A Facial Expression Recognition Method Based on Improved VGG19 Model

Lihua Bi[1], Shenbo Tang[2], Canlin Li[3]*
School of Software Engineering, Zhengzhou University of Light Industry, Zhengzhou, China[1]
School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou, China[2, 3]

***Abstract*—With the increasing demand for human-computer interaction and the development of emotional computing technology, facial expression recognition has become a major focus in research. In this paper, an improved VGG19 network model is proposed by involving enhancement strategies, and the facial expression recognition process with the improved VGG19 model is provided. We validated the model on FER2013 and CK+ datasets and conducted comparative experiments on facial expression recognition accuracy among the improved VGG19 and other classic models, including the original VGG19. Instance tests were also performed, using probability histograms to reflect the effectiveness of expression recognition. These experiments and tests demonstrate the superiority, as well as the applicability and stability of the improved VGG19 model on facial expression recognition.**

***Keywords*—Facial expression recognition; deep learning; VGG19 model**

I. Introduction

Emotional recognition is a dynamic process aimed at understanding a person's emotional state, meaning the feelings corresponding to each individual's behaviour vary [1]. Generally, people express their emotions in different ways. To ensure meaningful communication, accurate interpretation of these emotions is essential [2]. Facial expressions are a primary means by which people convey emotions [3-6]. Mehrabian [7] observed that 7% of knowledge is transmitted between people through writing [8], 38% through voice, and 55% through facial expressions. Ekman and Friesen published the Facial Action Coding System (FACS) in 1978, which describes the seven main facial expressions people express without language, such as fear, detachment, surprise, disgust, good fortune, sincerity, and neutrality. This system is considered the threshold for Facial Expression Recognition (FER) [9].

Various applications involve understanding human emotions through facial expressions, including human-computer interaction, robotics, and healthcare [10-12]. However, emotion recognition in our daily lives is important for social contact, as emotions play a significant role in determining human behaviour [13].

In the field of school education, the emotional state of elementary school students can be immediately interpreted through Facial Expression Recognition (FER). It allows teachers to recognize their students' academic interests, including appropriate teaching methods to improve teaching efficiency [14]. Monitoring the analysis process of human posture over a period is crucial, for example, in museums where visitors can reflect on and explore what they see through this method. Expressions such as "neutral, surprise, fear" will be adjusted. This action provides basic semantic details and temporal structure to determine the category of speech signal [15].

Additionally, facial expression recognition is widely applied in other areas, such as lie detectors, smart healthcare, and so on [16].

In summary, facial expression recognition technology has broad applications and important significance in today's society [17]. It plays a positive role in improving human-computer interaction experience, enhancing intelligence levels, and improving quality of life.

In recent years, deep learning has been widely applied in facial expression extraction, such as FNN (feedforward neural network), CNN (convolutional neural network), etc. CNN-based image recognition methods have achieved good results. The multi-layer convolutional networks of CNN can effectively extract high-level, multi-level features of the whole face or part of the face, and achieve good face classification. Experimental results show that compared to other neural networks, CNN has better image recognition capabilities. The VGG network model, as an excellent representative of CNN, has been widely used in research and applications of facial expression recognition by many researchers [18]. VGG19 is a larger convolutional neural network model that contains 19 convolutional and fully connected layers and therefore requires larger storage and computational resources. In addition, due to the fact that VGG19 has more parameters and a deeper network structure, it is prone to overfitting, especially in face expression recognition applications, where the dataset is usually small, which can easily lead to a model that performs well on the training set but overfits on the test set. This paper will make improvements on the model design and loss functions of the original VGG19 model in the VGG network, as well as conduct the corresponding comparative experiments and applications.

II. Improved VGG19 Model

A. Modelling Design

An improved VGG19 model based on deep convolutional neural network is designed for feature extraction and decision making.

1) Each small piece in the improved VGG19 consists of the following components: A convolutional layer for feature

*Corresponding author

extraction, a BatchNorm layer for accelerating the training process and increasing the convergence speed of the network, a relu layer for introducing nonlinearities to enhance the model representation, and an average pooling layer for reducing the spatial dimensionality and extracting features. These components interact with each other and together build the deep structure of the VGG19 network, enabling efficient feature extraction and classification of images.

