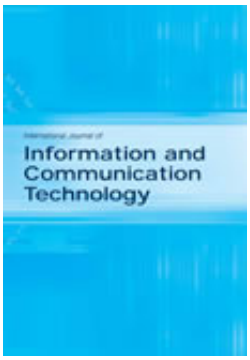
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Facial expression recognition based on YOLOv8 deep learning in complex scenes

Chujie Xu, Yong Du, Wenjie Zheng, Tiejun Li, Zhansheng Yuan

Article History:

Received:	27 October 2024
Last revised:	25 November 2024
Accepted:	25 November 2024
Published online:	20 January 2025

Facial expression recognition based on YOLOv8 deep learning in complex scenes

Chujie Xu, Yong Du*, Wenjie Zheng, Tiejun Li and Zhansheng Yuan

School of Ocean Information Engineering,
Jimei University,
Xiamen, 361021, China
Email: xcjjack@jmu.edu.cn
Email: duyong_jmu@163.com
Email: cin911@jmu.edu.cn
Email: litiejun@jmu.edu.cn
Email: yzsyymylf@jmu.edu.cn
*Corresponding author

Abstract: Accurate and effective facial expression recognition (FER) is of great significance in fields such as intelligent monitoring and emotional computing today. This article proposes a deep learning method based on You Only Look Once Version 8 (YOLOv8), which combines YOLOv8's real-time and efficient object detection capabilities with the feature extraction advantages of convolutional neural networks (CNNs) to improve facial expression recognition performance in complex environments. Firstly, YOLOv8 is used for precise facial detection. Then, the detected facial regions are fed into a feature extraction network, which extracts high-level features related to facial expressions through deep CNN, enhancing the robustness of the model to different complex scenes. The results of the experiment indicate that our approach performs well on publicly available facial expression datasets, especially in complex scenes where it significantly outperforms traditional expression recognition methods. This model provides new ideas for future applications in diverse, dynamic, and complex environments.

Keywords: facial recognition; facial expression recognition; FER; convolutional neural network; CNN; YOLOv8.

Reference to this paper should be made as follows: Xu, C., Du, Y., Zheng, W., Li, T. and Yuan, Z. (2025) 'Facial expression recognition based on YOLOv8 deep learning in complex scenes', *Int. J. Information and Communication Technology*, Vol. 26, No. 1, pp.89–101.

Biographical notes: Chujie Xu received his Master's degree from South China University of Technology. Currently, he works at Jimei University. His research interests include acoustic communication, underwater localisation, wireless communication and wireless sensor networks.

Yong Du received his Master's in Radio Physics from Lanzhou University in 2004. Currently, he is an Associate Professor at Jimei University. His current research interests include fibre-optic sensors, terahertz technology and artificial intelligence.

Wenjie Zheng received his Master's in Software Engineering from Xiamen University in 2010. Currently, he is a Senior Engineer at Jimei University. His current research interests include intelligent sensor, digital signal process and artificial intelligence.

Tiejun Li received his Master's and PhD degrees from the Chongqing University in 2005 and 2018. Currently, he is an Associate Professor at Jimei University. His research interests are in machine learning, terahertz technology, and sensors.

Zhansheng Yuan received his Master's in Measuring and Testing Technology and Instruments from JiLin University in 2001. Currently, he is an Associate Professor at Jimei University. His current research interests include semiconductor conductivity and integrated circuit design.

1 Introduction

With the sustained growth of artificial intelligence (AI) and computer vision technology, FER has become an important research topic in fields such as emotional computing, intelligent monitoring, medical diagnosis, and human-computer interaction. Facial expressions, as an important way of conveying and recognising human emotions, can reflect a person's psychological state and emotional changes. Therefore, how to accurately identify and classify facial expressions through automated means has become a research hotspot in recent years. However, facial expression recognition (FER) in complex scenarios continues to encounter numerous challenges, including variations in lighting, occlusion, and posture, which greatly affect the accuracy of FER models. To address these challenges, researchers have attempted to use deep learning techniques, especially CNN, aiming to enhance the accuracy and resilience of FER systems in challenging environments.

