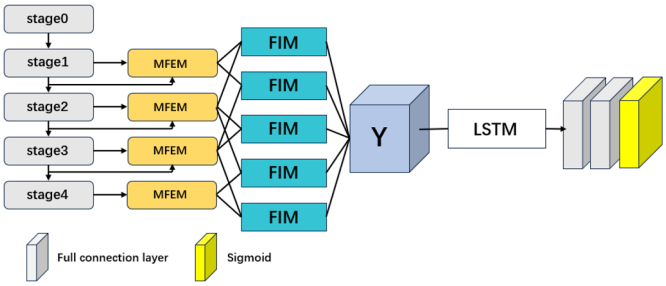

Fig. 6. Multi-Region Features and Feature Fusion [15]

The model's main network ResNet34 has five main stages, and except for the first stage, MFEM multi-region feature extraction modules are introduced in the other four stages, while sending data from the previous stage to the next stage for learning more advanced features.

Some AU occurrences in micro-expressions are in close and overlapping areas, and attention mechanisms can easily confuse, leading to judgment errors. Therefore, MFEM's multi-region feature extraction module uses attention mechanisms to focus on multiple local areas where micro-expressions occur [16]. As shown in Figure 7:

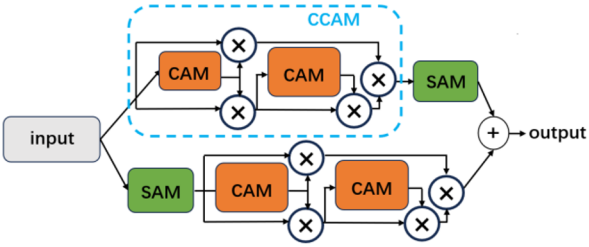

Fig. 7. MFEM Structure [15]

The module uses a parallel architecture to prevent interference that might occur in serial architectures. It uses SAM to obtain spatial attention and two CAMs to better obtain channel attention.

The Multi-Layer Feature Fusion Module (MFFM) module is used to fuse features from various levels. It learns interactively from features from different levels through FIM to produce more comprehensive and expressive feature representations. Since each frame image in micro-expression video sequences has dependencies on adjacent frames, the extracted features F are directly converted into an independent vector and input into the bidirectional LSTM module. Its data for micro-expression three-category comprehensive indicators Acc, F1-score are 84.96%, and 85.50% respectively, showing a high recognition rate overall.

5 Discussion and Analysis

The advantages of the ResNet50 algorithm improvement with integrated attention modules are quite evident. IRNet improves the stability and feature learning capabilities of traditional convolutional neural networks by combining attention mechanisms and residual networks. The use of multi-head cross-attention modules enhances the network's ability to handle complex inputs, and the design of attention modules is also very flexible, with good scalability, showing significant advantages in multi-region recognition tasks. However, although multi-head cross-attention modules can improve network performance, the parallel structure increases computational load and memory usage. Excessive parameters or improper configurations may lead to slower model convergence, and the complexity of the network also requires larger-scale training data for model training.

Deformable networks combine deformable convolution and CBAM, using deformable convolution to address complex changes in facial expressions and CBAM to weight adjust the feature map's channels and spatial dimensions. This not only improves the flexibility of the receptive field to handle complex facial deformations but also enhances the model's focus on key areas, improving its adaptability to different individuals and expressions. However, the additional parameters in deformable convolution increase the model's computational complexity, and overfitting issues may arise in smaller facial expression datasets. Training such complex models also has certain hardware requirements.

The ResNet50 algorithm with integrated attention modules emphasizes the learning ability of multi-region features and effectively solves the numerical stability issues of deep networks, making it suitable for multi-region facial expression recognition tasks. Deformable networks, by dynamically adjusting the receptive field of the convolution kernel, enhance sensitivity to local facial expression changes and combine hybrid attention mechanisms to further improve the model's detail modeling capabilities, making them suitable for complex facial expression changes. The two have different focuses; the ResNet50 algorithm with integrated attention modules is suitable for multi-region feature learning and scenarios requiring high network stability, while deformable convolution networks perform better in handling facial details and dynamic changes.

