ICIC Express Letters
Part B: Applications
Volume **15**, Number **4**, April **2024**

ICIC International ©2024 ISSN 2185-2766
pp. **327–334**

PARTIAL FACE EXPRESSION RECOGNITION USING DEEP LEARNING APPROACH

Chayanon Sub-r-pa[1], Chung-Yen Liao[2] and Rung-Ching Chen[1,*]

[1]Department of Information Management
Chaoyang University of Technology
No. 168, Jifeng East Road, Wufeng District, Taichung 413310, Taiwan
t5220317@gm.cyut.edu.tw; *Corresponding author: crching@cyut.edu.tw

[2]Department of Electronic Engineering
Feng Chia University
No. 100, Wenhua Road, Xitun District, Taichung City 407102, Taiwan
cyliao.t11@o365.fcu.edu.tw

Received July 2023; accepted September 2023

Abstract. *Facial Expression Recognition (FER) is essential information for many services and Artificial Intelligence (AI) applications. With a Convolutional Neural Network (CNN) and a large number of samples, expression recognition can be efficient. Existing facial expression datasets offer the full facial image with labeled expressions. However, some situations do not allow capturing the complete facial image, which can affect the performance of existing FER. This paper proposed Partial Face Expression Recognition (PFER) to recognize expressions from incomplete facial images. We explore the possibility of using incomplete facial images with expression recognition. In the experiment, we report the performance of PFER with different backbones, including CNN and a state-of-the-art Vision Transformer (ViT). Our experimental result shows that PFER has a high potential to recognize incomplete facial images.*
Keywords: Facial expression recognition, Partial facial expression recognition, Convolutional neural network, Vision Transformer

1. **Introduction.** Facial image information is crucial for facial or expression recognition [2-5]. FER and its application have gained attention from the prospective industry. FER is a complex model that requires knowledge from psychology, computer science, and AI. With a CNN, FER is shown efficient performance that can deploy to industrial applications, such as pain detection [1], which can report the real-time status of a patient to the hospital.

Existing FER [3-5] reports high accuracy while recognizing expressions with full facial images. In real-world applications, some situations cannot capture a full-facial image, such as the subject wearing a facial mask during the COVID-19 pandemic or wearing sunglass. An incomplete facial image affects the performance of the existing facial [6] and expression recognition model. However, facial expression recognition is a task that can be done by specialists in the physiological area, and some expressions can be described with incomplete facial images.

From physiological and psychological research, the definition of expression can be explained. Facial Action Coding System (FACS) [7] is a fine-grained and anatomically based coding system that differentiates between 44 facial movements known as Action Units (AU). Figure 1 shows the example of a list of upper and lower-face AUs and their interpretation. Coders are trained to apply specific operational criteria to determining the onset and offset as well as the intensity of the AUs. FACS showed that facial expressions

DOI: 10.24507/icicelb.15.04.327

FIGURE 1. A list of upper and lower-face AUs and their interpretation [5]

of pain are composed of a small subset of facial activities, namely lowering the brows (AU4), cheek raise/lid tightening (AU6-7), nose wrinkling/raising the upper lip (AU9-10), opening the mouth (AU25-27), and eye closure longer than 0.5s (AU43).

According to FACS, the differences between facial expressions usually locate in certain crucial regions, such as the eye and mouth. Moreover, studies have shown that attention naturally focuses on specific facial regions when humans recognize and distinguish different facial expressions [8,9]. For example, the eyes play an important role in fear analysis, while the mouth is vital for recognizing happiness.

Even a complete facial image is necessary for the existing FER model. However, an incomplete face image also has enough information to classify the expression. This research studies the partial face for expression recognition, considering the partial face as the upper and lower. We study the potential to implement the expression recognition model for incomplete facial images. Our experiment uses two different datasets: the controlled and in-the-wild facial expression datasets. The experiment uses different deep learning architectures, including the state-of-the-art ViT [10].

The experimental results confirm that using PFER can produce high accuracy closest to using full-face image in a controlled environment dataset. And in the in-the-wild dataset, PFER with the upper-face image can get an accuracy of 78.22% but is still significantly lower than FER, which has an accuracy of 87.28%.

The rest of this paper is organized as follows. In Section 2, we provide an overview of the related literature. Section 3 describes the PFER's detail. Section 4 presents the implementation, the data collection, data pre-processing, and experiment and analyzes the results obtained. Finally, in Section 5, we summarize the contributions of our study and suggest potential directions for future research.

