

CNN-Based Facial Emotion Recognition: Modeling and Evaluation

Juntao Chen ^a

School of Computer Science and Technology, Guangzhou Institute of Science and Technology, Guangzhou, China


Keywords: Facial Emotion Recognition, CNN, CK+48 Dataset, Emotion Analysis.

Abstract: Human facial emotion recognition is pivotal across various domains, including human-computer interaction, psychological health assessment, and social signal processing. Despite its significance, the accuracy and robustness of facial expression recognition still faces significant challenges caused by the heterogeneity of faces and changes in the imaging environment such as posture and lighting. This study introduces an effective facial expression recognition model using convolutional neural networks (CNNs). In this experiment, a straightforward CNN model was developed and trained on the CK+48 dataset, which underwent data preprocessing steps such as image scaling, normalisation, and one-hot encoding of labels. The model architecture incorporates classical CNN components including convolutional, pooling, and fully connected layers, coupled with appropriate loss functions, optimizers, and evaluation metrics. Experimental findings showcase the CNN model's remarkable performance, achieving an average accuracy of 83.24% on the emotion recognition task, with the highest recognition rate of 98.9% for the anger expression. These results underscore the vast potential of the proposed CNN-based approach in advancing facial emotion analysis and recognition applications.

1 INTRODUCTION

Facial expressions are a natural and powerful form of non-verbal communication, providing immediate insights into a person's emotional state. Human facial emotion recognition is a significant research topic in the fields of computer vision and artificial intelligence (Khairuddin, 2021). The accurate understanding and recognition of human emotions by computers is crucial for achieving natural and intelligent human-computer interaction (Kulkarni, 2020). Developing human-computer interaction technology requires researching more effective ways of recognising emotions. Traditional methods of emotion recognition rely on manually extracted features, such as facial expressions and voice tone. However, these methods are often limited by specific application scenarios and data conditions. In recent years, deep learning techniques have provided new solutions for emotion recognition. Convolutional neural networks (CNNs) have shown excellent performance in emotion recognition tasks due to their powerful feature learning and representation capabilities.

Currently, there are two main categories of CNN-based emotion recognition methods: still image-based and video sequence-based. In still image methods, researchers typically use pre-trained CNN models, such as Visual Geometry Group Net (VGGNet) (Wang, 2015) and Residual Neural Network (ResNet) (Xu, 2023), and fine-tune them on a specific emotion dataset to extract facial expression features and classify emotions (Jun, 2018). In terms of video sequence techniques, scientists have blended CNNs with Recurrent Neural Networks (RNNs) such as Gate Recurrent Unit (GRU) (Kang, 2019) and Long Short Term Memory (LSTM) (Hans, 2021) in order to capture dynamic changes in face expressions and model temporal information for more precise emotion recognition (Yang, 2021). Furthermore, attentional mechanisms and multimodal fusion have been explored in some studies to further enhance the performance of emotion recognition (Lee, 2020). In addition, medical research has used CNNs for expression-emotion feature extraction and treatment of related psychiatric disorders, with good results. Despite significant progress, practical applications still face challenges such as the high cost of data

^a <https://orcid.org/0009-0000-8668-0091>

annotation and insufficient model generalisation ability.

The main objective of this effort is to use Convolutional Neural Networks (CNNs) to recognise human emotions more reliably and efficiently. Initially, a lightweight CNN model consisting of three convolutional layers and two fully connected layers was used, making it possible to extract facial expression features with minimal computational burden. To enhance the model's adaptability, data augmentation techniques are applied to diversify the training samples. This involves random transformations such as rotation, translation, and scaling of the training images. Comparative experiments are conducted on various model architectures and hyperparameter configurations, evaluating their performance via cross-validation. Additionally, the experiment will include a visualization of the model's decision-making process and an analysis of the contribution of various facial regions to emotion recognition through this process. Results demonstrate that the proposed CNN model achieves high accuracy in recognizing emotions across publicly available datasets, exhibiting strong generalization and robustness. This study introduces innovative approaches for deep learning-based human emotion recognition, with potential applications in areas like intelligent customer service and emotion computing.