Initial studies on FER primarily depended on conventional machine learning approaches and manual feature extraction methods, such as LBP (Ojala et al., 1996) and SIFT (Lowe, 2004) methods can achieve good FER results under ideal conditions. These methods extract facial textures, edges, or local features, and combine them with classifiers such as support vector machines (SVM) for facial expression classification. However, traditional methods based on manual features exhibit significant limitations when dealing with changing lighting, facial expressions, and complex backgrounds in real-world scenarios. The manually designed features are difficult to effectively capture the complex changes in facial expressions, and their robustness is insufficient to adapt to complex and ever-changing practical application scenarios.

In recent years, the emergence of deep learning technology, particularly the extensive use of CNNs, has gained prominence, FER technology has made significant progress. CNN avoids the limitations of manual feature design in traditional methods by automatically learning feature representations from low-level to high-level. Kahou et al. (2013) introduced a deep neural network model that combines multimodal inputs to recognise facial expressions in videos. The experimental results demonstrated that the method achieved high recognition accuracy under multimodal fusion. Mollahosseini et al. (2017) further proposed a large-scale facial expression dataset called AffectNet and

trained a multimodal deep convolutional network model based on it, which achieved leading recognition results on multiple publicly available datasets.

Although CNN-based expression recognition methods have to some extent addressed the limitations of traditional methods, they still face challenges in complex scenarios. In practical applications, changes in lighting, partial facial occlusion, different shooting angles, and complex backgrounds often lead to a decline in the performance of FER. To this end, several researchers have introduced a method that integrates object detection and expression recognition, using face detection as a prerequisite for expression recognition, in order to improve the performance of the model in complex scenes. Zhao et al. (2021) introduced a multi-task learning model utilising faster R-CNN, which demonstrated good performance on public datasets by simultaneously performing face detection and expression recognition. Though their processing speed is slow and challenging to satisfy the high time-sensitive needs of application scenarios, models based on faster R-CNN can efficiently detect faces in complex backdrops.

The YOLO object identification technique has progressively become a useful tool in this field to solve the challenges of FER in complex circumstances and satisfy the needs of real-time processing. Fast and precise object identification features of the YOLO model make it extensively applied in several computer vision applications. As the most recent series version, YOLOv8 has been polished in speed and accuracy even further. Combining self attention mechanism, feature pyramid network (FPN), and other optimisation approaches (Wang et al., 2023a) YOLOv8 may attain quicker inference speed while preserving high detection accuracy. Thus, in order to extract and classify expression features, this article proposes an expression recognition method based on YOLOv8, which uses YOLOv8's efficient and accurate face detection capabilities, and further combines convolutional neural networks (CNNs) to extract and classify expression features, so improving expression recognition performance in complex scene.

First of all, this paper presents a face detection model based on YOLOv8, which lays the basis for later expression recognition by first rapidly and precisely identifying facial areas in challenging backgrounds. Second, this paper enhances the model's robustness to various complicated environments by using deep CNNs to capture high-level traits of expressions. Our method was tested on several publicly available facial expression datasets in the experiment, and the results revealed that in complex scenarios, our method outperforms conventional FER methods in terms of recognition accuracy and real-time performance, so displaying good generalising ability.

This paper suggests an expression recognition method combining the effective face detection capacity of YOLOv8 with the strong feature extraction capability of CNN to handle challenging surroundings. This approach offers a fresh way to identify facial emotions applications in complicated settings in the future since it operates well under conditions like lighting changes, partial occlusion, and several positions. This paper will include a thorough introduction to the design concept and implementation details of the approach, thereby verifying its advantages and efficacy by means of tests.

2 Relevant technologies

2.1 YOLOv8

Continuating the YOLO model and obtaining quick end-to-end object identification by a single forward propagation, YOLOv8 is the most recent iteration of the YOLO series of object detection models (Li et al., 2023). Particularly by merging self attention mechanism with FPN, which improves its capacity to identify multi-scale targets (Wang et al., 2023b), YOLOv8 has been further refined in terms of speed and accuracy against past generations.

Object detection requires locating the bounding box that encases the target object within the image and classifying it. For each detected target, YOLOv8 outputs the predicted category c, position (x, y), in addition to the bounding box's width w and height h. Define the detection target as a vector:

$$p = (x, y, w, h, c) \tag{1}$$

where coordinates (x, y) specify the centre of the bounding box, and w and h denote its width and height, and c is the probability distribution of the target category.