2) A dropout strategy is introduced between the final convolutional layer and the fully-connected layer, and this strategy allows the model to maintain a stable performance on new data, significantly enhancing the robustness of the model.

3) Instead of using several fully-connected layers, we ended up adding just one fully-connected layer, then another fully-connected layer, and used softmax to classify the input into one of seven expression categories involving: anger, disgust, fear, happiness, sadness, surprise and neutral.

Fig. 1 illustrates the structure of the improved VGG-19 model.

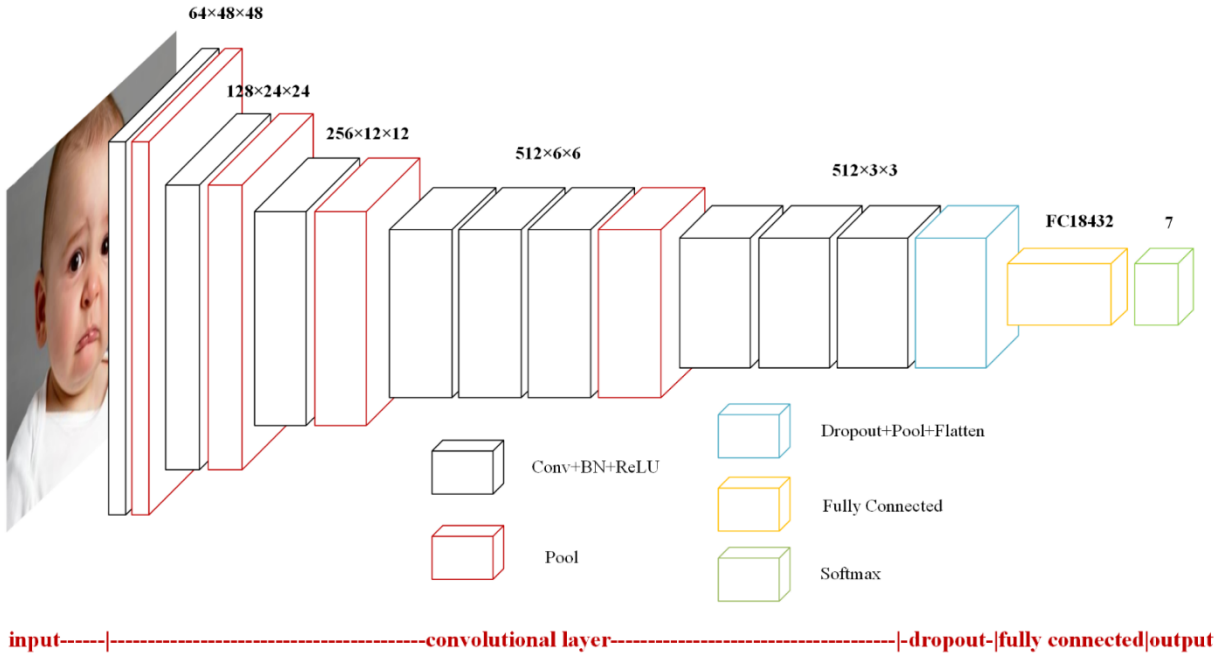

Fig. 1. The architecture of the improved VGG-19 model.

B. Loss Function Design

During the design process, the cross-entropy loss function is employed for calculations. A softmax layer is used to normalize the output probabilities of each class from the fully connected layer to 1, making data processing easier. Calculating the cross-entropy loss function is shown as Formula (1).

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^i \log\left(h_\theta(x^i)\right) + (1-y^i)\log\left(1-h_\theta(x^i)\right)\right] \quad (1)$$

In Formula (1), x^i represents the data for each category, y^i represents the correct answer for each category, $h_\theta(x^i)$ represents the predicted value obtained after processing with the improved VGG-19, and m represents the number of categories.

For the softmax regression multi-classification problem, this section solves it by using the normalized probabilities. The class label y can take k different values.