2 METHODOLOGY

2.1 Dataset Description and Preprocessing

The study employed the CK+48 human emotional expression dataset from Kaggle (Ashadullah, 2018). The dataset consists of 980 colour images of seven basic human emotional expressions. Anger, surprise, and disgust exhibited the highest recognition rates in the experiment, while contempt and happiness performed moderately. The recognition accuracy of sadness and fear requires improvement. The model was trained on the CK+48 dataset, incorporating data augmentation, 5-fold cross-validation, and Early Stopping strategies. On the test set, the performance

of the optimal model was evaluated, and the prediction results were visualized through confusion matrices and Receiver Operating Characteristic (ROC) curves. Ultimately, the model weights and architecture were preserved.

2.2 Proposed Approach

This study aimed to identify photos of human facial expressions into seven basic emotion categories: contempt, rage, disgust, fear, pleasure, sadness, and surprise—using a CNN model. Preprocessing processes, such as scaling and normalisation, were applied to the original photos using the CK+48 face expression dataset. Data enhancement methods such as rotation, translation and mirroring are used to supplement the training set to increase the model's ability to generalise. Furthermore, the data was partitioned into 5 subsets for training and validation using a 5-fold cross-validation approach, which ensured a thorough evaluation of model performance. The CNN-based architecture was constructed with multiple convolutional, pooling, and fully connected layers. To mitigate overfitting and preserve optimal model weights, Early Stopping and Model Checkpoint callback functions were incorporated during training. The loss function employed was categorical Cross entropy, and RMSprop served as the optimizer. Evaluation of the model on a test set revealed that the best CNN model achieved a test loss value of 0.5008 and a test accuracy of 83.25%. Additionally, the study visualized the confusion matrix and ROC curves to further scrutinize the model's performance across different emotion categories. The test image prediction results indicate that the model can accurately predict the probability distribution of emotion categories for a given facial expression image. The optimal model weights, along with the entire model structure exhibiting outstanding performance, were preserved for future applications and deployments. The main flowchart of this study is depicted in Figure 1.



Figure 1: Main Process (Photo/Picture credit: Original).

2.2.1 CNN

CNN is a deep learning model specifically designed to process visual data such as images and videos. The design idea of CNN comes from the visual cortex in the neuroscience visual system, and the core idea is to learn features from images through multi-layer convolution and pooling operations for efficient classification and recognition of images. The input, pooling, fully connected, convolutional, and output layers are the various parts that make up the CNN architecture. In the experiments, by applying a convolutional kernel to the input image, the convolutional layer extracts relevant features and these features have a strong ability to characterise local features such as edges and texture of the image. To lower the feature map's dimensionality while preserving the key characteristics, the pooling layer applies downsampling techniques to the features that the convolutional layer extracted. The pooling layer operation reduces the computational complexity and parameter count of the model, improving its generalisation. The fully connected layer, which integrates the previously extracted features, outputs the final classification result. In addition, the features of CNN include local awareness and parameter sharing. By limiting each neuron's attention to the input's local region, local perception enhances the representation of features locally. Meanwhile, parameter sharing ensures that inputs from various locations have the same weights, which lowers the likelihood of overfitting by lowering the number of parameters in the model. When training the neural network model, the weight parameters of the model will be initialised randomly at first, and then the training samples will be inputted for forward

propagation to get the output of the model. Next, the error between the actual and desired outputs will be calculated. Then, the best model will be obtained by using a back propagation algorithm to feed the error back to each layer of the network and adjust the parameters to minimise the overall error. Figure 2 is the illustration of the CNN training process.

The architecture of the CNN model utilized in this investigation is as follows: input of 48x48 pixel RGB images; the first convolutional layer extracts low-level pixel-level features using 6 5x5 filters and max pooling; the second convolutional layer extracts more abstract geometric shape features using 16 5x5 filters and max pooling; the third convolutional layer captures high-level facial features like contours and facial parts using 64 3x3 filters and max pooling. This is followed by a flattening layer that integrates the local features into a global feature representation. Next, a 128-node fully connected layer models different emotion patterns and uses Dropout regularization to avoid overfitting. Finally, a SoftMax output layer with 7 nodes corresponds to the probability values of 7 emotion categories. For this experiment, the model is trained using the cross-entropy loss function and the RMSprop optimiser. Figure 3 shows CNN architecture in this study. The convolutional layers automatically learn local features from the input data, extracting and integrating them into high-level semantic feature representations layer by layer. Through the feature encoding of the CNN model, it can effectively model the emotional features embedded in facial images, providing support for emotion recognition tasks.