The loss function of YOLOv8 integrates bounding box regression loss, classification loss, as well as the loss associated with target confidence. Generally, the loss function is expressed as:

$$L = \lambda_{bbox} L_{bbox} + \lambda_{cls} L_{cls} + \lambda_{conf} L_{conf} \tag{2}$$

where L_{bbox} is the regression loss of the bounding box; L_{cls} represents the classification loss; L_{conf} is the target confidence loss; λ_{bbox}, λ_{cls}, and λ_{conf} are the weight hyperparameters of the loss.

YOLOv8 uses complete IoU (CIoU) loss to optimise the position of bounding boxes. It is a comprehensive indicator that considers the overlap rate, distance, and aspect ratio of bounding boxes. The equation is as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b')}{c^2} + \alpha \cdot v \tag{3}$$

where IoU represents the Intersection over Union of the predicted box and the ground truth box; $\rho = (b, b')$ represents the Euclidean distance between the centres of two boxes; c denotes the diagonal measurement of the least enclosing area that can enclose two boxes; v represents the measure of consistency for aspect ratio, and α is the adjustment parameter.

IoU is used to measure the degree of overlap between predicted boxes and real boxes, and its equation is:

$$IoU = \frac{Area\ of\ overlap}{Area\ of\ union} = \frac{B_p \cap B_g}{B_p \cup B_g} \tag{4}$$

where B_p and B_g refer to the predicted bounding box and the actual bounding box, respectively. Classification loss is usually measured using cross entropy loss, which measures the variation in the predicted class distribution and the true class. The equation is:

$$L_{cls} = -\sum_{i=1}^{C} y_i \log\left(\hat{y}_i\right) \tag{5}$$

where C indicates the count of categories, y_i is the indicator of the true category label, and $\hat{y}_i$ is the predicted category probability by the model

The YOLO series models use target confidence to represent the model's prediction of the existence of a target object (Wu and Dong, 2023). The loss of target confidence can be measured by binary cross entropy (BCE):

$$L_{conf} = -\left(y \log(p) + (1-y)\log(1-p)\right) \tag{6}$$

where y the designation for whether the real target exists or not, and p is the confidence level of the prediction.

YOLOv8 adopts a dynamic anchor box mechanism to match targets of different scales, and the size of the anchor box is generated through a pre-set clustering algorithm. For each anchor box, the model predicts the offset (Δx, Δy, Δw, Δh):

$$x = x_a + \Delta x \cdot w_a \tag{7}$$

$$y = y_a + \Delta y \cdot h_a \tag{8}$$

$$w = w_a + \exp(\Delta w) \tag{9}$$

$$h = h_a + \exp(\Delta h) \tag{10}$$

where (x_a, y_a, w_a, h_a) refers to the centre coordinates, width, and height of the anchor box, and the predicted offset of the model is used to adjust the anchor box to better fit the target.

To avoid multiple detection boxes overlapping on the same target, YOLOv8 uses NMS to suppress highly overlapping prediction boxes. The equation for NMS is:

$$\hat{B} = \left\{B_i : IoU\left(B_i, B_j\right) < threshold, \forall j \neq i\right\} \tag{11}$$

where $\hat{B}$ is the set of suppressed bounding boxes, and IoU measures the degree of overlap between a pair of bounding boxes.

YOLOv8 uses a FPN for combining features across various scales, with the equation:

$$P_l = Conv\left(C_l\right) + Upsample\left(P_{l+1}\right) \tag{12}$$

where P_l is the pyramid feature map of the l^{th} layer, C_l is the convolutional feature of the l^{th} layer, and *Upsample* represents the upsampling operation used to enhance the detail information of the high-resolution layer.

The final output of the YOLOv8 model includes three sections: category prediction, bounding box regression, and target confidence. The total output of the model is:

$$O = (c, b, p) \tag{13}$$

where c is the category prediction vector, b denotes the parameter for bounding box regression, and p is the target confidence level.