We use cross-entropy as the loss function, which corresponds to the softmax classifier we choose in the last layer. The softmax classifier is a logical classifier that is oriented towards multiple classes. Its normalized classification probabilities are more direct and sum up to 1. Cross-entropy can to some extent solve the problem of noisy labels [19], and using cross-entropy error functions can speed up training and have better generalization effects than sum-of-squares function [20].

III. Face Expression Recognition Process with Improved VGG19 Model

As shown in Fig. 2, the image is first taken as input and undergoes preprocessing including face alignment, data augmentation, normalization, etc. The resulting data is then fed into our improved VGG19 network model, which is trained on emotion class labels obtained from datasets such as CK+ and FER2013. After training, the best improved VGG19 model is obtained, and the model is then evaluated and tested. From the emotion input to the model's prediction output, scores are obtained for each category, and the final prediction is made based on the highest score value to obtain the result of emotion classification.

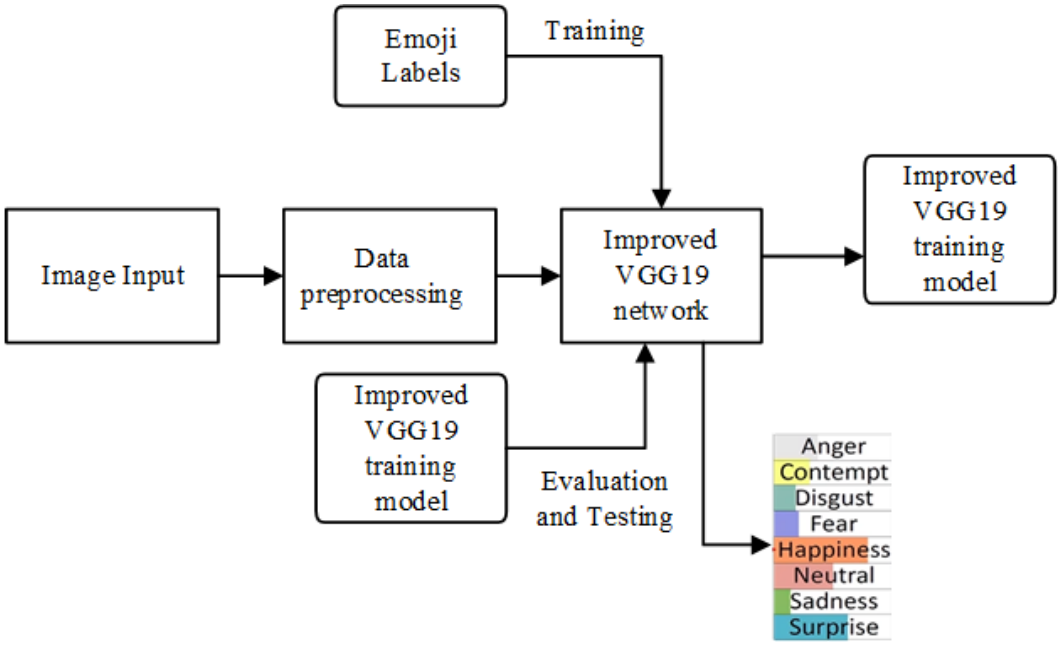

Fig. 2. Facial expression recognition process of the improved VGG19 model.

IV. Design of Experiments

A. Experimental Environment

The computer operating system used for the experiments in this paper is Windows 10，64-bit, 16G RAM. The CPU is 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30GHz, the GPU is NVIDIA GeForce RTX3060 with 8G of graphics memory. The Python version used is 3.6, the deep learning framework is Pytorch 1.1, the CUDA version is 11.8, and the Python IDE is Pycharm version 2022.1.2. The specific hardware and software configurations are shown in Table I.

TABLE I. Experimental Hardware and Software Configuration

Items	Configuration
operating system	Windows10
RAM	16GB
CPU	Intel Core i7-11800H
GPU	NVIDIA RTX3060
graphics memory	8G
software framework	Anaconda Pytorch1.1
Python	3.6
CUDA	11.8
Main usage libraries	Numpy，H5py，Sklearn

B. Dataset

FER2013 and CK+ datasets were chosen for the experiments.

