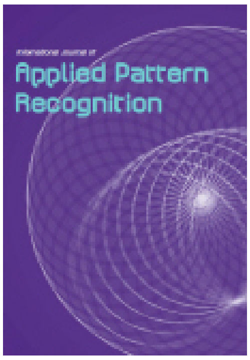
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Applied Pattern Recognition

ISSN online: 2049-8888 - ISSN print: 2049-887X
https://www.inderscience.com/ijapr

Face expression classification and recognition based on LBP+GLCM features and attention mechanism in CNN

Xia Zhang, Caini Yan

DOI: 10.1504/IJAPR.2025.10075277

Article History:

Received:	10 February 2025
Last revised:	11 July 2025
Accepted:	26 August 2025
Published online:	07 January 2026

Face expression classification and recognition based on LBP+GLCM features and attention mechanism in CNN

Xia Zhang*

School of Intelligent Manufacturing,
Longdong University,
Qingyang City, Gansu Province 745000, China
Email: zhx8613@126.com
*Corresponding author

Caini Yan

School of Literature and History,
Longdong University,
Qingyang City, Gansu Province 745000, China
Email: 578840364@qq.com

Abstract: We propose a lightweight facial expression recognition (FER) model that integrates local binary pattern (LBP) and grey level co-occurrence matrix (GLCM) hybrid features with a channel attention mechanism based on the ECA.Net variant to effectively balance the accuracy-efficiency trade-offs. This model leverages texture features extracted via LBP and LBP+GLCM, combined with attention mechanisms to enhance the focus on salient facial cues. Built upon the VGG16 CNN architecture within the TensorFlow framework, the model was tested on the FER2013 and RAF-DB datasets for cross-validation, achieving a recognition accuracy of 79.89% and 86.77%, respectively. To validate its practical effectiveness, a deployable QT-based user interface system was developed, enabling real-time FER from images, videos, and live camera feeds. This approach aims to meet the increasing demand for lightweight, reliable solutions, particularly in scenarios with limited computational resources, while maintaining high recognition accuracy.

Keywords: attention mechanism; facial expression recognition; FER; convolutional neural network; CNN; FER2013 dataset; RAF-DB dataset; local binary pattern; LBP.

Reference to this paper should be made as follows: Zhang, X. and Yan, C. (2025) 'Face expression classification and recognition based on LBP+GLCM features and attention mechanism in CNN', *Int. J. Applied Pattern Recognition*, Vol. 8, No. 1, pp.1–15.

Biographical notes: Xia Zhang works in Intelligent Manufacturing, Longdong University. Her research interest includes computational intelligence, nonlinear signal processing and image processing.

Caini Yan works in School of Literature and History, Longdong University. Her research interests include foreign language education and international relations.

1 Introduction

Facial expressions play an extremely important role in human emotional communication activities. They can effectively convey human thoughts, views, and physical states. Studying human facial expressions to understand human intentions and purposes has become a research trend in fields such as automated driving, human-computer interaction, and medical treatment. As early as the early 1970s, psychologists Ekman and Friesen (1978) studied six expressions – happy, surprise, sadness, fear, anger and disgust – and established an image database based on their research, pioneering the study of human facial expressions.

In recent years, with the rapid development of computer technology, more and more scholars have focused on facial expression recognition (FER). Early FER primarily employed algorithms such as local binary pattern (LBP) (Hu and Liu, 2004), Gabor wavelet methods (Ojala et al., 2002), and scale-invariant feature transform (SIFT) (Pang et al., 2002) to extract facial expression information. These methods were then combined with traditional machine learning algorithms – such as the Bayes classifier, support vector machines (SVM), and other classifiers – to achieve recognition, yielding good results at that time. For example, in the literature (Lowe, 2004), the Gabor method was used to extract features, along with LBP and local phase quantisation (LPQ) encoding, combined with principal component analysis (PCA) for dimensionality reduction. Experiments on the JAFFE dataset achieved high recognition accuracy. Vishwakarma and Mishra (2019) proposed a multi-level sparse (MLS) classifier based on sparse representation (SR) and multi-modal feature extraction, and verified the reliability of this method on four standard databases: ORL, YALE, AR, and Georgia Tech. Liu et al. (2016) combined Gabor filters with LBP to capture more subtle appearance details across various scales, achieving an average recognition rate of 96.3% on the CK+ dataset. Salur and Aydin (2020) processed facial expressions using LBP to obtain recognisable texture features. By employing an SVM classifier, they achieved a recognition rate of 91.4% on the CK+ dataset. Features extracted by traditional expression recognition algorithms are considered shallow and require manual labelling, which can affect recognition accuracy to some extent.

With the rapid development of machine learning, the field of FER has opened new opportunities. Deep learning models such as convolutional neural networks (CNNs), deep belief networks (DBNs), and stacked autoencoders (SAEs) (Vaswani et al., 2017) are commonly used in FER. These models effectively address the shortcomings of traditional algorithms and can extract deep features with richer informational content. Gu et al. (2016) combined DBNs with multi-layer perceptrons (MLPs) to propose a deep learning-based FER method, validating its performance on the CK+ and JAFFE datasets. The results showed that this approach outperformed classification methods such as nearest neighbour, MLP, and SVM. Alshahrani et al. (2020) integrated CNN with long short-term memory (LSTM) networks to develop a new FER framework, which has been validated on datasets such as CK+, MMI, SFEW, and a self-built database, demonstrating strong processing effectiveness.