2. **Related Work.** Deep learning and CNN have become state-of-the-art for FER. [3-5] used the CNN-based approach to classify facial expressions. [11] used a Transformer-based model to improve the accuracy. However, most of the existing methods are trained with full facial images. Classifying with incomplete facial images affects the performance of the existing methods.

[12] proposed recognizing expressions in partial facial images with a rapid feature vector technique. They developed a feature extraction technique from the upper area of the facial image, and then classified expression using a model based on CNN and LSTM. They found a high potential to use only the upper facial area in expression recognition.

[13] used the CNN model to classify facial expressions with the full face, half face (left and right side), eyes, single eye, mouth, and half of the mouth. They experiment and report the high potential to use haft face for expression recognition. However, the experiment is with small-scale datasets, which need more investigation.

This paper considers using full, upper, and lower facial images for FER. Our experiment uses large-scale datasets, including RafD [14] controlled environment facial image dataset and Raf-DB [15,16] facial expression in the wild dataset.

3. **Proposed Work.** FER requires a full facial image for classification. However, many situations cannot capture all facial images. Using PFER increases flexibility in this task. PFER is a model that uses partial face images as input for expressions recognition, which can be helpful in many real-world applications.

3.1. **Partial Face Expression Recognition (PFER).** PFER requires the partial face image to recognize the expressions. The partial face image can be categorized into two categories. First is the upper facial image, which considers the upper area from the tip of the nose. Second is the lower facial image, which considers the facial image in the lower area from the tip of the nose.

3.2. **Classification model.** Several research uses CNN architecture in FER, such as VGG16 [17] or EfficientNet [18]. This experiment includes EfficientNet architecture as the CNN-based model. From [4,19] report the effective performance to use EfficientNet in FER. Moreover, our experiment also includes ViT [10], a state-of-the-art Transformer-based model.

The models used in our experiment modify only the output layer to match the expressions available in each dataset. FER models are trained with full facial images. And PFER is trained with upper facial images or lower facial images. Then the experimental results are used to confirm the results when using partial face images as input to the models.

4. **Experiment.** We designed the experiment to evaluate the potential of PFER based on the deep learning approach. This section includes implementation detail of datasets and classification models. Then evaluate and analyze the experimental results.

4.1. **Implementation.**

4.1.1. *Dataset.* This paper uses two facial image datasets to train PFER models. The first dataset is Radboud Faces Database (RafD) [14]. The RafD is a high-quality faces images database that contains pictures of eight emotional expressions (anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral). Each emotion is shown with three different gaze directions, and all pictures were taken from five camera angles simultaneously (0-180 degrees). RafD includes side facial images, which is uncommon in existing facial expression datasets. We separate the experiments with this dataset by including side facial images (0-180 degrees) and excluding side facial images (45-135 degrees). The example from 0-180 degrees is shown in Figure 2. To verify the experimental results, we randomly split this dataset into a training set (80%), validation set (10%), and testing set (10%).

FIGURE 2. Sample from RafD [14] with a different angle

The second dataset, Real-world Affective Faces Database (Raf-DB) [15,16] is a large-scale facial expression database with 29,672 facial images downloaded from the Internet. Based on the crowdsourcing annotation, each image has been independently labeled into seven expressions (anger, disgust, fear, happiness, sadness, surprise, and neutral). This dataset is already split into the training set and the testing set. For Raf-DB, we use the testing set as the validation set.

Figure 3 shows samples from RafD and Raf-DB used in our experiment. Each image is processed and cropped for upper and lower facial images. The cropped upper and lower facial image is done by finding the facial landmarks with YuNet [20]. We use a nose landmark in the y-axis of each image as a separate point. From the top of the image to the nose landmark is considered the upper face, and from landmarks to the bottom of the image, it is considered the lower face.