2.2 CNN

CNN replaces fully connected layers with convolutional operations, allowing the model to extract local features more effectively and have translation invariance (Chua, 1997; Kattenborn et al., 2021). CNN is thus extensively used since it has obtained outstanding leads to progress in jobs including image classification, object detection, and speech recognition. Typically consisting of many layers – the convolutional layer, activation function layer, pooling layer, and fully connected layer (Alzubaidi et al., 2021; Chua and Roska, 1993).

The convolutional layer is the core component of CNN, mainly used to extract local features of the input image. Using filters – also called convolution kernels – convolutions operations locally weight and sum the input to produce feature maps (Shin et al., 2016; Bhatt et al., 2021). The convolution operation's equation is:

$$S(i,\ j) = (X * K)(i,\ j) = \sum_{m} \sum_{n} X(i+m,\ j+n) \cdot K(m,\ n) \tag{14}$$

where $X(i, j)$ is the pixel value of the input image; $K(m, n)$ is the weight matrix of the convolution kernel; $S(i, j)$ is the output feature map. An important feature of convolutional layers is parameter sharing, where the same convolutional kernel slides across different regions of the input image, allowing the same kernel to extract similar features from different positions in the image.

Nonlinear transformations help the layer of activation function let the network learn and depict increasingly complicated features. ReLU and the sigmoid function are the most often utilised activation functions.

The most often used activation function is ReLU; its equation is:

$$f(x) = \max(0,\ x) \tag{15}$$

ReLU only stores positive values and sets negative values to 0, hence adding nonlinearity.

Usually applied in binary classification issues, the sigmoid function maps the input to the range of (0, 1). The recipe calls for:

$$f(x) = \frac{1}{1+e^{-x}} \tag{16}$$

Among the usual pooling operations are max pooling and average pooling (Cai and Vasconcelos, 2019).

To reduce the dimensionality of feature maps and lower both model parameters and computational complexity, the pooling layer is employed, all while maintaining key feature information. Maximising pooling is choosing the maximum value within the pooling window; the formula is:

$$S(i,\ j) = \max_{m,n}(i+m,\ j+n) \tag{17}$$

Usually employed for image classification problems, maximum pooling helps to retain most important characteristics.

The equation is average pooling – that is, choosing the mean value from the pooling window:

$$S(i,\ j) = \frac{1}{|W|}\sum_{m,n} X(i+m,\ j+n) \tag{18}$$

Average pooling can smooth the feature map and reduce the impact of noise.

Typically, the fully connected layer is utilised at the end of a network to map extracted features into the final category prediction. Each neuron has connections to every neuron in the previous layer (Zheng et al., 2017), and the resulting output can be described by the equation below:

$$y = \sum_i w_i x_i + b \tag{19}$$

where x_i is the input; w_i is the weight; b is bias; y is the output.

3 FER model based on YOLOv8 and CNN

In order to achieve face detection based on YOLOv8 and recognition of facial expressions through CNN, the entire process can be divided into two main stages: detection of faces and recognition of expressions. Firstly, YOLOv8 is used to accurately locate and detect facial regions in complex backgrounds, extracting face frames in real-time and efficiently. Next, the detected facial regions are input into a CNN, which extracts high-level features related to facial expressions through multi-layer convolution operations. Finally, CNN classifies these features and outputs the categories of facial expressions. Figure 1 illustrates the model framework diagram:

YOLOv8 is used to detect the position of faces from complex backgrounds. The feature of YOLOv8 is its ability to process and output detected bounding boxes in real-time, representing the position of the face in the image. YOLOv8 accepts an input image I with a size of $H \times W \times 3H$ (where H and W represent the dimensions of the image, including height and width, and three represents the RGB channel). YOLOv8 uses multiple predefined anchor boxes for multi-scale detection.

Given an input image I, YOLOv8 will output N detection boxes, each consisting of position parameters (x, y, w, h) and category confidence c.

$$D = \left\{\left(x_i,\ y_i,\ w_i,\ h_i,\ c_i\right)\right\}_{i=1}^{N} \tag{20}$$

where (x_i, y_i) refers to the centre point of the i^{th} detection box; w_i, h_i is the width and height of the detection box; c_i is the category confidence level (in this case, the confidence level of 'face').