The FER2013 database is derived from the Representation Learning Challenge of 2013 ICML. FER2013 is a large, unrestricted database of all images registered through Google's Image Retrieval API and resized to 48*48 pixels after eliminating mislabelled frames and adjusting the cropping region. FER2013 has 28709 training images, 3589 validation images and 3589 test images with seven expressions including: anger, disgust, fear, happiness, sadness, surprise and neutral as shown in Fig. 3. Some of these images have watermarks and noise etc., as shown in Fig. 4.

The Cohn-Kanade (CK+) database, released in 2010, is an extension of the CK database. It is the most extensive laboratory-controlled database for evaluating FER (Facial Expression Recognition) systems. CK+ includes 123 subjects and a total of 593 video segments, ranging from 10 to 60 frames in length. These videos contain 327 sequences labeled with seven basic expression labels: anger, contempt, disgust, happiness, fear, surprise and sadness. The labels are based on the Facial Action Coding System (FACS). That is to say, compared with the FER2013 dataset, it replaces neutrality with contempt. Since CK+ does not provide a specific training set, validation set, and test set, the evaluation methods for this database are not unified. Fig. 5 shows some examples of expressions in the CK+ dataset.

C. Data Processing

To enhance the data used in this section and maximize the avoidance of overfitting, we enhanced the robustness of our predictions by performing data augmentation. Specifically, for each original image with a size of 48×48, we randomly created 10 cropped images with a size of 44×44. In addition, we also collected 10 processed images for each facial expression, which are cropped from the upper left, lower left, upper right, lower right, and center, and then their reflections are extracted from each cropped image for testing. To reduce classification errors, we used the average score of the 10 images as the final result.

Fig. 3. FER Examples of dataset expressions (anger, disgust, fear, happiness, neutral, sadness, surprise).

Fig. 4. Example of noise in the FER2013 dataset.

V. Experimental Results and Comparative Analysis

A. Experimental Analysis of Improved VGG19 vs. VGG19

As shown in Table II, the improved VGG19 model exhibits better accuracy on the FER2013 dataset compared to the original VGG19 model. The accuracy of the improved VGG19 model on the public and private FER2013 datasets is 70.911% and 73.029% respectively, and higher than that of original VGG19 model. The learning rate for the experiments was set at 0.01, with 250 epochs for the FER2013 dataset and 60 epochs for the CK+ dataset.

TABLE II. Comparison of Recognition Accuracy of VGG19 and Improved VGG19

Model	FER2013 public dataset	FER2013 private dataset
VGG19	68.821%	70.995%
Improved VGG19	70.911%	73.029%

B. Comparative Experiments on the FER2013 Dataset

Based on FER2013 dataset, we compared the accuracy in face expression recognition of the improved VGG19 network model and the top ten algorithms in the 2013 Kaggle facial expression recognition competition as well as DNNRL[21], CPC [22] methods, as shown in Table III.

According to Table III, the improved VGG19 network model presented in this paper achieved an impressive facial expression recognition accuracy of 73.029% on the FER2013 dataset. This result surpasses the accuracy of the top ten algorithms from the 2013 Kaggle competition, as shown in the first ten rows of Table III. It also outperforms the recognition effects of the DNNRL and CPC network structures, which are newer models listed from rows 11 to 12. The comparison clearly indicates that the improved VGG19 model has a significant advantage in accuracy.

The reason for such an achievement is the introduction of a Batch Normalization (BN) layer into the original VGG19 network structure. Additionally, a Dropout strategy is applied between the final convolutional layer and the fully connected layer. These enhancements effectively prevent overfitting issues that can arise from the deep nature of the network and also improve the training convergence speed of the model.