In recent years, researchers have combined traditional feature extraction algorithms with deep learning networks, leading to continuous improvements in recognition rates. For example, the combination of Gabor wavelet transform with CNN achieved a recognition accuracy of 96.81% on the CK+ dataset, surpassing the performance of either method alone (Qin et al., 2020). Connie et al. (2017) studied and compared the performance of Dense SIFT+CNN and SIFT+CNN on FER-2013 and CK+ datasets. Results showed that Dense SIFT+CNN outperformed both traditional CNN and SIFT+CNN models. Aouani and Ayed (2024) extracted histograms of oriented gradients (HOG) and compared the use of SVM, CNN, and a hybrid CNN-SVM classifier. Validation on the Ryerson Multimedia Laboratory (RML) dataset showed that the CNN-SVM classifier outperformed the CNN-MLP classifier. Xu and Sun (2021) proposed a self-attention deep auto-encoder (SA-DAE) network for expression recognition. This network can extract facial expression features in an unsupervised manner, maximising the acquisition of global information without increasing the parameter.

To reduce computational overhead, a series of compact and efficient models have emerged: MobileNet (Nan et al., 2022) is a lightweight deep neural network that employs depthwise separable convolution; GhostNet (Luo et al., 2024) generates more feature maps through simple operations, extracting relevant information from subtle features at a low computational cost; Wang et al. (2021) employed a dual-path attention network to ensure the algorithm's real-time performance while reducing model parameters and significantly boosting accuracy, all at a lower computational expense.

We incorporate an attention module into the convolutional layer of the CNN to enable the network to focus more on important and useful features, thereby improving recognition efficiency and achieving model light-weighting. Finally, based on this model, we design a FER system using a UI interface in QT. This system can recognise facial expressions from three aspects: images, videos, and live camera feeds, thereby verifying the practicality of the model.

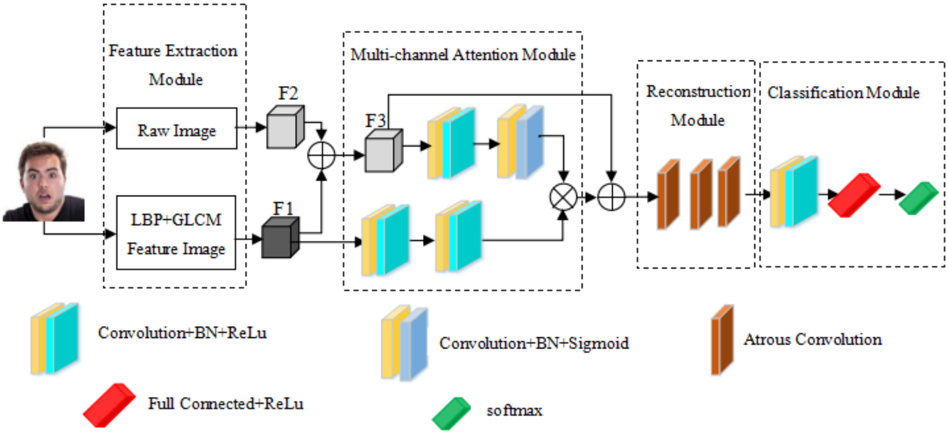
Using CNNs for FER requires significant computer memory and incurs considerable computational costs. However, many researchers focus on improving accuracy while overlooking the limitations of memory and processing speed, especially when deploying models on mobile devices. In this paper, we combine CNNs with LBP+grey level co-occurrence matrix (GLCM) to ensure good recognition performance while maintaining efficiency. Additionally, we incorporate an attention module into the CNN's convolutional layers to enable the network to focus more on important features, thereby improving recognition efficiency and enabling model lightweighting. Finally, based on this model, we design a FER system using a QT-based UI interface. This system can recognise facial expressions from images, videos, and live camera feeds, thus demonstrating the practicality of the proposed model.

2 Mathematical model

A lightweight facial expression classification and recognition model based on LBP+GLCM features, combined with a multi-channel attention mechanism, is proposed to maintain the accuracy of existing models. Firstly, the combination of CNN neural network and LBP+GLCM ensures that the model has a strong recognition capability. Secondly, the multi-channel attention mechanism is introduced to further enhance the classification features extracted by depth-separable convolutions, focusing on the most significant region classification features. The model structure is illustrated in Figure 1.

The model is reconstructed by adding attention mechanism to the convolutional layer. It is divided into four parts: the feature extraction module, multi-channel attention module, reconstruction module and classification module. The feature extraction module consists of two independent CNNS, one for the original image and the other for the LBP+GLCM feature image. A pure convolutional layer is used as the backbone for feature extraction. To prevent the network from becoming too complicated, the first six layers of the VGG-16 network are utilised in this paper to extract features from both the original image and the LBP+GLCM features. The convolution of 1 × 1 × 64 is then used to reduce the LBP+GLCM feature dimension to match that of the original image. The network fuses the feature F1 extracted from the original image with the feature F2 extracted from the LBP+GLCM feature image.

Figure 1 Network model (see online version for colours)



The attention module functions by increasing the weight of useful features, enabling the network to focus more on the features that are essential for expression recognition. This approach allows the network to recognise different expressions more efficiently. After the attention module, the attention graph is adjusted using atrous convolution as the reconstruction module. The reconstruction module consists of three 3 × 3 Atrous with expansion coefficients of 2, 3 and 4, progressing from the bottom to the top. Compared with traditional operators, atrous convolution can achieve a larger receptive field without losing information due to pooling, while maintaining the same computational cost.