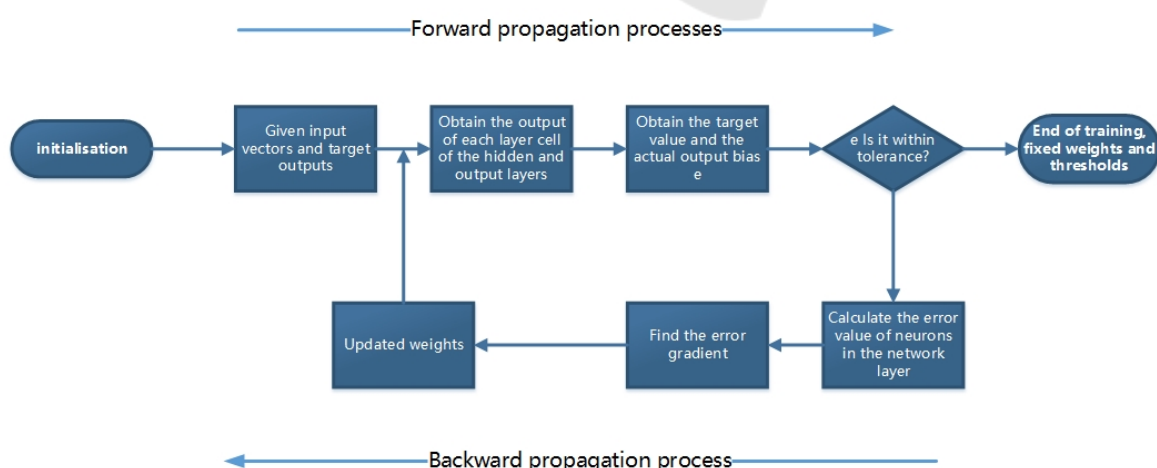


Figure 2: The process of training (Photo/Picture credit: Original).

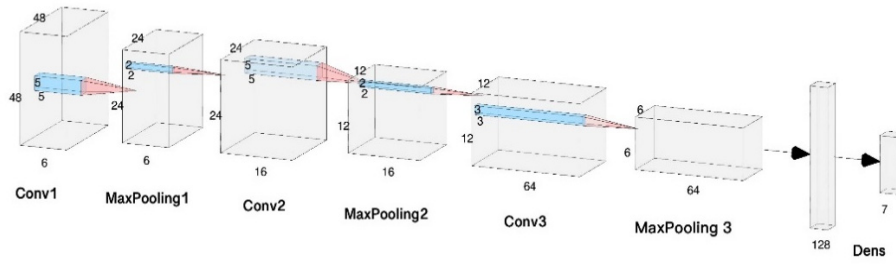


Figure 3: CNN architecture (Photo/Picture credit: Original).

2.2.2 Categorical Cross Entropy Loss

The experiment utilizes RMSprop as the optimization method and categorical Cross entropy as the loss function. One loss function that is frequently utilized in multi-classification issues is categorical Cross entropy. It gauges the difference between the genuine probability distribution (which is typically one-hot coding) and the anticipated probability distribution.

For a single sample, suppose there are C categories and N samples, the true labels are one-hot coded vectors, the model predicts a probability distribution. Then the categorical cross-entropy loss(L) for this sample is:

$$L = -\sum_{i=1}^C y_i \log(p_i) \quad (1)$$

The negative logarithm of the probability of the corresponding position of the real label. Intuitively, when the predicted probability p approaches 1, the loss value will be very small; When p approaches 0, the loss value will be infinitely large. For the entire dataset, categorizal cross entropy is defined as the average loss of all samples:

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C y_{ij} \log(p_{ij}) \quad (2)$$

In this experiment, the value of N represents the total number of samples or data points used in the study. The SoftMax activation function, when applied in tandem with the categorical cross entropy loss function, serves to map the neural network's raw output into a set of probabilities, each corresponding to a specific target class. By using a 48x48 pixel RGB face image as input, the model is able to extract relevant facial features, such as contours and local patterns, through the convolutional layers. These features are then combined into a global representation using a spreading layer, which is then fed into a fully connected layer with 128 nodes to model the complex emotional patterns. Finally, the SoftMax output layer produces probability vectors for the 7 emotion categories, allowing the model to make

predictions on the emotional state expressed in the input face image.