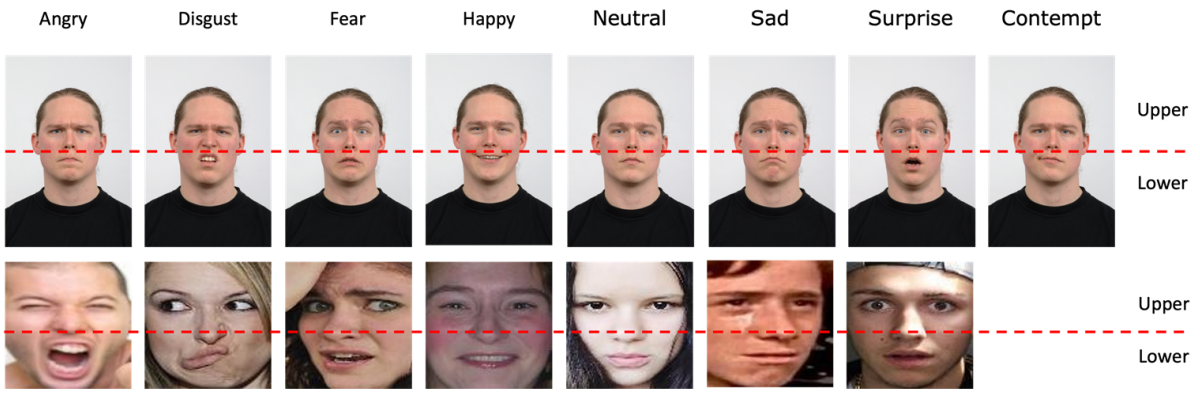

FIGURE 3. Sample images from RafD [12] (1st row) and Raf-DB [13,14] (2nd row)

4.1.2. *Classification models.* We build models for PFER using the CNN-based and Transformer-based models. In CNN-based, we used EfficientNet. EfficientNet is a method for uniformly scaling all depth/width/resolution dimensions using a simple yet highly effective compound coefficient. EfficientNet uses version numbers to denote the scaling of the network respectively. In this experiment, we use EfficientNet version B0-B2 since it is commonly used with facial images [4,19].

ViT [10] is state-of-the-art for image classification based on the Transformer model. We compare and analyze the results of EfficientNet and ViT to find the suitable mode in PFER.

In our experiment, all models used pre-trained weights with classification from ImageNet. The classifier is designed for eight classes output in RafD and seven in Raf-DB.

4.1.3. *Training parameters.* All models are trained with the same set of parameters. We trained each model and dataset for 50 epochs. We did a grid search to find the suitable learning rate for our task, including learning rate = 0.001, 0.0005, and 0.0001. We use state-of-the-art SAM optimizer [21] to optimize our training instead of Adam or SGD. SAM can improve model generalization compared to another optimizer. We use the cross-entropy loss as a loss function, which is standard in the classification model. And we use a batch size of 32.

4.2. **Evaluation and analysis.** During the training, we evaluate the model with the validation set for the RafD dataset and the testing set for the Raf-DB dataset. The model with the highest accuracy is considered the best model for each architecture and dataset pair. We evaluate and analyze the best model in each dataset with a testing set focusing on accuracy, loss, and number of parameters.

4.2.1. *Learning rate.* We experiment with each model and dataset with different learning rates and aim to find the suitable learning rate for PFER. All models are used cross-entropy as a loss function. Figure 4 shows loss during the training for the Raf-DB dataset with models EfficientNet B0 and ViT. In EfficientNet, B0 shows insignificant results in differences in learning rate. The results from EfficientNet B1 and B2 follow the same trend. However, ViT shows significantly different when using a learning rate equal to

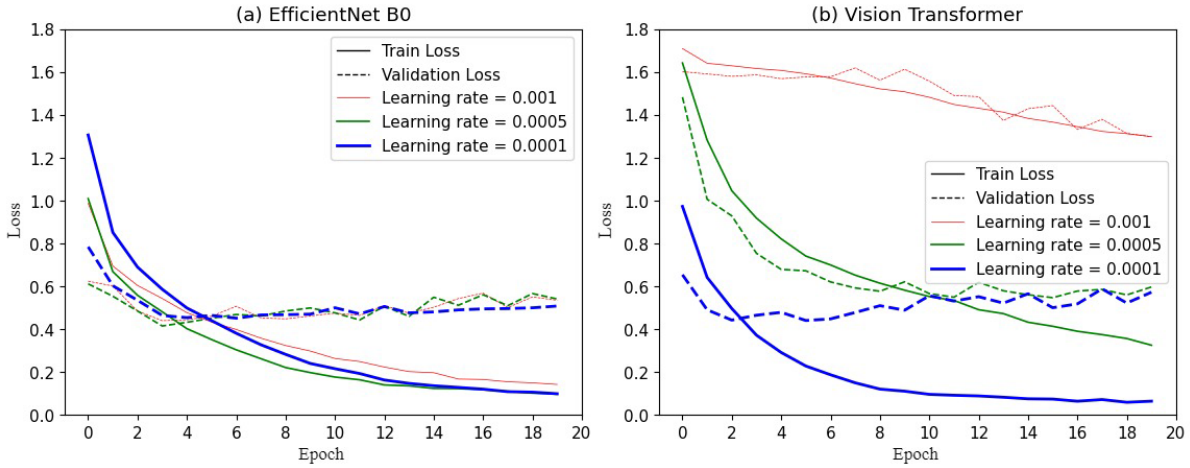

FIGURE 4. Training and validation loss: (a) EfficientNet B0 [16]; (b) ViT [8]

0.0001. ViT cannot converge with a high learning rate for FER and PFER. The results for analysis in the following section are from models trained with a learning rate equal to 0.0001.