The feature map from the attention module contains important information for expression recognition. For each layer, all the feature maps from the previous layers and feature F1 are concatenated as inputs, with the feature maps from the current layer used as inputs for all subsequent layers. The fused feature map F3 performs element summation through the output of the attention module. This architecture not only extracts deeper features but also helps alleviate the issue of reusing useful features. Finally, classification is performed using softmax fully connected layer, which incorporates a batch normalisation (BN) layer after each layer to enhance the stability and speed of neural network training, as well as to mitigate the vanishing gradient problem.

2.1 *Attention mechanism*

To enhance the model's ability to refine multiple feature maps, such as facial images, a multi-channel attention mechanism is embedded into the model, allowing it to focus only on important features in the channel dimension (Gao et al., 2021). The attention module consists of two branches: one is the main branch F_P, which is used to acquire features, and the mask F_m, which integrates LBP+GLCM features to acquire attention graphs F_m. The update formula for the attention module is shown in equation (1).

$$F_{refine} = F_P \otimes F_m \tag{1}$$

Here, ⊗ denotes F_P and F_m, multiplied by the corresponding positional elements.

In the main branch, the input feature F3 is simultaneously max-pooled and globally average-pooled to obtain different spatial feature descriptors, denoted as F_{avg} and $F_{\max}$. Inspired by Wang et al. (2020), for the spatial feature map of CNNs, the correlation between adjacent channels is the greatest. Therefore, a one-dimensional convolution with a kernel length of k is designed to aggregate channel feature information across the attention of one-dimensional channels. The two channel attention feature vectors, F_{avg} and $F_{\max}$ are fused by addition, and the Sigmoid activation function is used to normalise the weights, and the main branch F_P is obtained. The F_P can be calculated according to equation (2).

$$F_P = Sigmoid\left(f\left(F_{avg}\right) + f\left(F_{\max}\right)\right) \tag{2}$$

where f denotes a one-dimensional convolution operation.

The mask branch takes the input the last layer and generates attention weights F_m as shown in equation (3):

$$F_m = Sigmoid\left(W * \left(F_m\right) + b\right) \tag{3}$$

where W is weights, b is biases of the convolution layer, * is the convolution operation. The *Sigmoid* provides a probability distribution in the range of (0, 1).

2.2 Local binary pattern

LBP is an algorithm used for image feature extraction. It determines the pattern category (such as texture, shape, etc.) of a pixel location by comparing the neighbourhood of a certain size around each pixel in the image. The texture features extracted by LBP exhibit grey-level invariance and rotation invariance. LBP consists of two steps: threshold processing and coding (Huang et al., 2011). First, a 3×3 pixel window is defined, and the grey value $H_c(x, y)$ of the central pixel point (x, y) in the pixel window is set as the threshold. This value is then compared with the adjacent eight pixel values $H_1(x - 1, y - 1)$, $H_2(x, y - 1)$, $H_3(x + 1, y - 1)$, $H_4(x + 1, y)$, $H_5(x + 1, y + 1)$, $H_6(x, y + 1)$, $H_7(x - 1, y + 1)$, $H_8(x - 1, y)$. If these eight pixel values are greater than or equal to the threshold, the corresponding neighbour is assigned a value of 1; otherwise, it is assigned 0. In the encoding step, the binary numbers obtained from the threshold processing step are converted to decimal numbers. LBP can be calculated according to equation (4).

$$LBP = \sum_{p=0}^{p\text{-}1} 2^p s\left[H_c(x, y) - H_p(x, y)\right] \tag{4}$$

where $s(x)$ is the symbolic function defined as: $s(x)=\begin{cases}1, & x\geq 0\\ 0, & x<0\end{cases}$; $H_c(x, y)$ is the grey value of the central pixel, and $H_p(x, y)$ is the grey value of the neighbouring pixel.

When the LBP operator is applied, the LBP value of the centre point is first calculated using the above method. Then, all pixels in the image are traversed with a step size of 3 to obtain the corresponding LBP feature image.

2.3 GLCM features

The GLCM reflects image texture features by calculating how often pixel pairs with specific grey levels appear in the image. By counting the occurrence of pixel pairs with a certain intensity values at a specified distance and direction, texture features such as entropy, energy, contrast, and homogeneity are extracted from the GLCM of the image. The GLCM considers pixel pairs $C(x, y)$ and $C'(x + dx, y + dy)$ with grey levels i and j, and calculates the occurrence probability $P(i, j)$ of all pixel pairs based on displacement $d = (dx, dy)$ and grey levels i and j. Seven commonly used features for facial texture analysis are extracted: mean, standard deviation (Std), heterogeneity (Diss), homogeneity (homo), energy, maximum correlation coefficient (Ma), and entropy (Ent) (Wu et al., 2015).

The formulas for feature extraction are given in equations (5) to (11).

$$Mean=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}P(i, j)i \tag{5}$$

$$Std=\sqrt{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}P(i, j)(i-Mean)^2} \tag{6}$$

$$Diss=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}P(i, j)\,|\,i-j\,| \tag{7}$$

$$Homo=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\frac{P(i, j)}{1+(i-j)^2} \tag{8}$$

$$Energy=\sqrt{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}P(i, j)^2} \tag{9}$$

$$Ma=\sum_{n=0}^{N\text{-}1}\sum_{i=0}^{N\text{-}1}\sum_{j=0}^{N\text{-}1}\frac{P(i, n)P(j, n)}{P_x(i)P_y(i)} \tag{10}$$

$$Ent=\sum_{i=0}^{N\text{-}1}\sum_{j=0}^{N\text{-}1}P(i, j)\log P(i, j) \tag{11}$$

Here, $P(i, j)$ is the probability of the pixel pair (i, j), and the logarithm operation is denoted as log. In the literature, the sub-window size is typically 5×5, the grey level range is [0, 255], and the quantisation level is 8.

