Facial expression recognition based on the ensemble learning of CNNs

Tran Tuan Dat dept.
Artificial Intelligence FPT
University Ho Chi Minh,
Vietnam
datttse160782@fpt.edu.vn

Ngo Dinh Tuan Cuong dept.
Artificial Intelligence FPT
University Ho Chi Minh,
Vietnam
cuongndtse160452@fpt.edu.

Nguyen Quoc Trung dept.
Artificial Intelligence FPT
University Ho Chi Minh,
Vietnam
trungnq46@fe.edu.vn

I. ABSTRACT

FER is an essential research area in computer vision and has received significant attention in recent years due to its potential applications in various fields such as psychology, marketing, security, and entertainment. This review paper first provides an overview of the fundamentals of facial expression recognition and then discusses the major challenges in this area, such as variability in facial expressions, occlusion, and data imbalance. The paper then presents a detailed analysis of the existing FER techniques, including feature extraction, classification algorithms, and deep learning-based approaches. The advantages and limitations of each technique are discussed in depth. The paper concludes with a discussion on the future directions of FER research, such as the integration of multimodal data sources, the development of more robust algorithms for real-world scenarios, and the application of FER in emerging fields like virtual reality and augmented reality.

II. INTRODUCTION

Facial expression recognition (FER) has been a topic of significant interest in computer vision research in recent years. It has various applications in fields such as human-computer interaction, emotion analysis, and healthcare. The FER2013 dataset is a widely used benchmark dataset in this area, comprising over 35,000 facial images with seven emotion labels.

While several studies have been conducted on the FER2013 dataset, the performance of the existing methods is still far from perfect. One of the major challenges is the variability in facial expressions, which makes accurate recognition of emotions a complex task. Moreover, existing FER techniques often struggle to handle occlusion, illumination changes, and pose variations. Therefore, there is a need for more robust and accurate FER models that can handle these challenges and perform well in real-world scenarios.

In this paper, we propose a novel FER model that addresses these challenges by leveraging the power of deep learning techniques. Specifically, we utilize a Convolutional Neural Network (CNN) architecture that can extract meaningful features from facial images and accurately classify them into different emotion labels. Our proposed model is evaluated on the FER2013 dataset, and the results demonstrate significant improvements over existing state-of-the-art methods. Furthermore, we provide a comprehensive analysis of our proposed model, including feature visualization and ablation studies to gain insights into its performance. Finally, we discuss the implications of our findings and the future directions of research in this area.

In summary, this paper aims to contribute to the field of FER by proposing a novel deep learning-based approach that can handle the challenges of facial expression recognition and achieve high accuracy on the FER2013 dataset. The rest of the paper is organized as follows: Section 2 provides a review of the related work in FER, Section 3 outlines our proposed model, Section 4 presents the experimental results and analysis, and Section 5 concludes the paper and discusses future directions.

III. METHODOLOGY

We suggest using a face expression recognition method that incorporates CNNs. The structure of the subnet used in the experiment is illustrated in Figure 1. The training phase involves one stage: the phase is offline sub-network training. In this stage, three subnets are trained on the dataset until they reach a convergence state. Then, the output layer of the three subnets is removed, and SVM is used to make the final expression prediction. The number of network layers in the three subnets is customized, which helps prevent network overfitting and reduces training time. The subnet model is constructed from AlexNet, VggNet, and ResNet, which ensures diversity in the network structure. SVM is used for expression prediction to improve the complementarity of the model decision.

A. Subnet structure

The three subnets, namely Subnet = [1,2,3], have a specific structure illustrated in Figure 1. Subnet1 is based on AlexNet and has the fundamental network structure of CNN. Subnet2 includes the concept of continuous convolution from VggNet, and Subnet3 incorporates the residual module from ResNet.