TABLE III. Accuracy of each Method on the FER2013 Dataset

	Method	Accuracy
1	RBM	71.161%
2	Unsupervised	69.267%
3	Maxim Milakov	68.821%
4	Radu+Marius+Cristi	67.483%
5	Lor.Voldy	65.254%
6	Ryank	65.087%
7	Eric Cartman	64.474%
8	Xavler Bouthller	64.224%
9	AlehandroDubrovsky	63.109%
10	Sayit	62.190%
11	DNNRL	70.60%
12	CPC	71.36%
13	Improved VGG19	73.029%

C. Comparative Experiments on the CK+ Dataset

Table IV clearly shows the accuracy of different methods for facial expression recognition on the CK+ dataset. The improved VGG19 model addresses the issue of overfitting, which can occur due to the small scale and limited number of samples in the CK+ dataset, by employing Dropout and Batch Normalization (BN) strategies.

For the CK+ dataset, we used a tenfold cross-validation method. The dataset is randomly divided into 90% for training and 10% for testing. The highest accuracy achieved in the tests is 93.939%. As can be seen from the comparative results in Table IV, although the accuracy of the improved VGG19 network model is not the highest, it has reached a relatively high level.

Fig. 5. Examples of expressions from the CK+ dataset (anger, contempt, disgust, fear, happiness, sadness, surprise).

Looking at Tables III and IV, there is a significant difference in the training effects of the improved VGG19 model on the FER2013 and CK+ datasets, with a considerable gap in recognition rates. Moreover, the results on the CK+ dataset are notably better than those on the FER2013 dataset. The CK+ dataset was obtained in a laboratory environment, where factors such as background lighting and camera quality were controlled and standardized, making the dataset cleaner and more reliable. As a result, samples are more easily recognized accurately. Additionally, the dataset has undergone augmentation, resulting in higher image quality. Therefore, the algorithm has demonstrated a high facial expression recognition accuracy on the CK+ dataset.

TABLE IV. ACCURACY COMPARISON ON THE CK+ DATASET

Method	Accuracy
Shan et al.[23]	89.1%
Jeni et al. [24]	96%
Kahou et al.[25]	91.3%
Improved VGG19	93.939%

D. Confusion Matrix Analysis

Fig. 6 illustrates that the accuracy for recognizing happiness and surprise is higher than for other emotions. However, the accuracy for recognizing fear is somewhat lower. There are two reasons for this issue.

Firstly, the dataset has an imbalance in the number of images with different emotion categories. There are as many as 7,215 images for happiness, but only 436 for disgust, while the average number of images for each category is around 4,000. Such an imbalance is sufficient to cause classification errors.

Secondly, some emotions have connections with each other. For instance, anger, disgust, fear, and sadness are often difficult to distinguish in real life, especially when people do not know each other well. Furthermore, misjudgments often occur with certain categories, perhaps because some categories are indeed hard to differentiate and are easily confused.

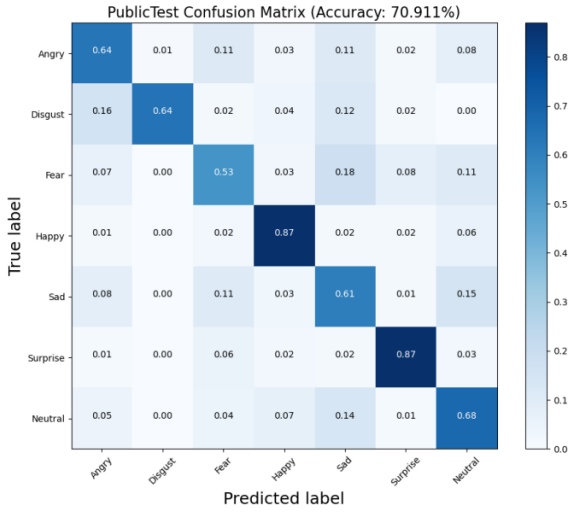

Fig. 6. Confusion matrix for the improved VGG19 model on the FER2013 PrivateTest dataset.

The next research direction will focus on modules that pay attention to specific expressions. By focusing on detailed information, the classification ability of the model can be further improved, providing more support for enhancing classification accuracy.

VI. EXAMPLE TEST

We conducted a facial expression recognition experiment using the improved VGG19 neural network and validated it. After training the best model on the FER2013 dataset, we tested it on test images to obtain the probabilities of various expressions. The probabilities of the images in each category and the model's predictions were visualized. The specific verification process includes as follows.