4.2.2. *Models accuracy.* For model accuracy, we evaluate model accuracy with testing sets from RafD and Raf-DB datasets with full facial images, upper facial images, and lower facial images. Table 1 shows each model's and dataset's accuracy. The best accuracy for the two datasets comes from full facial images. However, partial face images and RafD (images with a controlled environment) can get similar accuracy compared to full-facial images. We can conclude with PFER lower facial image it is more necessary for expression recognition.

TABLE 1. Accuracy (the best results are shown in **bold**)

Dataset	Area	Model			
		EfficientNet B0	EfficientNet B1	EfficientNet B2	ViT [8]
RafD(0-180) [12]	Full	0.9934	0.9947	**0.9960**	0.9842
	Upper	0.9406	**0.9591**	0.9565	0.9024
	Lower	**0.9934**	0.9894	0.9881	0.9815
RafD(45-135) [12]	Full	**0.9979**	**0.9979**	0.9958	0.9937
	Upper	0.9540	0.9623	**0.9665**	0.9498
	Lower	**0.9979**	0.9937	0.9958	0.9916
Raf-DB [13,14]	Full	0.8618	0.8651	**0.8728**	0.8698
	Upper	0.7682	0.7792	**0.7822**	0.7729
	Lower	0.7579	0.7596	0.7666	**0.7702**
Number of parameters (Million)		4	6.5	7.7	85.6

However, with facial expressions in the wild or the Raf-DB dataset, full facial images are significantly more accurate than partial facial images. Unlike a controlled environment, PFER with upper facial has slightly more accuracy than lower facial images.

We compare the result when using CNN-based and Transformer-based. The results show that ViT can classify facial expressions with high accuracy. However, compared to CNN-based, the results are slightly lower (1%) with full and partial face images.

From the results, each model has a high potential to implement in real-world applications. We analyze the parameter number of each model. The lower parameter number can

gain the advantage of less computation time, and it is possible to show better performance in real-time applications (e.g., real-time FER from a webcam).

The end of Table 1 shows the number of parameters of each model. The fewer parameters require less computation power and memory, allowing the model to run in smaller