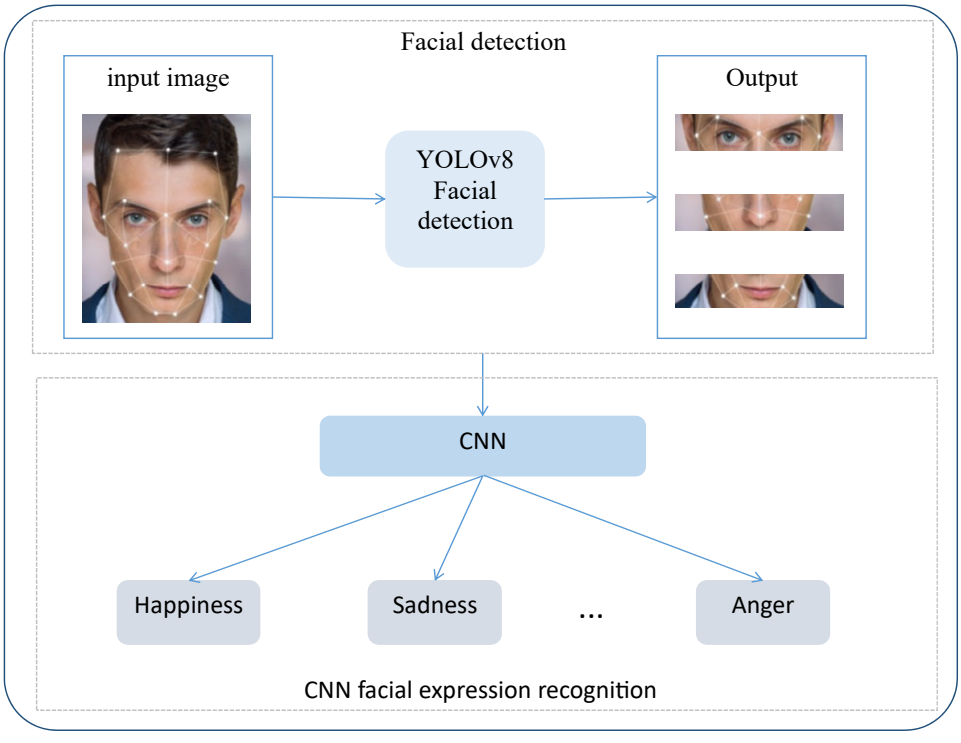
YOLOv8 will output a matrix containing the position information and category probability of each detection box. For each detection box i, we have:

$$D_i = \left(x_i,\ y_i,\ w_i,\ h_i,\ c_i\right) \tag{21}$$

After removing highly overlapping frames through non-maximum suppression (NMS), we finally obtained a set of detected face frames.

$$\hat{D} = \left\{\left(x_i,\ y_i,\ w_i,\ h_i\right)\right\}_{i=1}^{\hat{N}} \tag{22}$$

Figure 1 Model framework diagram (see online version for colours)



After detecting the facial regions, these regions are cropped from the original image I. For each detection box (x_i, y_i, w_i, h), face region I_i can be represented as:

$$I_i = I\left[x_i - \frac{w_i}{2} : x_i + \frac{w_i}{2}, y_i - \frac{h_i}{2} : y_i + \frac{h_i}{2}\right] \tag{23}$$

In this way, we obtain the set $\{I_i\}_{i=1}^{\hat{N}}$ of all facial regions. Next, the facial regions extracted from YOLOv8 will be fed into CNN to extract high-level features related to facial expressions and perform expression classification. Each cropped face image I_i will be input into a CNN. The convolutional layers in CNN are capable of extracting local facial features, such as the shape of the eyes and mouth, changes in eyebrows, etc.

The convolution operation can be expressed as:

$$F_i = W * I_i + b \tag{24}$$

where W is the convolution kernel, representing the learned filter; * representing convolution operations; B is bias; F_i represents the feature map corresponding to the i^{th} face.