The advantages of large kernel convolutional neural networks are quite evident. First, by expanding the receptive field with large kernels, they can better capture global image features while using Inception depthwise separable convolution to reduce computational costs and enhance recognition capabilities. However, although LKCNN reduces computational costs to some extent with Inception depthwise separable convolution, the use of large kernels and multi-branch structures still leads to increased computational overhead, especially when processing high-resolution images, which may require higher hardware resources.

The multi-region feature and feature fusion model focuses on multiple aspects of micro-expressions through MFEM. Its parallel architecture helps prevent interference that might occur during serial processing, and when combined with LSTM for temporal feature modeling, it can capture spatial features while considering the temporal sequence information of video sequences. However, the overall complexity of the model is high, and it has high requirements for data quality. On data with low video resolution or inconspicuous micro-expressions, it may be difficult to fully utilize its feature extraction and recognition capabilities. In scenarios with limited training data, attention mechanisms may lead to overfitting, and the use of multiple modules also increases the overall number of model parameters, making the parameter-tuning process very complex.

Both models use attention mechanisms, but the focuses of the two attention mechanisms are different. CBAM focuses on the attention to micro-expressions, while MFEM focuses on the model's stronger focusing ability on specific facial areas. LKCNN is suitable for static micro-expression recognition tasks, enhancing the capture of global features through the combination of large kernel convolution and depthwise separable convolution. The two have different focuses in design; the former is more suitable for lightweight static tasks, while the latter is suitable for applications requiring high recognition accuracy and temporal modeling.

6 Conclusion

This paper reviews the development process of expression recognition, introduces datasets and data processing methods for facial recognition, and focuses on several of the latest research achievements in expression recognition, analyzing their strengths, weaknesses, and application scenarios. The development of expression recognition technology has made significant progress, with improvements in model architecture, optimization of data processing methods, and the introduction of more flexible feature extraction mechanisms, leading to greater accuracy and broader application of expression recognition.

In the future, with further development of computational power and algorithms, the field of expression recognition still has broad development prospects. Firstly, how to further improve the real-time performance and lightweight nature of models will become an important research direction for application promotion. Secondly, enhancing the model's robustness to complex environments such as occlusions and lighting changes is also a major challenge. On this basis, exploring the integration of multimodal data (such as speech, posture, etc.) to improve the accuracy and diversity of expression recognition is expected to push this field into a more mature application stage.

References

1. P. Ekman, W. V. Friesen, Constants across cultures in the face and emotion, Journal of Personality and Social Psychology 17(2), 124–129 (1971).
2. P. Ekman, W. V. Friesen, Nonverbal leakage and clues to deception, Psychiatry 32(1), 88–106 (1969).

3. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Computation 1(4), 541–551 (1989).
4. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1–9 (2015).
5. K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016).
6. D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by backpropagating errors, Nature 323(6088), 533–536 (1986).
7. W. S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, The Bulletin of Mathematical Biophysics 5(4), 115–133 (1943).
8. K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biological Cybernetics 36(4), 193–202 (1980).
9. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 25, 1097–1105 (2012).
10. M. Qiu, Research on facial expression recognition algorithms based on convolutional neural networks (Master's thesis), Nanchang University (2024).
11. H. Ren, Research on facial expression recognition methods based on improved convolutional neural networks (Master's thesis), Guizhou Normal University (2024).
12. S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, In Proceedings of the 32nd International Conference on Machine Learning, 448–456 (2015).
13. Z. Cao, Research and application of micro-expression recognition algorithms based on deep learning (Master's thesis), Qilu University of Technology (2024).
14. C. Cao, D. Zhang, Micro-expression recognition combining multi-region features and feature fusion, Small and Micro Computer Systems, 1–9 (2024).
15. S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9(8), 1735–1780 (1997).
16. T. Y. Hou, X. H. Wu, A multiscale finite element method for elliptic problems in composite materials and porous media, Journal of Computational Physics 134(1), 169–189 (1997).