Facial Expression Recognition Based on Various Deep Neural Networks

Group 11: Yuan Zhao
Khoury College of Computer Sciences
Northeastern University
Boston, United States of America
zhao.yuan2@northeastern.edu

***Abstract*—Facial expression recognition is crucial for enhancing human-computer interaction in computer vision. Traditional approaches depend on handcrafted features and conventional machine learning, but they are limited by environmental factors and issues of accuracy. The introduction of deep learning has transformed facial expression recognition by facilitating the concurrent extraction of features and classification of images. This paper introduces a real-time model employing depthwise separable convolutional networks to address the challenges and inefficiencies associated with traditional Convolutional Neural Networks (CNN) in recognizing facial expressions and gender. Initially, a Multi-Task Convolutional Neural Network (MTCNN) is utilized to detect faces at various scales, supplemented by Kernelized Correlation Filters (KCF) for enhanced detection speed. The efficacy of three neural network architectures is evaluated using the FER-2013 dataset, achieving accuracies of 57.15%, 58.83% with VGG, and 60.00% with ResNet.**

***Index Terms*—facial expression, FER-2013 dataset, convolutional neural networks, VGG, ResNet**

I. Introduction

As sensing technologies and artificial intelligence advance, human feature detection and recognition, particularly through facial expressions, have become a research hotspot, vital for enhancing human-computer interactions. Facial features are key communicators of rich emotional data and gender specifics, making image processing and deep learning essential for facial expression and gender recognition across domains such as smart education, public safety monitoring, and telemedicine. Facial expression recognition, broadly an interdisciplinary field, involves computer vision, image processing, and psychology. It plays a critical role in human-computer interaction by: 1) improving the understanding of human emotions to enhance interaction experiences; 2) enabling tracking and recognition of facial expressions in video streams; and 3) studying facial expression encoding patterns, which aids in the efficient transmission and storage of images containing facial expressions. The applications of facial expression recognition are vast, spanning security, psychology, healthcare, customer satisfaction analysis, and online education, indicating its broad prospective utility.

II. Related work

In the previous works, Jeon et al. [1] employed Histogram of Oriented Gradients (HOG) features for face detection to mitigate the impact of uneven lighting on expression recognition, achieving a 70.7% recognition rate on the FER-2013 dataset with SVM; yet, this method exhibits weak interference resistance and poor adaptability. Zhang Yanliang et al. [2] proposed identifying micro-expressions by linking seven local regions related to facial keypoints into a feature vector, though this approach suffers from low recognition rates in localized areas. Luo Zhenzhen et al. [3] used Conditional Random Forests and Support Vector Machines (SVM) to detect facial features indicative of smiling emotions. Dai Yixiang et al. [4] utilized wearable devices to gather bio-signals such as EEG, pulse, and blood pressure, analyzing multimodal emotions with sparse autoencoding methods. However, due to the high costs associated with equipping each participant with devices, this method is not feasible for large-scale use.

Currently, effective methods for image classification and object detection in natural scenes predominantly involve traditional machine learning and Convolutional Neural Networks (CNN). Traditional machine learning approaches typically use manually designed features and apply classifier algorithms for expression determination. Typical methods for facial feature extraction include Principal Component Analysis (PCA) [5], Local Binary Pattern (LBP) [6], Gabor Wavelet Transform [7], and Scale Invariant Feature Transform (SIFT) [8]. Commonly used classification techniques include Hidden Markov Models (HMM) [9] and K-Nearest Neighbor (KNN) algorithms [10].