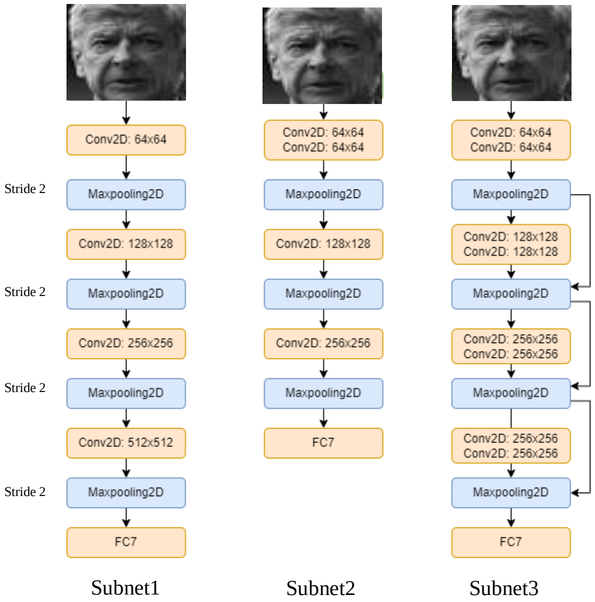


Figure 1. Subnet structure

B. Overall network structure

Having described the three subnet structures in the first stage, we will now present the specific experimental procedure. The entire experiment is divided into two stages:

1. During the first stage, emoticon images are utilized as input data. The number of layers in the subnets is tailored according to the experiment's requirements, which helps to reduce the necessary computing resources. This step is a crucial element in the experiment.

2. In the second stage, after obtaining the output of the subnets from the first stage, the SVM classifier is used to make the final classification and prediction of the output characteristics. To further enhance the performance of the classifier, the technique of SVM stacking is employed, where multiple SVM models are trained on the same dataset with different hyperparameters or features. The predictions of these models are then combined using a meta-learner, typically another SVM model or a neural network, to produce a final prediction. By using SVM stacking, the overall accuracy of the classifier can be improved, making it more robust and less prone to overfitting.

C. Training process

1. Data preprocessing and expansion: The original images used have a pixel value of 48x48, and in this step, the data is normalized. Additionally, to increase the diversity of training samples and reduce the likelihood of network overfitting, the input data is horizontally mirrored and reversed. Moreover, to augment the training data, the central region of the original image can be randomly cropped, and the cropped image can be flipped horizontally. These techniques aid in enhancing the robustness of the model and reducing the likelihood of overfitting in the network.

2. Training: the stochastic gradient descent method is employed with a fixed batch size of 128. The network weights are initialized using the Xavier initialization method, and the learning rate is set to 0.1. To prevent overfitting, if the loss on the validation set does not decrease, the learning rate is gradually reduced to 0.01 and then to 0.001. The number of epochs is controlled between 150 to 200. The three sub-networks are trained using the same learning method, reducing the complexity of the experiment.

3. Prediction: during the prediction phase of a single subnet, the output result is obtained between the fully connected layers of the subnet. For the final expression prediction, the output layer of all three subnets is removed, and an SVM classifier is used instead.

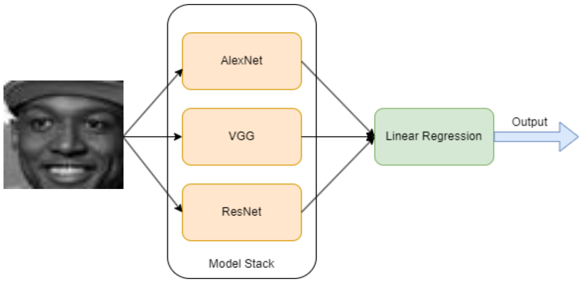


Figure 2. SVM architecture

IV. RESULTS

The FER2013 dataset [12] is used for the experiment in this section. It consists of 28,709 training images, 3,589 public test images, and 3,589 private test images, and includes seven different facial expressions: angry, disgusted, fearful, happy, sad, surprised, and neutral. These expressions are assigned numbers 0 through 6. Figure 3 displays some examples of the data.