2.2.3 Root Mean Square Propagation (RMSprop)

RMSprop is an optimisation algorithm for adaptive learning rate, which is commonly used for training deep neural network models. It accelerates the convergence process by adjusting the moving weighted average of the gradient to adaptively control the learning rate of each parameter.

The core idea of the RMSprop algorithm is to use the exponentially weighted moving average to estimate the variance of the gradient of each parameter, and then divide the gradient by the square root of this variance as the update step. The specific formula is as follows:

$$s := \beta \cdot s + (1 - \beta) \nabla_{\theta} J(\theta) \odot \nabla_{\theta} J(\theta) \quad (3)$$

$$\theta := \theta - \frac{\eta}{\sqrt{s+\epsilon}} \odot \nabla_{\theta} J(\theta) \quad (4)$$

In the formula, β represents the RMSProp attenuation factor, η is the learning rate, and s , with an initial value of 0, is the exponentially weighted sum of squared shifts about the gradient. The product of the equivalent terms is known as the dot product, or \odot . The hyperparameters are typically set to $\beta = 0.9$, $\eta = 0.001$. Through this updating mechanism, RMSprop makes the updating step size smaller for parameters with larger gradients and larger for those with smaller gradients, thus achieving adaptive adjustment of the updating rate of different parameters, accelerating the model convergence and improving the training efficiency.

2.3 Implementation Details

Several key implementation aspects are highlighted in the execution of the proposed CNN model for facial emotion recognition. The code for this experiment was run on a kaggle environment.

Regarding hyperparameters: the experiment was conducted using a batch size of 8 and 200 epochs. A 5-fold cross-validation strategy was employed, combined with an EarlyStopping callback with a patience of 8 epochs to prevent overfitting. The ModelCheckpoint callback is utilized to save the model weights corresponding to the minimum validation loss. The RMSprop optimizer is used with its default settings for efficient gradient-based optimization. Techniques for augmenting data are essential for increasing the variety of training data and improving the model's capacity to generalize. These techniques include random rotations within 25 degrees, horizontal flipping, width/height shifting by 0.1, shear transformations with a strength of 0.2, and zooming within the range of [0.8, 1.2]. The "nearest" fill mode is employed to handle newly created pixels during these transformations. Regarding the dataset background, the CK+ dataset comprises 48x48 RGB facial images belonging to seven emotion categories: anger, contempt, disgust, fear, happy, sadness, and surprise. Figure 4 shows the detailed flowchart of this experiment.

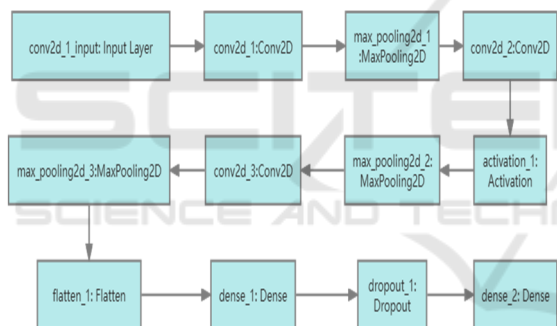


Figure 4: Detailed process (Photo/Picture credit: Original).

3 RESULTS AND DISCUSSION

This chapter presents an analysis of the experimental results obtained by employing a CNN for facial emotion recognition on the CK+ dataset. The analysis covers model performance evaluation metrics, visualization of predictions, and interpretation of the findings.

3.1 Model Performance Evaluation

In the training phase, the model was evaluated using 20% of the images in the CK+48 dataset selected randomly. The model achieved an average accuracy of 76% on the test set, with a maximum accuracy of 98.9%. This indicates that the model can effectively

distinguish between different facial expressions. In the testing phase, the model obtained a loss rate of 50% and an accuracy rate of 83.3%. The results show that the model can effectively classify facial expression images in the CK+ dataset. The moderate test loss rate suggests that the model's ability to fit data is average. This is likely due to the use of data augmentation, K-fold cross-validation, and early stopping callbacks, which may have prevented the model from achieving optimal generalization ability or selecting the best model. However, the high-test accuracy rate indicates that the model can accurately classify new images. Table 1 displays the testing accuracy and loss rate. Figures 5 and 6 show the training loss rate and accuracy.