Compared to traditional machine learning, deep neural networks can autonomously learn features, reducing the incompleteness caused by manually designed features. Tang [11] proposed combining CNN with SVM and replaced the cross-entropy loss minimization typically used in fully connected CNNs with standard hinge loss to minimize margin-based losses, achieving a 71.2% recognition rate on their test set. MobileNet-V2 [12] utilizes multi-scale kernel convolution units primarily based on depthwise separable convolutions, with linear bottleneck layers in branches, achieving a 70.8% recognition rate for facial expression classification. Li et al. [13] introduced a novel depth-local CNN approach aimed at enhancing discriminability between facial expression categories by maximizing inter-class differences while maintaining local compactness. Kample et al. [14] improved facial expression recognition accuracy by constructing cascaded CNNs. Xu Linlin et al. [15] addressed issues of prolonged network

training times by proposing a facial recognition method based on parallel convolutional neural networks, achieving a 65.6% accuracy. CNNs are often used as black boxes that obscure learned features, complicating the trade-off between classification accuracy and unnecessary parameter count. To address this, Szegedy et al. [16] suggested using guided gradient backpropagation for real-time visualization to validate the features learned by CNNs. Recognizing expressions such as 'anger', 'disgust', 'fear', 'happiness', 'sadness', 'surprise' and 'neutral' in the FER-2013 dataset [16] is challenging, requiring models for facial analysis and gender recognition to exhibit robustness and high computational efficiency.

III. Methods

A complete real-time model for facial expression recognition encompasses three stages: face detection, feature extraction, and classification. Multi-Task CNN To meet the high accuracy and fast response requirements of practical applications, the Multi-Task CNN (MTCNN) network is employed for face detection in input images. Subsequently, the Kernelized Correlation Filter (KCF) tracker is utilized for precise facial tracking and localization. The facial images are then normalized and input into a depthwise separable convolutional neural network for classification. Ultimately, the networks for expression and gender recognition are parallelly integrated. Figure 1 illustrates the overall framework of the real-time facial expression recognition model.

A. face detection

MTCNN The MTCNN algorithm employs an image pyramid to adapt to facial images of varying scales. The algorithm consists of a three-tiered cascaded network: the Proposal Network (P-Net) for rapid candidate window generation, the Refine Network (R-Net) for high-precision candidate window filtering and selection, and the Output Network (O-Net) for producing the final bounding boxes and facial detection points. The network, progressing from coarse to fine, aligns faces at different angles by reducing the number of convolution kernels and size, increasing network depth, and incorporating candidate boxes with classification for fast and efficient face detection.

KCF The integration of the Kernelized Correlation Filters (KCF) tracking algorithm not only addresses the challenges of detecting faces with varying angles and occlusions in practical applications but also enhances face detection speed. This algorithm utilizes a circulant matrix of the region surrounding the target to gather positive and negative samples, employs ridge regression to train the detector, and converts matrix operations into vector Hadamard products—element-wise multiplication in the Fourier space—thereby reducing computational load. Initially, the MTCNN algorithm detects the face, passing the coordinate information to the KCF tracking algorithm, which uses it as the basis for the face detection sample box. Employing a tracking strategy of detecting one frame and tracking five, the last detected face frame is updated to refresh the MTCNN model, preventing tracking losses.

B. Deep Neural Network Models

CNN The proposed Convolutional Neural Network (CNN) model, designed for facial expression recognition, exhibits a structured hierarchy of layers to process and classify input images effectively. The network architecture is composed of sequential layers detailed as follows:

1) **Convolutional Layers**:
   Conv1: 64 filters (3x3), stride 1, padding 1. Batch normalization and ReLU activation are applied. Max-pooling (2x2 window, stride 2) reduces spatial dimensions. *Conv2*: 128 filters (3x3), same padding. Batch normalization and ReLU activation are applied. Max-pooling reduces feature map size. *Conv3*: 256 filters (3x3), same padding. Batch normalization and ReLU activation are applied. Max-pooling further decreases feature dimensions.
2) **Weight Initialization**: Gaussian distribution is used to initialize parameters for *Conv1*, *Conv2*, and *Conv3*.
3) **Fully Connected Layers**: Dropout of 0.2 is applied to reduce overfitting. Linear transformation flattens feature maps to a vector and connects to 4096 neurons with ReLU activation. Dropout of 0.5 is applied before subsequent linear transformations, reducing neuron count to 1024, then 256. Final layer with 7 neurons corresponds to the emotion categories.