3 Results

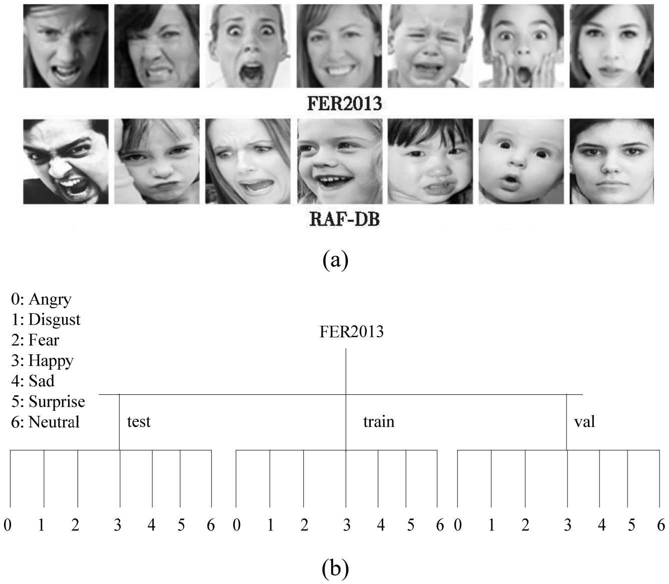
In the experiment, this paper did not expand the number of samples in the datasets. First, the original image was transformed through grey value and size normalisation operation. Then, the LBP operator was used to extract LBP texture features, and GLCM features were also extracted to form the feature vector set. The tensorflow architecture was used to build the CNN model. The dataset was divided into three parts: training set, validation set and test set. After training and testing, to verify the practicality of the model, QT software was used to establish a UI interface. Three methods of network picture, video and real-time screen were employed to validate the effectiveness of the expression recognition method.

3.1 Experimental conditions

Experiments were conducted on two challenging real-world datasets, FER2013 and RAF-DB, both containing seven basic expressions: anger, disgust, fear, happy, sad, surprise, and neutral. These datasets also include factors such as occlusion, lighting and age.

The FER2013 dataset consists of 35,887 greyscale images of size 48×48, as shown in the upper part of Figure 2(a). It is divided into 28,708 training images, with 3,589 images each for the public test and private test sets. The images are stored as pixel data in an Excel file. The dataset was loaded using Python libraries like pandas or NumPy, then divided into training, validation, and testing sets in an 8:1:1 ratio to enhance model evaluation. Pixel values were converted from strings to integer lists, and emotion labels were encoded as numerical values for machine learning purposes. Figure 2(b) shows the datasets directory labels.

Figure 2 Fer2013 and RAF-DB datasets, (a) dataset samples, (b) catalogue label diagram of the dataset



(a)

(b)

RAF-DB is a real-world facial expression dataset comprising approximately 30,000 facial images, all of size 100×100 pixels. Sample images are shown in the lower part of Figure 2(b). Based on provided facial bounding boxes, 15,339 images depicting the seven basic expressions were cropped from the original images for experiments, with 12,271 images in the training set and 3,068 in the testing set.

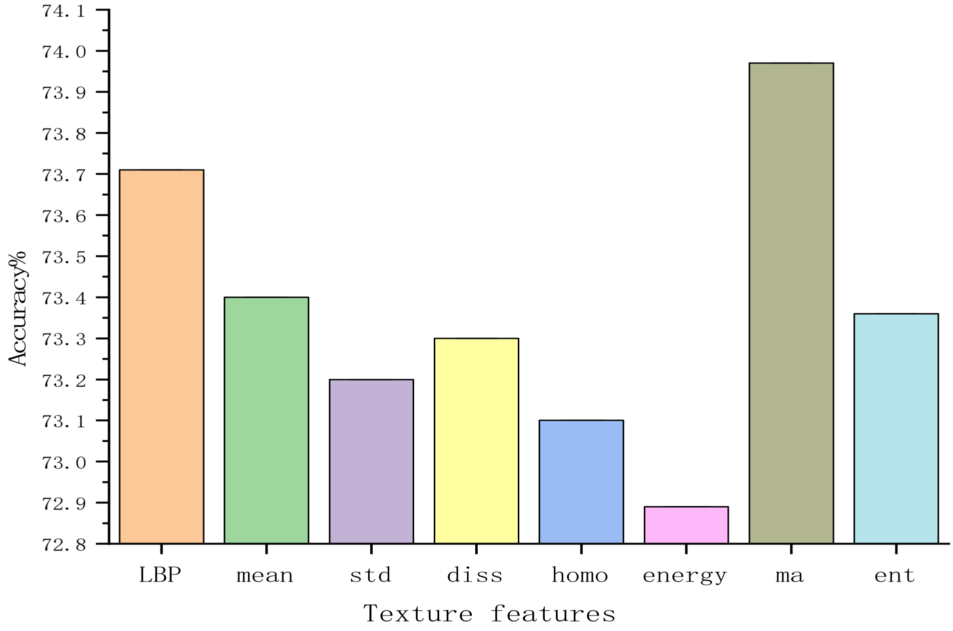
The experiment was conducted on Windows 10 Pro, using an Intel Core i7-4970 CPU with a main frequency of 3.6 GHz and 6 GB of memory. The basic environment for program execution was Python 3.9, with the PyCharm integrated development environment. The version of PyTorch used was 1.7.1. The SGD optimiser was used to dynamically adjust the learning rate during model training, with the initial learning rate set to 0.01. For FER2013 dataset, a learning rate decay was applied every five batches starting from the 80th batch, with a decay rate of 0.9 and a total of 250 training batches. To prevent overfitting, a data augmentation strategy was adopted before training.