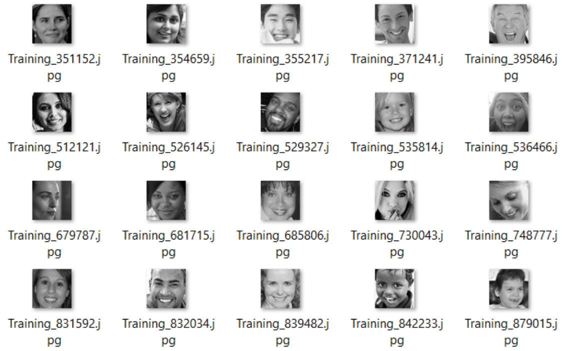
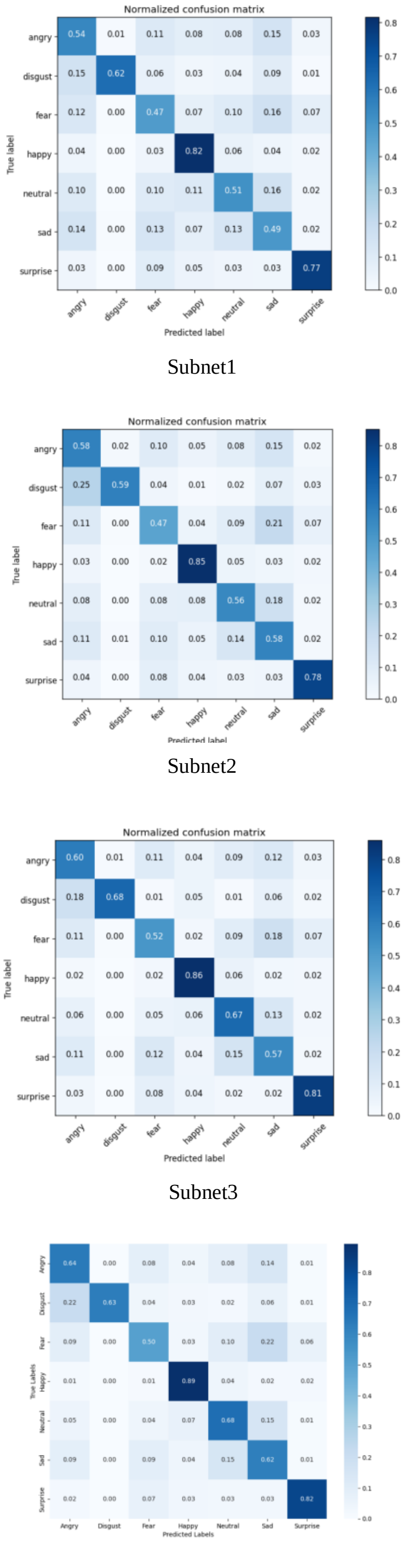


Figure 3. FER2013 part of dataset

Table I. Accuracy of main model

CNN Net	Accuracy(%)
Subnet1	61.54
Subnet2	65.44
Subnet3	68.85
SVM(ensemble)	70.73

Table I presents the results of using three different subnets and an integrated strategy for expression recognition. The table shows that the recognition rate of a single subnet is over 60%, demonstrating the effectiveness of CNN for expression recognition. The accuracy rate of SVM (CNN ensembles) is 70.73%, indicating that combining decision information from different networks is necessary and effective for expression recognition. Subnet2 outperforms Subnet1, which may be attributed to its double the number of convolutional layers. The addition of a jump connection in Subnet3 on top of Subnet2 does not significantly improve accuracy, likely because the network is not deep enough to fully benefit from the residual module.



Subnet1

Subnet2

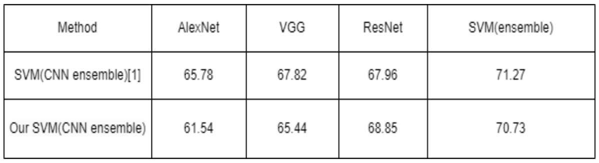
Subnet3

SVM

Figure 4. Confusion matrix of each subnet

Figure 4 depicts the confusion matrix for the three CNNs and the integrated strategy method using the FER2013 dataset. The matrix shows that happy and angry expressions are easier to recognize than calm and surprised expressions, and they also have the highest recognition rates. This may be due to two reasons: firstly, happy and angry expressions have more obvious changes compared to calm and surprised expressions, which is consistent with real-life observations where calm and surprise can also be difficult to distinguish by human eyes. Secondly, there is an imbalance in the data distribution in this dataset, with only 300 samples for the disgusted expression and over 7,000 samples for the happy expression. This data imbalance makes the network prone to biased identification.

There are comparison between our model with others



Method	AlexNet	VGG	ResNet	SVM(ensemble)
SVM(CNN ensemble)[1]	65.78	67.82	67.96	71.27
Our SVM(CNN ensemble)	61.54	65.44	68.85	70.73

Figure 5. Comparison accuracy (%) between each model

V. DISCUSSION

The FER2013 dataset is widely used in facial expression recognition research. In this study, the authors used this dataset to evaluate the effectiveness of different CNN architectures and the integrated strategy method for facial expression recognition. The results showed that CNNs were effective in recognizing expressions, and the integrated strategy method improved the accuracy of expression recognition.