VGG-Inspired CNN Architecture The VGG-inspired convolutional neural network (CNN) designed for facial expression recognition utilizes a hierarchical structure to process and classify input images effectively. The network architecture consists of multiple layers as outlined below:

1) **Convolutional Blocks**: *Block 1*: Consists of two convolutional layers with 32 output channels, each using 3x3 filters, stride 1, and padding to maintain spatial dimensions. Batch normalization and ReLU activation are applied, followed by a max-pooling layer with a 2x2 window and stride 2 to reduce spatial dimensions.
   Block 2 and Block 3: Each block contains three convolutional layers with increasing output channels from 32 to 64, and then to 128. All layers use 3x3 filters with padding. Batch normalization and ReLU activation are applied in each layer. Each block concludes with a max-pooling step, similar to Block 1, to further reduce feature map sizes.
2) **Fully Connected Layers**: This section integrates dropout (rate 0.5) to mitigate overfitting, followed by a linear transformation to flatten the feature maps from the final convolutional block into a vector. This vector connects to two hidden layers, each consisting of 4096 neurons, with ReLU activation. Subsequent linear transformations reduce the neuron count to 1024, then to 256, and finally to the output layer with 7 neurons corresponding to the facial expression categories.

The architecture is meticulously engineered to enable robust and precise recognition of facial expressions, leveraging its depth and uniformity for optimal performance.

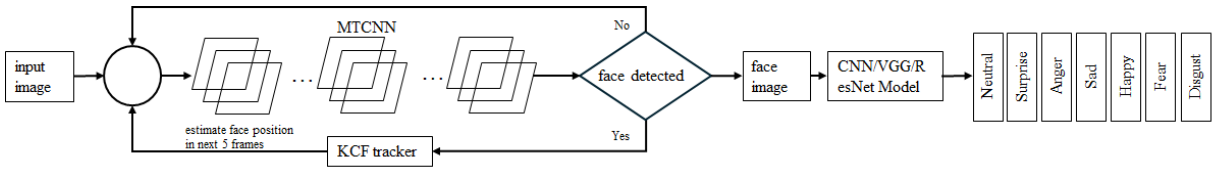

Fig. 1. Facial expression recognition framework

ResNet Architecture The ResNet model, tailored for facial expression recognition, leverages a deep architecture that efficiently addresses the vanishing gradient problem common in very deep networks. The design incorporates 18 layers, including convolutional layers, residual blocks, and fully connected layers, structured as follows:

1) **Initial Convolutional Layer**: Begins with a 7×7 convolutional layer, followed by batch normalization, ReLU activation, and a 3×3 max-pooling layer with a stride of 2 for down-sampling.
2) **Residual Blocks**: Comprises four sets of residual blocks, each containing two residual units. These units maintain the output size while progressively increasing the channel count from 64 to 512. Each residual unit includes:
   - Convolutional layers with 3×3 kernels and padding of 1
   - Batch normalization and ReLU activation
   - Shortcut connections to facilitate gradient flow during back propagation
3) **Global Average Pooling**: Follows the residual blocks with a global average pooling layer that reduces spatial dimensions to a 1×1 grid.
4) **Output Layer**: Concludes with a fully connected layer with 7 neurons for classifying into facial expression categories.

This configuration not only enhances feature extraction and classification capabilities but also supports the training of much deeper networks without performance degradation. ResNet's introduction of residual connections revolutionizes the training dynamics, enabling significantly deeper networks that offer improved performance and convergence in facial expression recognition tasks.

IV. Experiments and Results

A. Dataset

This project focuses on facial expression recognition research based on Convolutional Neural Network (CNN) models. To maximize the accuracy of expression recognition, extensive training and optimization are required, utilizing the FER2013 dataset. This dataset comprises 35,886 facial expression images, including 28,708 training images and 3,589 images each for public and private validation, featuring seven distinct expressions. Originating from the 2013 Kaggle competition, this dataset, primarily sourced through web scraping, exhibits inherent inaccuracies with a human accuracy rate of approximately $65\% \pm 5\%$.

B. Process dataset

To ensure consistency in facial size and position, preprocessing of images in the dataset is necessary. The primary preprocessing steps include face detection, face alignment, and image size normalization. During preprocessing, images are normalized to 48×48 pixels.