1) Input the test image into the improved VGG19 network to get the corresponding predicted values through the network's forward propagation.

2) Use the cross-entropy loss function to find the difference between the predicted values and the actual values.

3) Update the parameters of the network model at various levels using the backpropagation method.

Fig. 7 shows the specific test results for one of the test examples. The image of a sad expression, when identified and classified by the improved VGG19, yielded a fear probability of 0.2, a sadness probability of 0.7, and a minimal probability of neutrality. Since the highest probability was for sadness, the model outputted a sad expression, which is consistent with the test image. The analysis process for other expression examples such as Fig. 8 is similar to that shown in Fig. 7.

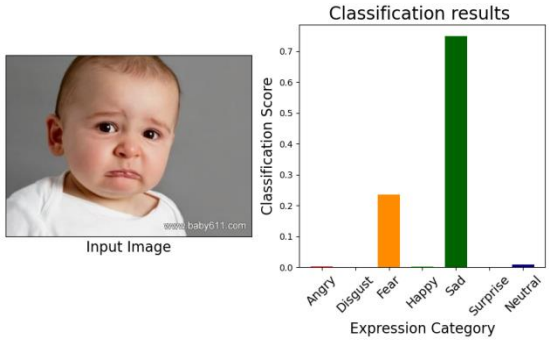

Fig. 7. Example of a sad expression.

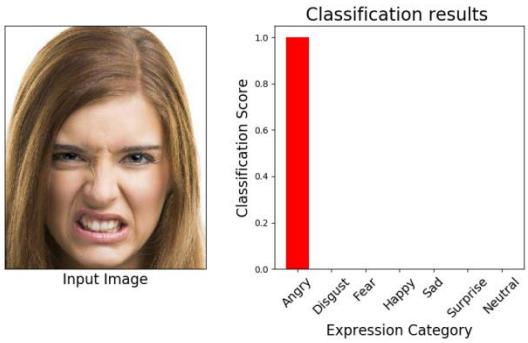

Fig. 8. Example of an angry expression.

VII. CONCLUSION

This paper provides an in-depth exploration of facial expression recognition and its applications using an improved VGG19 model. We designed the structure of the improved VGG-19 model involving enhancement strategies as well as the loss function, and describe facial expression recognition process with the improved VGG19 model. In the experimental design section, we detailed the selection of the dataset and the methods of data processing. In the section on experimental results and comparative analysis, we conducted comparative experiments on accuracy between the improved VGG19 and other classic models, including the original VGG19. We also performed instance tests, using probability histograms to reflect the effectiveness of expression recognition. These tests demonstrate the superiority, applicability, and stability of the improved VGG19 model. However, the accuracy of distinguishing some expressions, such as sadness, disgust and fear, could be improved. In this regard, future directions could focus on designing sub-networks for each expression that are dedicated to recognising specific expressions. For example, specific sub-network structures are designed for sad and upset expressions to better capture the features of these expressions.

REFERENCES

[1] Hu, L., Li, W., Yang, J., Fortino, G., Chen, M. (2019). A sustainable multi-modal multi-layer emotion-aware service at the edge. IEEE Transactions on Sustainable Computing, 7(2), 324-333.

[2] Shrivastava V, Richhariya V, Richhariya V. Puzzling Out Emotions: A Deep-Learning Approach to Multimodal Sentiment Analysis[A]. 2018 International Conference on Advanced Computation and Telecommunication (ICACAT)[C]. Bhopal, India, 2018, 1-6.

[3] Perveen N, Roy D, Chalavadi K M. Facial Expression Recognition in Videos Using Dynamic Kernels[J]. IEEE Transactions on Image Processing, 2020, 29:8316-8325.

[4] Zhi, R., Zhou, C., Li, T., Liu, S., Jin, Y. (2021). Action unit analysis enhanced facial expression recognition by deep neural network evolution. Neurocomputing, 425, 135-148.

[5] Li S, Deng W. Deep facial expression recognition: A survey[J]. IEEE Transactions on Affective Computing, 2020, 13(3):1195-1215.