3.2 Experimental results

3.2.1 Ablation experiment

To evaluate the effectiveness of adding channel attention mechanism during the training of fused texture feature set, the effectiveness of each texture feature was tested. LBP texture features were extracted respectively and seven types of GLCM texture features: mean, standard deviation, dissimilarity, homogeneity, energy, maximum correlation coefficient, and entropy were incorporated into the feature extraction process. As shown in Figure 3, when GLCM texture features are combined with the maximum correlation coefficient feature, the highest expression recognition accuracy is achieved.

Figure 3 Accuracy of expression recognition using fused GLCM texture features (see online version for colours)



Four types of features, LBP, mean, ma, and ent, were selected to form the feature vector to verify the effectiveness of incorporating the channel attention mechanism. Our method was evaluated and compared with other existing algorithms on the FER2013, as shown in Table 1. The recognition accuracy of our approach is 78.89%, which outperforms four other state-of-the-art algorithms. Additionally, the recognition accuracy decreases when LBP+GLCM and the channel attention mechanism are not used.

Table 1 FER2013 data set comparison of different methods

Methods	*Recognition rate (%)*
Wang et al. (2019)	71.21
Saurav et al. (2022)	72.77
Li et al. (2021)	73.11
This article (excluding LBP+GLCM)	67.73
This article (does not include channel attention machine)	73.96
This article	79.89

3.2.2 Comparative experiment

To further validate the lightweight performance of the proposed model, several classic lightweight models in the field of image classification, including CERN (Gera et al., 2022), EfficientNetB0 (Bodavarapu and Srinivas, 2021) and MobileViT-S (Wang and Zhang, 2024), were selected for comparison across three metrics: model parameters, average predicted time and recognition accuracy. As shown in Table 2, the proposed method achieves the highest accuracy on the FER2013 and RAF-BD datasets, with 79.89% and 86.77%, respectively.

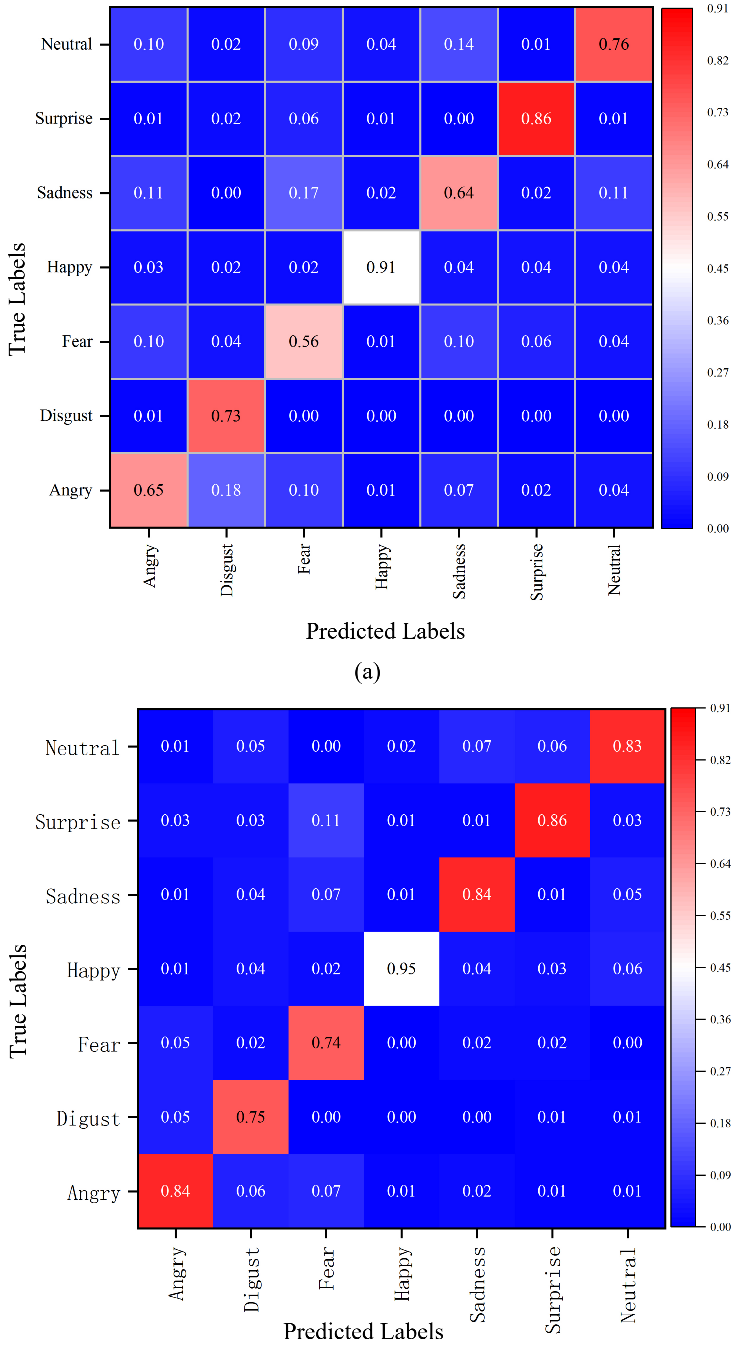
Table 2 Performance comparison of various models