After convolution operation, non-linear activation functions are usually applied, such as ReLU:

$$F_i^{'} = \max(0, F_i) \tag{25}$$

Then, the feature map is downsampled through pooling layers (such as max pooling) to reduce its dimensionality, prevent overfitting, and preserve the most important features:

$$P_i = MaxPool\left(F_i^{'}\right) \tag{26}$$

After feature extraction through multiple convolutional and pooling layers, the features are flattened and passed to the fully connected layer:

$$y_i = W_{fc} P_i + b_{fc} \tag{27}$$

where W_{fc} is the weight of the layer that is fully connected; b_{fc} is bias; y_i is the output vector representing the probability distribution of the expression category.

Convert the output to probability using the *Softmax* function:

$$\hat{y}_i = Softmax\left(y_i\right) = \frac{e^{y_i^k}}{\sum_{j=1}^{C} e^{y_i^j}} \tag{28}$$

where C is the number of categories of expressions.

Throughout the training phase, the cross-entropy loss function is employed to assess the disparity between the predicted output $\hat{y}_i$ of the model and the true label y_i:

$$L_{cls} = -\sum_{i=1}^{C} y_i \log\left(\hat{y}_i\right) \tag{29}$$

By using backpropagation and gradient descent, adjust the weight parameters W and bias b through the network to decrease the loss function.

4 Experiment

4.1 Dataset

The dataset used in this article is RAF-DB, which was collected in a non-experimental environment with diverse sample sources and wide coverage. There are significant differences between samples in terms of age, gender, race, head posture, lighting, and occlusion. Therefore, the RAF-DB dataset is selected as the model training and validation dataset for this section. The database contains 29,670 facial images, all of which are filtered on the internet. There are a total of 15,339 images used for facial expression classification, the training set contains 12,271 images, while the testing set has 3,068 images. In the data pre-processing stage, each image is aligned and cropped, resulting in a processed size of 112 × 112 pixels. This dataset contains a total of seven basic expression labels, and the expression labels corresponding to the images are manually annotated. Table 1 shows the distribution of the number of each type of facial expression found in the RAF-DB dataset.

Table 1 The proportion of each type of expression in the RAF-DB dataset

Emoji category	*Proportion of image quantity (%)*
Surprise	10.55
Fear	2.31
Disgust	5.72
Happiness	38.84
Sadness	16.04
Anger	5.65
Neutral	20.89
Total	15,339

4.2 Experimental results and analysis

The overall expression recognition accuracy of the model on the test set reached 91.5%. This result indicates that the model can still maintain high recognition performance when facing complex environmental conditions such as lighting changes, pose changes, occlusion, etc. Specifically for each expression category, Table 1 shows the distribution of accuracy, as shown in Figure 2.

Table 2 Distribution of accuracy rates for different facial expressions

Emoji category	*Accuracy (%)*
Surprise	88.2
Fear	87.8
Disgust	89.0
Happiness	94.1
Sadness	90.5
Anger	91.0
Neutral	93.0

From these data, it can be seen that the model performs best in recognising happy and neutral expressions, while performing relatively poorly in recognising surprise and fear. Possible factors contributing to this situation include uneven sample sizes of corresponding expressions in the dataset, background complexity, and subtle differences in expressions.

To thoroughly assess the performance of the proposed model, we conducted comparative experiments with several mainstream FER methods, including established machine learning practices such as SVM and directional gradient histogram (HOG) feature extraction. Basic CNN: using classic CNN structures such as VGG16 for FER. YOLOv5: as the previous generation model of the YOLO series, compared with YOLOv8 and EfficientNet. The comparison of experimental results is shown in Table 3 and Figure 3.

Figure 2 Distribution map of accuracy of different facial expressions (see online version for colours)