In conducting research on facial expression recognition with depthwise separable convolutional neural networks, initializing network weights is critical. The *gaussian_weights_init()* function is employed to ensure that the weights of the neurons are randomized initially. This function assigns weights to each convolutional layer from a normal distribution with a mean of 0 and a standard deviation of 0.04. This approach follows the standard practice of initializing weights to be neither too large nor too small to avoid the problems of vanishing and exploding gradients. By sampling each neuron's weight vector from a multi-dimensional Gaussian distribution, the initialization ensures that neurons operate in diverse directions within the input space, thus preventing gradient issues and improving training speed and efficiency.

C. Models training

In the training of deep learning models, distinct behaviors are observed across architectures. The CNN model demonstrates that validation accuracy reaches a plateau by the 17th epoch.

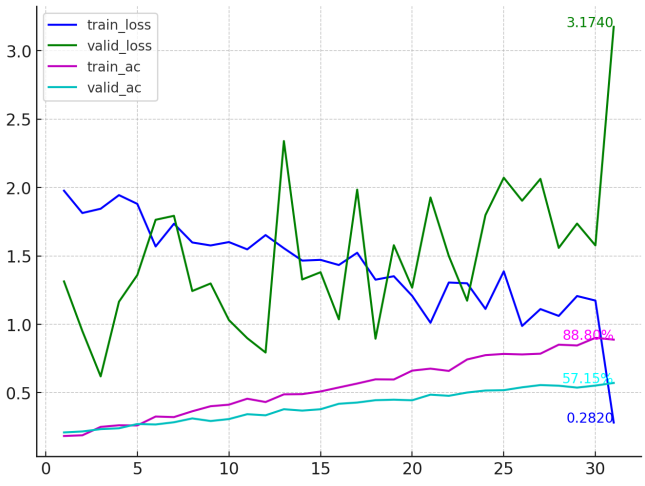

Fig. 2. Training result of CNN network

In contrast, the VGG model underwent an initial training phase of 30 epochs, followed by a subsequent 60 epochs,

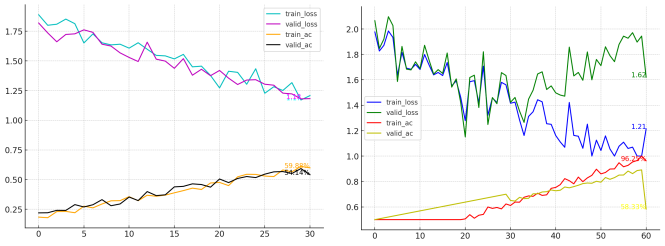

Fig. 3. Training result of VGG network

leveraging its architectural advantage to deepen the network rapidly with identical modules, which potentially improves learning outcomes. This extended training regimen significantly increased training accuracy and decreased training loss, although no significant changes in these metrics were observed after the initial 30 epochs.

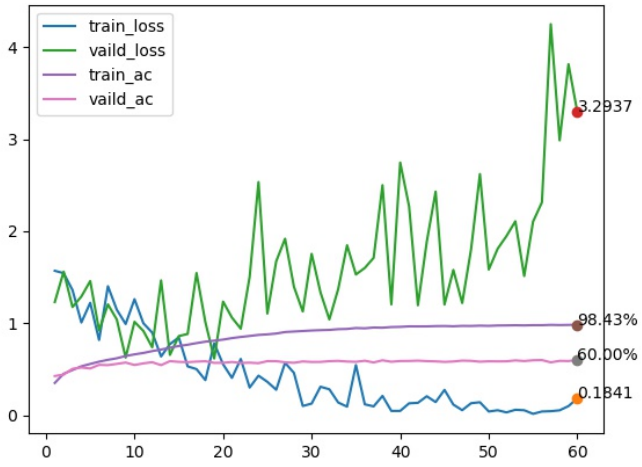

Fig. 4. Training result of ResNet network

Meanwhile, the ResNet model's use of residual blocks led to better data fitting, showcasing the benefits of its advanced structural features in enhancing model performance.