Model	*Parameters/M*	*Average predicted/s*	*FER2013 ACC/%*	*RAF-BD ACC/%*
CERN (Gera et al., 2022)	1.45	5.69	—	84.08
EfficientNetB0 (Bodavarapu and Srinivas, 2021)	5.30	2.05	70.80	84.21
MobileViT-S (Wang and Zhang, 2024)	2.31	0.71	60.30	77.05
This article	1.92	0.72	79.89	86.77

3.2.3 Confusion matrices

Figure 4 shows the confusion matrices of the model on the FER2013 and RAF-BD datasets. As shown in Figures 4(a) and 4(b), the recognition accuracy for the expression ‘happy’ is relatively high, reaching 91% and 95%, respectively. In the FER2013 confusion matrix, the recognition rates for ‘sad’ and ‘fear’ are the lowest, with the probability of ‘fear’ being misclassified as ‘surprise’ at 6%. In the RAF-BD confusion matrix, ‘disgust’ and ‘fear’ have the lowest recognition rates, with ‘disgust’ being misidentified as ‘anger’ at 7%, and ‘fear’ being misclassified as ‘surprise’ at 11%. The expressions ‘disgust’ and ‘anger’ are prone to mutual misclassification, as are ‘fear’ and ‘sadness’.

Figure 4 Confusion matrices on FER2013 and RAF-BD datasets, (a) confusion matrix of FER2013 dataset, (b) confusion matrix of RAF-BD dataset (see online version for colours)



(a)

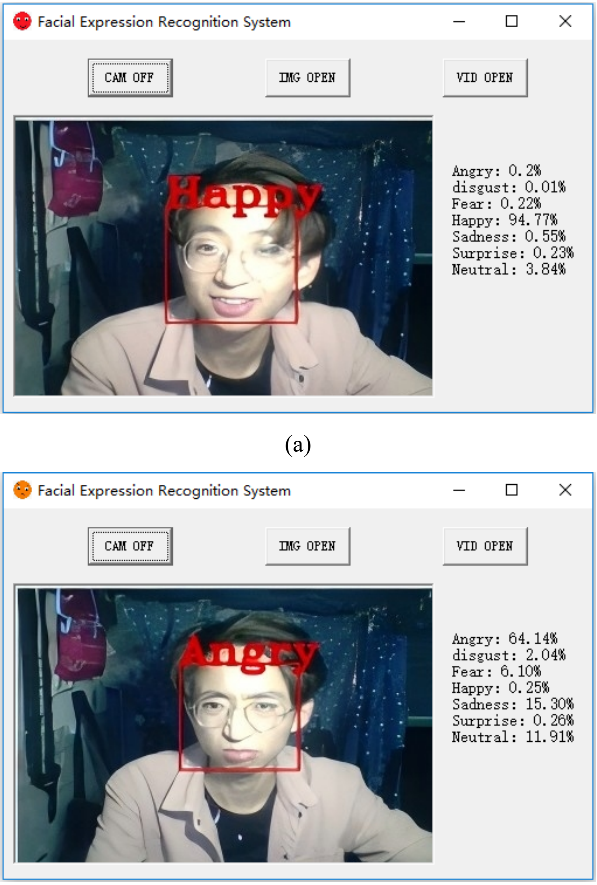
(b)

3.3 Design and implementation of the expression recognition application demonstration interface

To display expression recognition more intuitively, an expression recognition application demonstration interface based on this model was designed. The practicality and applicability of this model are demonstrated and verified through this interface.

The visual interface is designed on Windows 10, using QT Designer 5.14.1 to create the user interface, and PyQt5 to implement the expression recognition and application functions. The three buttons, from left to right on the main interface, are used to open the camera for real-time facial expression detection, select a picture in the file for facial expression analysis, and select a video from the file for facial expression detection. The detection results from these three image sources are shown in Figure 5.

Figure 5 Different expression demonstration interface, (a) the happy expression detected in live camera feeds, (b) the angry expression detected in live camera feeds, (c) the sad expression detected in the image, (d) the neutral expressions detected in the video (see online version for colours)


Facial Expression Recognition System
CAM OFF
IMG OPEN
VID OPEN
Happy

```
Angry: 0.2%
disgust: 0.01%
Fear: 0.22%
Happy: 94.77%
Sadness: 0.55%
Surprise: 0.23%
Neutral: 3.84%
```

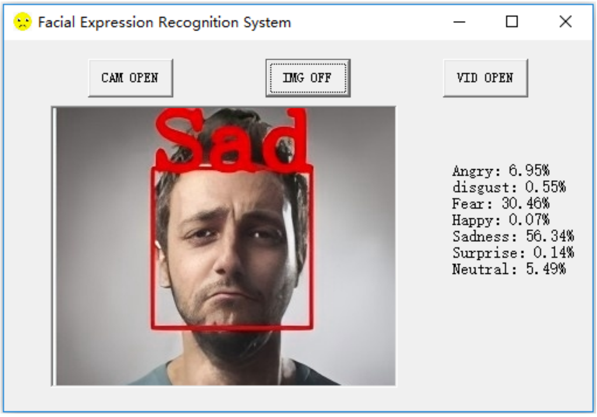
(a)

Facial Expression Recognition System
CAM OFF
IMG OPEN
VID OPEN
Angry

```
Angry: 64.14%
disgust: 2.04%
Fear: 6.10%
Happy: 0.25%
Sadness: 15.30%
Surprise: 0.26%
Neutral: 11.91%
```