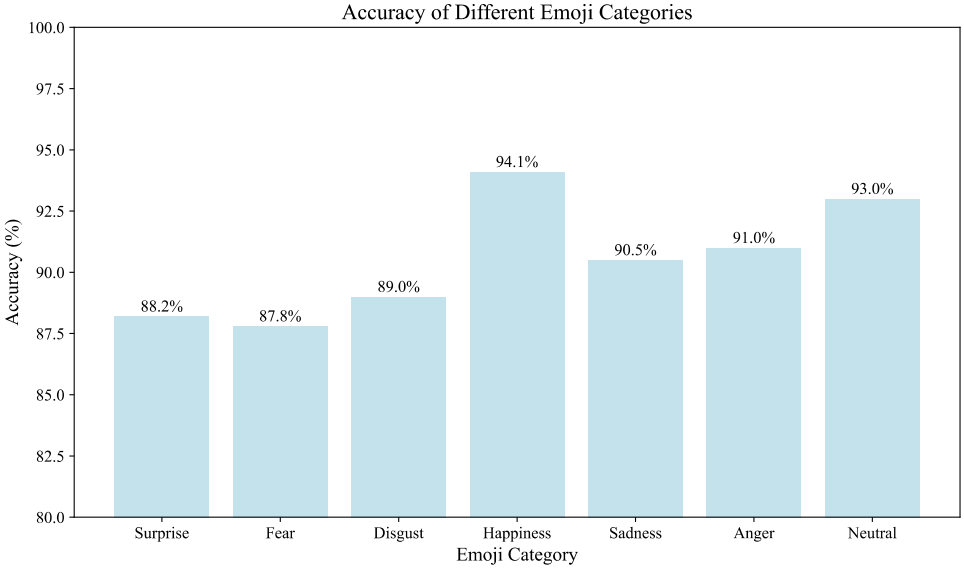
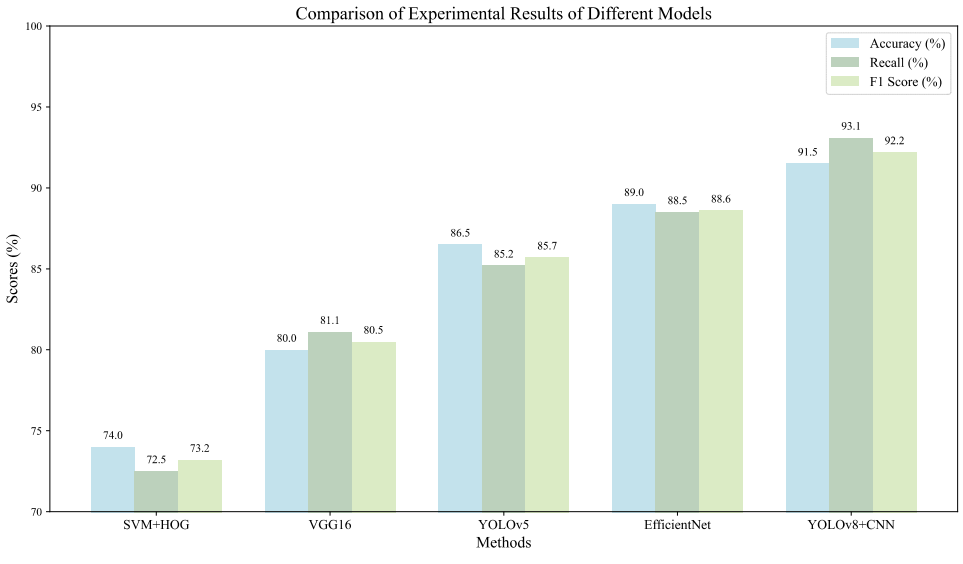


Table 3 Comparison of experimental results from different models

Method	*Accuracy (%)*	*Recall (%)*	*F1-score (%)*
SVM+HOG	74.0	72.5	73.2
VGG16	80.0	81.1	80.5
YOLOv5	86.5	85.2	85.7
EfficientNet	89.0	88.5	88.6
YOLOv8+CNN	91.5	93.1	92.2

In terms of accuracy, recall, and F1-score, the YOLOv8+CNN model performs well and outperforms other methods. Among them, the improvement in recall rate is particularly significant, indicating that the model has significantly enhanced its ability to recognise positive samples (i.e., expressions). The reasons for the performance improvement are as follows. Firstly, as the latest version of YOLO, YOLOv8 has faster speed and higher accuracy in object detection, which can quickly and accurately detect faces in complex backgrounds, reducing false positives and missed detections. Secondly, after combining YOLOv8 for face detection, a deep convolutional neural network is utilised to extract advanced features, enabling the model to capture more detailed facial expression features and improve classification ability. Additionally, several data augmentation techniques, including rotation, scaling, and flipping, were implemented during training to improve the model's robustness, enabling it to maintain good performance even in the face of different expressions and environmental conditions.