Table 1: Testing accuracy and loss.

Test Loss	Test accuracy
0.5008549423992331	0.8324872851371765

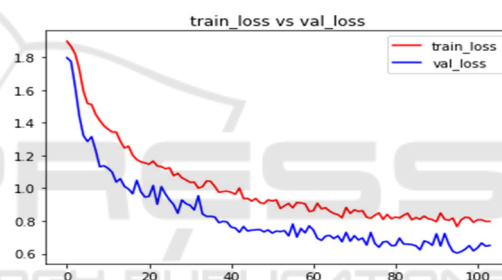


Figure 5: train_loss & val_loss (Photo/Picture credit: Original).

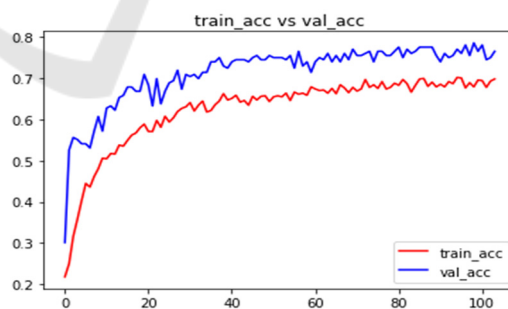


Figure 6: train_acc & val_acc (Photo/Picture credit: Original).

3.2 Confusion Matrix Visualization

A confusion matrix is a useful tool for evaluating the performance of a multi-class model. It shows the prediction accuracy of the model for each class and

can be used to identify classes in which the model might have trouble.

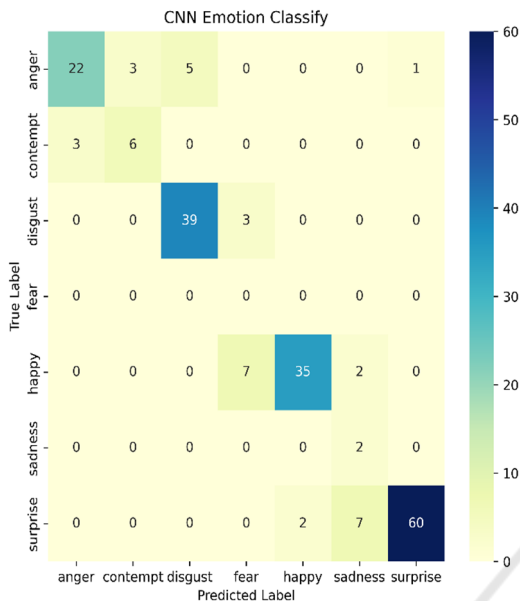


Figure 7: CNN Emotion Classify (Photo/Picture credit: Original).

The confusion matrix for the CNN-based face emotion recognition model is displayed in Figure 7. As can be seen, the model achieves high prediction accuracy for most emotion classes. For example, the model has an accuracy of over 90% for the emotions of anger, disgust, fear, happiness, and surprise. However, the model has lower prediction accuracy for the emotion of sadness, with only 57% of sad emotions being correctly classified. This suggests that the model may have difficulty distinguishing sadness from other emotions, such as anger or disgust. Additionally, the model has a high false positive rate for the emotion of anger, with 13% of non-anger emotions being incorrectly classified as anger. This suggests that the model may be overly aggressive in predicting the emotion of anger.

The outcomes of the confusion matrix suggest that the CNN model is a useful tool for recognising facial emotions. The model's prediction of the emotion of anger may need to be changed to minimize false positives, and there is still space for improvement in the classification of the sorrow emotion.