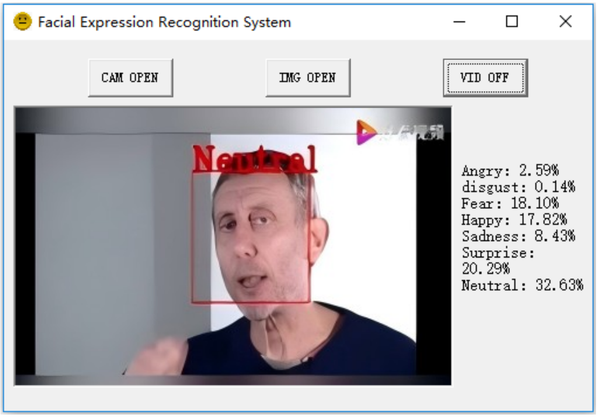
(b)

Figure 5 Different expression demonstration interface, (a) the happy expression detected in live camera feeds, (b) the angry expression detected in live camera feeds, (c) the sad expression detected in the image, (d) the neutral expressions detected in the video (continued) (see online version for colours)



```
Angry: 6.95%
disgust: 0.55%
Fear: 30.46%
Happy: 0.07%
Sadness: 56.34%
Surprise: 0.14%
Neutral: 5.49%
```

(c)



```
Angry: 2.59%
disgust: 0.14%
Fear: 18.10%
Happy: 17.82%
Sadness: 8.43%
Surprise:
20.29%
Neutral: 32.63%
```

(d)

4 Discussion

To analyse the effectiveness of different optimisation strategies in our method, experiments were conducted using CNN as the baseline network on the FER2013 dataset. The results are shown in Table 1. It can be seen that our model achieves an accuracy of 79.89% on the FER2013 dataset, representing an increase of 12.16% compared to FER based solely on fused LBP+GLCM texture features, and an improvement of 5.93% over the model incorporating channel attention mechanisms. This indicates that both the LBP+GLCM and the attention module improve recognition accuracy. Furthermore, our model outperforms methods such as Wang et al. (2019), Saurav et al. (2022), Li et al. (2021), with improvements of 8.68%, 7.12%, and 6.78% respectively, demonstrating the superiority of the proposed approach.

As shown in Table 2, compared with the EfficientNetB0 (Bodavarapu and Srinivas, 2021) network, our method reduces the number of parameters by 63.77% and the average predicted time by 64.88%. Compared with the MobileViT-S (Wang and Zhang, 2024) network, our method reduces the number of parameters by 16.88% and the average predicted time by 1.4%. The proposed method has fewer parameters, moderate average predicted time, and the highest accuracy, which demonstrates the effectiveness of the lightweight model and its generalisability across different datasets.

Figures 4(a) and 4(b) show that during the operation of the FER system, it performs well in recognising 'happy', 'sad', and 'angry' expressions, but often confuses 'fear' and 'surprise', and frequently misidentifies 'disgust' as a neutral expression. Using the facial action coding system (FACS) developed by psychology (Martinez et al., 2017) to explain this phenomenon, one possible reason is that the feature values of expressions such as 'fear', 'surprise', and 'disgust' are similar. Another reason could be the scarcity of training data for expressions like 'disgust' and 'surprise' compared to 'happy', which makes it difficult for the model to distinguish these expressions accurately.

This study has achieved multiple innovative breakthroughs in the field of FER, significantly advancing the theoretical development of this domain. First, by integrating LBP+GLCM, a multi-scale, multi-level texture feature extraction strategy was proposed, effectively enhancing the representation of subtle expressions. Second, inspired by the ECA network, a channel attention mechanism was introduced to strengthen the expressive capacity of key texture channels, improving the model's classification performance and robustness. Third, adhering to lightweight design principles, the model was optimised through parameter pruning, reducing parameters by 63.77% and significantly shortening predicted time by 64.88%, achieving efficient and accurate FER. This enhances the model's practicality and generalisability in mobile and edge computing scenarios.

5 Conclusions

In this paper, we propose a new method for FER that combines LBP+GLCM features with an attention mechanism. In this model, LBP+GLCM are used to extract the local texture features of the original image. These features are then fused with the original image features, and the classification features extracted by the depth separable convolution are further enhanced by the channel attention mechanism. The FER2013 and RAF-BD datasets were used to validate the model, achieving a recognition accuracy of 79.89% and 86.77%, respectively. Finally, to demonstrate the application of the method, an expression recognition application interface was designed to recognise facial expressions from three sources: Images, videos and live camera feeds. Traditional method of extracting manual features may contain more expression information. In future research, other traditional methods will be integrated with deep learning techniques for expression recognition. Additionally, by combining face detection, face alignment and other algorithms, variety of strategies will be adopted to improve the accuracy and robustness of expression recognition methods, enabling practical applications in real-world scenarios.

Acknowledgements

This work was supported by the Key Research and Development Program of Gansu Province-Industrial Project (Grant No. 25YFGM002) and the Campus Cooperation Project (Grant No. HXZK2426 and HXZK2458).

Declarations

All authors declare that they have no conflicts of interest.

References

Alshahrani, A.M., Alshahrani, M.A. and Alghamdi, A.A. (2020) 'Novel deep learning model for facial expression recognition based on maximum boosted CNN and LSTM', *IEEE Access*, Vol. 14, No. 7, pp.1373–1381.

Aouani, H. and Ayed, Y.B. (2024) 'Deep facial expression detection using Viola-Jones algorithm, CNN-MLP and CNN-SVM', *Social Network Analysis and Mining*, Vol. 14, No. 1, p.65.