Figure 3 Comparison chart of experimental results (see online version for colours)



5 Conclusions

As AI technology continues to advance, FER, a crucial emotion computing technology, is extensively applied in areas like intelligent monitoring and human-computer interaction, and mental health assessment. However, factors such as changes in lighting, posture, and occlusion in complex environments pose many challenges for FER. This study proposes a deep learning method based on YOLOv8, which combines YOLOv8's efficient face detection capability with the feature extraction advantages of CNN to improve FER performance in complex scenes. The model proposed in this article not only has good generalisation ability and real-time processing ability, but also provides new ideas for future applications in diverse, dynamic and complex environments. Future research can further explore the application of larger datasets, model integration, and real-time systems to enhance the adaptability and performance of models in practical scenarios.

Acknowledgements

This work is supported by the Department of Education of Fujian Province (No. JAT220194), the Department of Education of Fujian Province (No. JAT210232) and the National Natural Science Foundation Cultivation Program of Jimei University (No. ZP2020036).

References

Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. (2021) 'Review of deep learning: concepts, CNN architectures, challenges, applications, future directions', *Journal of Big Data*, Vol. 8, pp.1–74.

Bhatt, D., Patel, C., Talsania, H. et al. (2021) 'CNN variants for computer vision: history, architecture, application, challenges and future scope', *Electronics*, Vol. 10 No. 20, p.2470.

Cai, Z. and Vasconcelos, N. (2019) 'Cascade R-CNN: high quality object detection and instance segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 5, pp.1483–1498.

Chua, L.O. (1997) 'CNN: a vision of complexity', *International Journal of Bifurcation and Chaos*, Vol. 7, No. 10, pp.2219–2425.

Chua, L.O. and Roska, T. (1993) 'The CNN paradigm', *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, Vol. 40, No. 3, pp.147–156.

Kahou, S.E., Pal, C., Bouthillier, X. et al. (2013) 'Combining modality specific deep neural networks for emotion recognition in video', *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pp.543–550.

Kattenborn, T., Leitloff, J., Schiefer, F. et al. (2021) 'Review on convolutional neural networks (CNN) in vegetation remote sensing', *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 173, pp.24–49.

Li, Y., Fan, Q., Huang, H. et al. (2023) 'A modified YOLOv8 detection network for UAV aerial image recognition', *Drones*, Vol. 7, No. 5, p.304.

Lowe, D.G. (2004) 'Distinctive image features from scale-invariant keypoints', *International Journal of Computer Vision*, Vol. 60, pp.91–110.

Mollahosseini, A., Hasani, B. and Mahoor, M.H. (2017) 'Affectnet: a database for facial expression, valence, and arousal computing in the wild', *IEEE Transactions on Affective Computing*, Vol. 10, No. 1, pp.18–31.

Ojala, T., Pietikäinen, M. and Harwood, D. (1996) 'A comparative study of texture measures with classification based on featured distributions', *Pattern Recognition*, Vol. 29, No. 1, pp.51–59.

Shin, H.-C., Roth, H.R., Gao, M. et al. (2016) 'Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning', *IEEE Transactions on Medical Imaging*, Vol. 35, No. 5, pp.1285–1298.

Wang, G., Chen, Y., An, P. et al. (2023a) 'UAV-YOLOv8: a small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios', *Sensors*, Vol. 23, No. 16, p.7190.

Wang, X., Gao, H., Jia, Z. et al. (2023b) 'BL-YOLOv8: an improved road defect detection model based on YOLOv8', *Sensors*, Vol. 23, No. 20, p.8361.

Wu, T. and Dong, Y. (2023) 'YOLO-SE: improved YOLOv8 for remote sensing object detection and recognition', *Applied Sciences*, Vol. 13, No. 24, p.12977.

Zhao, R., Liu, T., Xiao, J. et al. (2021) 'Deep multi-task learning for facial expression recognition and synthesis based on selective feature sharing', *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, pp.4412–4419.

Zheng, L., Yang, Y. and Tian, Q. (2017) 'SIFT meets CNN: a decade survey of instance retrieval', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 5, pp.1224–1244.