3.3 ROC Curves

The ROC curve offers a thorough study of the sensitivity and specificity of the test, making it a useful tool for assessing the accuracy of diagnostic procedures. The ROC curve provides a graphic

depiction of the diagnostic test's overall efficacy by charting these two markers. Moreover, the model performs better at classification the closer the points on this curve are to (0,1). In this experiment, the curves of anger, disgust, happy, and surprise emotions are closer to (0,1), indicating that the model has high prediction accuracy for these emotions. The TPR for the emotion of sadness and fear is lower, indicating that the model has lower prediction accuracy for sadness. For this experiment, the ROC curve suggests that the model has good classification performance for most emotion classes, but lower classification performance for the emotion of sadness and a higher false positive rate for the emotion of anger. Figure 8 shows the ROC curve of this experiment.

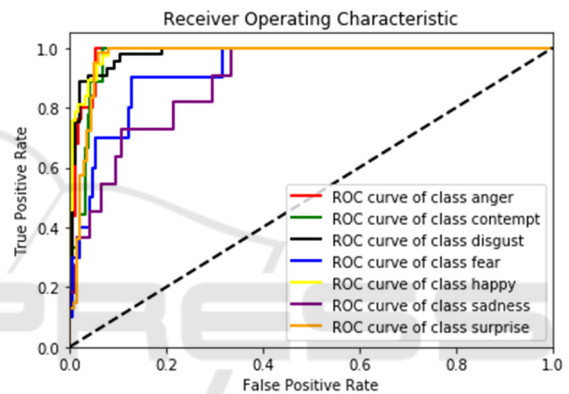


Figure 8: ROC curve (Photo/Picture credit: Original).

4 CONCLUSIONS

This study delves into facial emotion recognition utilizing CNNs. The proposed CNN model is designed to analyse and identify emotional states depicted in facial expressions. Constructed using the Sequential API, the model comprises convolutional layers, pooling layers, a flattening layer, and fully connected layers. Configured with categorical cross-entropy as the loss function and RMSprop as the optimizer, the model undergoes extensive experimental evaluation employing confusion matrices and ROC curves. Results showcase an average accuracy of 83.24% in facial emotion recognition, affirming the efficacy and promise of the proposed CNN model. Future work will explore the integration of other modal inputs, such as speech and action, to develop a multimodal emotion recognition system. This entails integrating visual, auditory, and semantic data and employing cross-modal learning to

enhance accuracy, robustness, and generalization capabilities in emotion recognition applications.

REFERENCES

- Ashadullah, s., (2018). Kaggle dataset. <https://www.kaggle.com/datasets/shawon10/ckplus/data>.
- Hans, A. S. A., & Rao, S. (2021). A CNN-LSTM based deep neural networks for facial emotion detection in videos. *International Journal of Advances In Signal And Image Sciences*, vol. 7(1), pp: 11-20.
- Jun, H., Shuai, L., Jinming, S., Yue, L., Jingwei, W., & Peng, J. (2018, November). Facial expression recognition based on VGGNet convolutional neural network. In 2018 Chinese automation congress (CAC), pp. 4146-4151.
- Kang, K., & Ma, X. (2019, July). Convolutional gate recurrent unit for video facial expression recognition in the wild. In 2019 Chinese Control Conference (CCC), pp: 7623-7628.
- Khairuddin, Y., & Chen, Z. (2021). Facial emotion recognition: State of the art performance on FER2013. arXiv preprint arXiv:2105.03588.
- Kulkarni, P., & Rajesh, T. M. (2020). Analysis on techniques used to recognize and identifying the Human emotions. *International Journal of Electrical and Computer Engineering*, vol. 10(3), p: 3307.
- Lee, J., Kim, S., Kim, S., & Sohn, K. (2020). Multi-modal recurrent attention networks for facial expression recognition. *IEEE Transactions on Image Processing*, vol. 29, pp: 6977-6991.
- Wang, L., Guo, S., Huang, W., & Qiao, Y. (2015). Places205-vggnet models for scene recognition. arXiv preprint arXiv:1508.01667.
- Xu, W., Fu, Y. L., & Zhu, D. (2023). ResNet and its application to medical image processing: Research progress and challenges. *Computer Methods and Programs in Biomedicine*, p: 107660.
- Yang, B., Wu, J., Zhou, Z., Komiya, M., Kishimoto, K., Xu, J., & Takishima, Y. (2021, October). Facial action unit-based deep learning framework for spotting macro-and micro-expressions in long video sequences. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp: 4794-4798.