D. Recognition result

The face recognition process initially employed the default OpenCV *haar cascade*, which, however, resulted in misidentifications that significantly affected the results of face expression recognition, as illustrated in Figure 5. Subsequently, this project implemented the *MTCNN* for face recognition and the *KCF* tracker to enhance the robustness of the face recognition processing.

As can be seen from Table I, the recognition rate for happy expressions using the method described in this paper is 81%. This high accuracy is primarily due to the facial features of happy expressions being more pronounced compared to other expressions, resulting in a higher probability of correct classification during the function categorization process. The recognition rates for sad and neutral expressions are 71% and 70%, respectively. Conversely, the recognition rates for

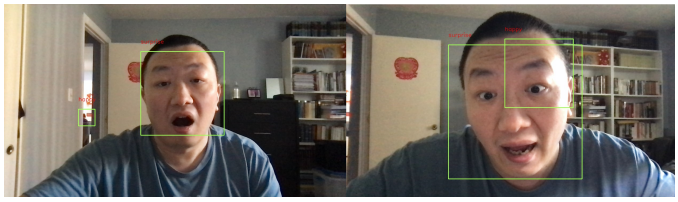

Fig. 5. Misidentified by using default OpenCV "haar cascade"

anger and fear expressions are relatively lower at 59% and 61%, respectively, where misclassifications are more likely to occur, as shown in Fig. 6. The reason for this is that both expressions involve considerable facial movements during feature extraction and learning processes, which may produce similar facial features.

TABLE I
CONFUSION MATRIX OF EXPRESSION RECOGNITION ON FER-2013 DATASET

Actual / Predicted	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral
Anger	0.59	0.04	0.21	0.04	0.12	0.03	0.07
Disgust	0.06	0.69	0.05	0.06	0.03	0.03	0.04
Fear	0.03	0.02	0.61	0.04	0.13	0.02	0.15
Happy	0.03	0.01	0.03	0.81	0.03	0.03	0.06
Sad	0.02	0.02	0.11	0.02	0.71	0.10	0.06
Surprise	0.04	0.01	0.21	0.06	0.03	0.61	0.04
Neutral	0.08	0.01	0.07	0.07	0.05	0.02	0.70

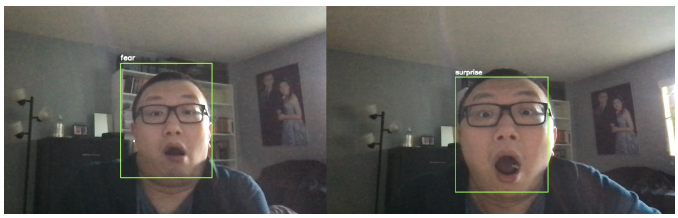

Fig. 6. Misclassification of facial expression

E. Comparison of Results Across Different Models

Comparing recognition results between CNN, VGG, and ResNet, no significant differences were observed. According to previous work [17], the most substantial factor influencing the recognition rate is the choice of dataset. Utilizing the same model, the validation accuracy on the FER2013 dataset is approximately 79%, while it exceeds 95% for the CK+ datasets. Furthermore, in my experiments, I found that the FER2013 database does not perform strongly in recognizing Asian faces, which may be attributed to the database's predominance of Caucasian and European subjects.

V. CONCLUSION

Addressing issues such as complexity, time consumption, and poor real-time performance in the training of convolu-

tional neural networks, this paper proposes a real-time facial expression and gender recognition method based on depthwise separable convolutional neural networks. The method utilizes MTCNN combined with KCF for face detection and tracking. By incorporating convolution blocks and residual blocks into the VGG and ResNet models, it is possible to enhance both training and validation accuracies. Ultimately, the model achieves a recognition rate of 60.0% on the FER-2013 dataset. The processing time for a single facial image is (0.22 ± 0.05) ms, with an overall processing speed of 80 frames per second. Experimental results demonstrate that the proposed model can be effectively used for multi-class classification while maintaining real-time prediction capabilities; it can perform face detection, and emotion classification within a single integrated module. Future work will expand the types of emotions recognized, enlarge the expression database, and train on datasets from real-world scenarios to further enhance recognition accuracy.