[6] Mahmood M R, Abdulazeez A M. A Comparative Study of a New Hand Recognition Model Based on Line of Features and Other Techniques[A]. Recent Trends in Information and Communication Technology: Proceedings of the 2nd International Conference of Reliable Information and Communication Technology[C]. Springer International Publishing, 2018, 420-432.

[7] Mellouk W, Handouzi W. Facial emotion recognition using deep learning: review and insights[J]. Procedia Computer Science, 2020, 175:689-694.

[8] Wen, G., Chang, T., Li, H., Jiang, L. (2020). Dynamic objectives learning for facial expression recognition. IEEE Transactions on Multimedia, 22(11), 2914-2925.

[9] Ekman P, Friesen W V, Ancoli S. Facial signs of emotional experience[J]. Journal of Personality and Social Psychology, 1980, 39(6):1125-1134.

[10] Tran, H. N., Phan, P. H., Nguyen, K. H., Hua, H. K., Nguyen, A. Q., Nguyen, H. N., Nguyen, N. V. (2024). Augmentation-Enhanced Deep Learning for Face Detection and Emotion Recognition in Elderly Care Robots.

[11] Liu, D., Ouyang, X., Xu, S., Zhou, P., He, K., Wen, S. (2020). SAANet: Siamese action-units attention network for improving dynamic facial expression recognition. Neurocomputing, 413, 145-157.

[12] Zhi R, Wan M. Dynamic facial expression feature learning based on sparse RNN[A]. 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)[C]. IEEE, 2019: 1373-1377.

[13] Wu, M., Su, W., Chen, L., Pedrycz, W., Hirota, K. (2020). Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition. IEEE Transactions on Affective Computing, 13(2), 805-817.

[14] Pabba C, Kumar P. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition[J]. Expert Systems, 2022, 39(1): e12839.

[15] Chen, L., Ouyang, Y., Zeng, Y., Li, Y. (2020, August). Dynamic facial expression recognition model based on BiLSTM-Attention. In 2020 15th International Conference on Computer Science & Education (ICCSE) (pp. 828-832). IEEE.

[16] Durga, B. K., Rajesh, V., Jagannadham, S., Kumar, P. S., Rashed, A. N. Z., Saikumar, K. (2023). Deep Learning-Based Micro Facial Expression Recognition Using an Adaptive Tiefes FCNN Model. Traitement du Signal, 40(3).

[17] Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., Mellit, A. (2023). A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. Sustainability, 15(7), 5930.

[18] Mahendar M, Malik A, Batra I. Facial Micro-expression Modelling-Based Student Learning Rate Evaluation Using VGG–CNN Transfer Learning Model[J]. SN Computer Science, 2024, 5(2): 204.

[19] Bishop C M. Pattern recognition and machine learning[J]. Springer google schola, 2006, 2: 1122-1128.

[20] Simard P Y, Steinkraus D, Platt J C. Best practices for convolutional neural networks applied to visual document analysis[A]. In Proceedings of the Seventh International Conference on Document Analysis and Recognition[C]. USA: IEEE Computer Society, 2003, 958-963.

[21] Kim, B. K., Roh, J., Dong, S. Y., Lee, S. Y. (2016). Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. Journal on Multimodal User Interfaces, 10, 173-189.

[22] Chang, T., Wen, G., Hu, Y., Ma, J. (2018). Facial expression recognition based on complexity perception classification algorithm. arXiv preprint arXiv:1803.00185.

[23] Xie, S., Shan, S., Chen, X., Meng, X., Gao, W. (2009). Learned local Gabor patterns for face representation and recognition. Signal Processing, 89(12), 2333-2344.

[24] Jeni, L. A., Takacs, D., Lorincz, A. (2011, November). High quality facial expression recognition in video streams using shape related information only. In 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops) (pp. 2168-2174). IEEE.

[25] Kahou, S. E., Froumenty, P., Pal, C. (2014, September). Facial expression analysis based on high dimensional binary features. In European Conference on Computer Vision (pp. 135-147). Cham: Springer International Publishing.