Bodavarapu, P. and Srinivas, P. (2021) 'Facial expression recognition for low resolution images using convolutional neural networks and denoising techniques', *Indian Journal of Science and Technology*, Vol. 14, No. 12, pp.971–983.

Connie, T., Al-Shabi, M., Cheah, W.P. and Goh, M. (2017) 'Facial expression recognition using a hybrid CNN-SIFT aggregator', *Multi-disciplinary Trends in Artificial Intelligence*, No. 10607, pp.139–149.

Ekman, P. and Friesen, W.V. (1978) 'Facial action coding system (FACS): a technique for the measurement of facial action', *Rivista di Psichiatria*, Vol. 47, No. 2, pp.126–138.

Gao, M., Chen, Y.H. and Zhang, Z.H. (2021) 'Classification algorithm of garbage images based on novel spatial attention mechanism and transfer learning', *System Engineering Theory and Practice*, Vol. 41, No. 2, pp.498–512.

Gera, D., Balasubramanian, S. and Jami, A. (2022) 'CERN: compact facial expression recognition net', *Pattern Recognition Letters*, Vol. 155, pp.9–18.

Gu, Y., Zhang, Y., Huang, H., Zhang, D. and Wang, J. (2016) 'Facial expression recognition via deep learning', *IEEE Transactions on Affective Computing*, Vol. 7, No. 3, pp.201–212.

Hu, M.Q. and Liu, B. (2004) 'Mining and summarizing customer reviews', *10th ACM SIGKDD International Conference on Knowledge Discovery and Date Mining*, pp.22–25.

Huang, D., Shan, C., Ardabilian, M. et al. (2011) 'Local binary patterns and its application to facial image analysis: a survey', *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 41, No. 6, pp.765–781.

Li, H., Wang, N., Yu, Y. et al. (2021) 'LBAN-IL: a novel method of high discriminative representation for facial expression recognition', *Neurocomputing*, Vol. 432, pp.159–169.

Liu, Y., Cao, Y., Li, Y. et al. (2016) 'Facial expression recognition with PCA and LBP features extracting from active facial patches', *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pp.368–373.

Lowe, D. (2004) 'Distinctive image features from scale – invariant key-points', *International Journal of Computer Vision*, Vol. 60, No. 2, pp.91–110.

Luo, Y., Guo, F. et al. (2024) 'Cockpit facial expression recognition model based on attention fusion and feature enhancement network', *Qiche Gongcheng/Automotive Engineering*, Vol. 46, No. 9, pp.1697–1706.

Martinez, B., Valstar, M.F., Jiang, B. et al. (2017) 'Automatic analysis of facial actions: a survey', *IEEE Transactions on Affective Computing*, Vol. 10, No. 3, pp.325–347.

Nan, Y., Ju, J., Hua, Q. et al. (2022) 'A-MobileNet: an approach of facial expression recognition', *Alexandria Engineering Journal*, Vol. 61, No. 6, pp.4435–4444.

Ojala, T., Pietikäinen, M. and Harwood, D. (2002) 'Multiresolution gray scale and rotation invariant texture classification with local binary patterns', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 7, pp.971–987.

Pang, B., Lee, L. and Vaithyanathan, S. (2002) 'Thumbs up sentiment classification using machine learning techniques', *Conference on Empirical Methods in Natural Language Processing*, pp.79–86.

Qin, S., Zhu, Z., Zou, Y. and Wang, X. (2020) 'Facial expression recognition based on Gabor wavelet transform and 2-channel CNN', *International Journal of Wavelets, Multiresolution & Information Processing*, Vol. 18, No. 2, p.2050003.

Salur, M.U. and Aydin, I. (2020) 'A novel hybrid deep learning model for sentiment classification', *IEEE Access*, No. 8, pp.58080–58093.

Saurav, S., Gidde, P., Saini, R. et al. (2022) 'Dual integrated convolutional neural network for real-time facial expression recognition in the wild', *The Visual Computer*, Vol. 38, No. 3, pp.1083–1096.

Vaswani, A., Shazeer, N., Parmar, N. et al. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, pp.5998–6008.

Vishwakarma, V.P. and Mishra, G. (2019) 'A robust multi-level sparse classifier with multi-modal feature extraction for face recognition', *Int. J. Applied Pattern Recognition*, Vol. 6, No. 1, pp.76–102.

Wang, J. and Zhang, Z. (2024) 'Facial expression recognition in online course using light-weight vision transformer via knowledge distillation', *Lecture Notes in Computer Science*, No. 14327, pp.247–253.

Wang, L., He, Z., Meng, B. et al. (2021) 'Two-pathway attention network for real-time facial expression recognition', *Journal of Real-Time Image Processing*, Vol. 18, No. 4, pp.1173–1182.

Wang, Q., Wu, B., Zhu, P. et al. (2020) 'ECA-Net: efficient channel attention for deep convolution neural network', *IEEE Proceedings of CVF Conference on Computer Vision and Pattern Recognition*, pp.11531–11539.

Wang, W., Sun, Q., Chen, T. et al. (2019) *A Fine-grained Facial Expression Database for End-to-end multi-pose Facial Expression Recognition*, ArXiv Preprint, No. 1907, p.10838.

Wu, Q., Gan, Y., Lin, B. et al. (2015) 'An active contour model based on fused texture features for image segmentation', *Neurocomputing*, Vol. 151, No. 3, pp.1133–1141.

Xu, Q. and Sun, B. (2021) 'An expression recognition model based on deep learning and evidence theory', *Computer Engineering & Science*, Vol. 43, No. 4, pp.704–711.