REFERENCES

[1] J. Jeon, J. C. Park, Y. Jo, et al., "A real-time facial expression recognizer using deep neural network," in *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, New York: ACM, 2016, No. 94.

[2] Y. L. Zhang, B. Lu, X. P. Hong, et al., "Micro-expression recognition based on local region method," *Journal of Computer Applications*, vol. 39, no. 5, pp. 1282-1287, 2019.

[3] Z. Z. Luo, J. Y. Chen, L. Y. Liu, et al., "Conditional random forests for spontaneous smile detection in unconstrained environment," *Acta Automatica Sinica*, vol. 44, no. 4, pp. 696-706, 2018.

[4] Y. X. Dai, X. Wang, P. Dai, et al., "Stacked auto-encoder optimized emotion recognition in multimodal wearable biosensor network," *Chinese Journal of Computers*, vol. 40, no. 8, pp. 1750-1763, 2017.

[5] Y. Luo, T. Zhang, Y. Zhang, "A novel fusion method of PCA and LDP for facial expression feature extraction," *Optik – International Journal for Light and Electron Optics*, vol. 127, no. 2, pp. 718-721, 2016.

[6] P. Kumar, S. L. Happy, A. Routray, "A real-time robust facial expression recognition system using HOG features," in *Proceedings of the 2016 International Conference on Computing Analytics and Security Trends*, Piscataway: IEEE, 2016, pp. 289-293.

[7] S. S. Liu, Y. T. Tian, C. Wan, "Facial expression recognition method based on GABOR multi-orientation features fusion and block histogram," *Acta Automatica Sinica*, vol. 37, no. 12, pp. 1455-1463, 2011.

[8] V. D. A. Kumar, S. Malathi, et al., "Facial recognition system for suspect identification using a surveillance camera," *Pattern Recognition and Image Analysis*, vol. 28, no. 3, pp. 410-420, 2018.

[9] Y. T. Tseng, S. Kawashima, S. Kobayashi, et al., "Forecasting the seasonal pollen index by using a hidden Markov model combining meteorological and biological factors," *Science of the Total Environment*, vol. 698, Article No. 134246, 2020.

[10] K. Sun, H. Kang, H. H. Park, "Tagging and classifying facial images in cloud environments based on KNN using MapReduce," *Optik – International Journal for Light and Electron Optics*, vol. 126, no. 21, pp. 3227-3233, 2015.

[11] Y. Tang, "Deep learning using linear support vector machines," 2013. [Online]. Available: http://deeplearning.net/wp-content/uploads/2013/03/dlsvm.pdf. [Accessed: Apr. 10, 2019].

[12] M. Sandler, A. Howard, M. Zhu, et al., "Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection, and Segmentation," 2018. [Online]. Available: https://arxiv.org/pdf/1801.04381v1.pdf. [Accessed: Jun. 22, 2019].

[13] S. Li, W. Deng, J. P. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Piscataway, IEEE, 2017, pp. 2584-2593.

[14] C. Pramerdorfer, M. Kampel, "Facial expression recognition using convolutional neural networks: state of the art," 2016. [Online]. Available: https://arxiv.org/pdf/1612.02903.pdf. [Accessed: Apr. 10, 2019].

[15] L. L. Xu, S. M. Zhang, J. L. Zhao, "Expression recognition algorithm for parallel convolutional neural networks," *Journal of Image and Graphics*, vol. 24, no. 2, pp. 227-236, 2019.

[16] C. Szegedy, S. Ioffe, V. Vanhoucke, et al., "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, Palo Alto, CA: AAAI, 2017, pp. 4278-4284.

[17] A. Kandeel, M. Rahmanian, F. Zulkernine, H. M. Abbas, and H. Hassanein, "Facial Expression Recognition Using a Simplified Convolutional Neural Network Model," in *Proc. 2020 Int. Conf. Commun., Signal Process., and their Appl. (ICCSPA)*, Sharjah, United Arab Emirates, 2021, pp. 1-6, doi: 10.1109/ICCSPA49915.2021.9385739.