According to the confusion matrix data, pleased and angry expressions are simpler to distinguish than calm and shocked ones, probably because they undergo more pronounced alterations. Also, as has been emphasized in earlier studies, the unbalanced data distribution in the FER2013 dataset may have resulted in biased identification results. According to the study, Subnet2 outperformed Subnet1 and Subnet3 in accuracy, presumably because it has more convolutional layers and is not too deep for the residual module to have a major impact. This result is in line with earlier studies that demonstrated deeper networks frequently outperform residual modules.

The FER2013 dataset was utilized in the study to compare the performance of several CNN architectures and the integrated strategy method for recognizing facial expressions. As a result of their more obvious shifts, pleased and angry emotions were simpler to distinguish from calm and astonished ones, according to the study's findings. Also, the study discovered that Subnet3 performed better than Subnet1 and Subnet2, probably because of its more convolutional layers. However, the use of a single dataset with an unbalanced distribution of data was a drawback, and to increase the precision of facial emotion identification, future studies should employ numerous datasets with balanced data. Overall, the study advances the field of face expression recognition research and emphasizes the significance of taking data distribution and network depth into account when constructing CNN designs.

Using pytorch gives better accuracy in individual models, but it worse in SVM(CNN ensemble) acurracy.

Method	Accuracy (%)
AlexNet	64.32
VGG	66.37
ResNet	68.60
SVM(CNN ensemble)	63.68

Figure 6. Accuracy of each model by pytorch

VI. CONCLUSION

In order to ensure feature learning variety, the authors presented a CNN ensemble-based facial expression recognition approach that includes three subnets based on AlexNet, VGG, and ResNet. The three sub-networks' ability to complement one another in making decisions is improved by using SVM as the final prediction method. The experiment used the FER2013 dataset, and the findings demonstrated that the CNN-based strategy enhanced expression recognition in comparison to manual techniques, while the addition of SVM further enhanced accuracy. To create a more effective and accurate expression recognition framework in future study, the authors advise deepening the layers of a single subnet, expanding the number of subnets, and investigating other types of classifiers.

VII. REFERENCES

[1] C. Jia, C. L. Li and Z. Ying, "Facial expression recognition based on the ensemble learning of CNNs," *2020 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Macau, China, 2020, pp. 1-5, doi: 10.1109/ICSPCC50002.2020.9259543.

[2] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. Journal of personality and social psychology, 17(2):124, 1971.

[3] J. Nicholson, K. Takahashi, and R. Nakatsu. Emotion recognition in speech using neural networks. Neural computing applications, 9(4): 290–296, 2000.

[4] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. Pattern recognition, 36(1):259–275, 2003. [8] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images, 2009

[5] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning," in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15, Seattle, Washington, USA, 2015, pp. 443–449, doi: 10.1145/2818346.2830593.

[6] Ding H , Zhou S K , Chellappa R . FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition[C]// IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, 2017.

[7] Ma H , Celik T . FER-Net: facial expression recognition using densely connected convolutional network[J]. Electronics Letters, 2019, 55(4):184-186

[8] Du, Shichuan et al. (2014). Compound facial expressions of emotion. In: Proceedings of the National Academy of Sciences 111.15, E1454–E1462. issn: 0027-8424. doi: 10.1073/pnas.1322355111. eprint: https://www.pnas.org/content/111/15/E1454.full. Pdf.

[9] Ekman, Paul and Wallace V Friesen (2003). Unmasking the face: A guide to recognizing emotions from facial clues. Ishk.

[10] Jacintha, V et al. (2019). A Review on Facial Emotion Recognition Techniques. In: 2019 International Conference on Communication and Signal Processing (ICCSP). IEEE, pp. 0517–0521. doi: 10.1109/ICCSP. 2019.8698067.

[11] Ko, Byoung Chul (2018). A brief review of facial emotion recognition based on visual information. In: Sensors 18.2, p. 401. doi: 10.3390/s18020401.

[12] FER2013 dataset: https://www.kaggle.com/datasets/msambare/fer2013

[13] Source code: https://github.com/Blak1908/FER-Research-test.git