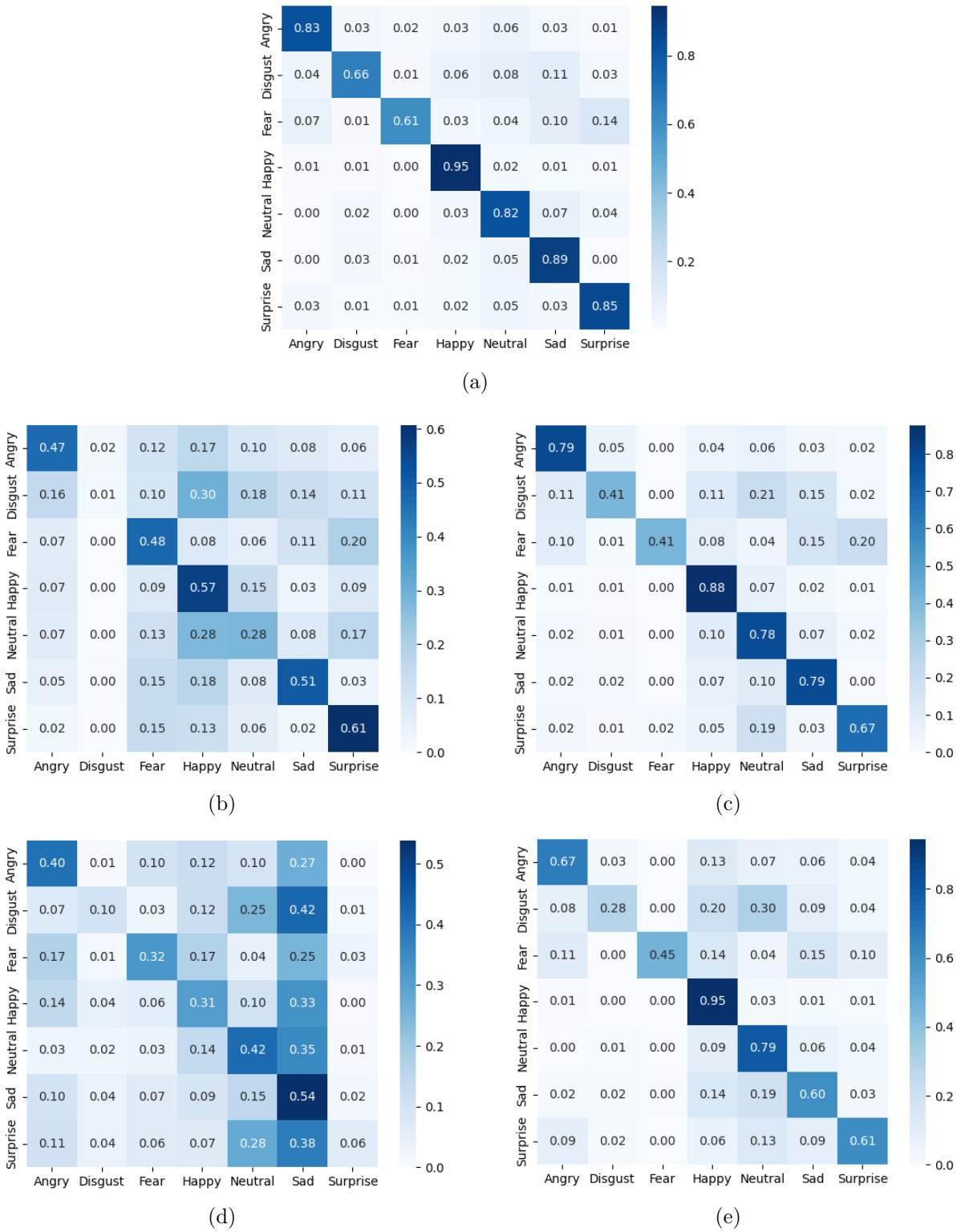

FIGURE 5. Confusion matrix of EfficientNet B2 on Raf-DB: (a) Trained and tested with full-face images; (b) trained with full face, tested with the upper; (c) trained and tested with the upper; (d) trained with full face, tested with the lower; (e) trained and tested with lower

environments like IoT. We conclude that ViT is unsuitable for real-time applications or low computation power environments. And EfficientNet B2 shows the best accuracy, but EfficientNet B0 can classify faster with slightly lower accuracy.

4.2.3. *Confusion matrix.* Figure 5 shows the confusion matrix with the cross-test dataset. The results using a model trained with a full facial (Figure 5(b) and Figure 5(d)), and then testing with partial face image shows a significant error. This is because the testing dataset (partial face image) differs significantly from the training dataset (full face image).

The results from the model trained and tested with the same partial face image dataset show better performance than those trained with a full face. However, the accuracy from the model trained with the partial image is still significantly lower than that trained and tested in full face. Partial face images can accurately predict happy, neutral, and sad expressions, but they have significant errors in classifying disgust and fear expressions.

5. **Conclusions.** This paper explores the potential of using PFER. We experimented with our method with facial expressions in the wild and controlled environment datasets. The PFER can get high accuracy compared to FER in a controlled environment. Expression recognition with the lower facial image is 0.3% different compared to recognition with full facial images, and it has the same accuracy when excluding side facial images. This implies that getting full facial images for expression recognition is unnecessary, especially when we can access lower facial images. However, with facial images in the wild, FER is still necessary since facial in the wild has more complex information.

Our experiment uses CNN-based (EfficientNet B0-B2) and Transformer-based (ViT). We found that in FER and PFER, using CNN-based is slightly better than Transformer-based. And ViT needs a low learning rate to train the model for this problem.

PFER shows the potential to use incomplete facial images. In future work, we plan to explore the different partial faces, such as upper facial images with incomplete lower facial images. The necessary ratio of facial images for expression recognition can be used to improve data augmentation and model accuracy.

Acknowledgment. This paper is supported by the Ministry of Science and Technology, Taiwan. The Nos are MOST-111-2622-E-324-002- and MOST-111-2221-E-324-020, Taiwan.

REFERENCES

[1] T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J.-U. Garbas and U. Schmid, Automatic detection of pain from facial expressions: A survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.43, no.6, pp.1815-1831, 2021.

[2] W. Cheng, H. Hsiao and D. Lee, Face recognition system with feature normalization, *International Journal of Applied Science and Engineering*, vol.18, no.1, 2021.

[3] Y. Guo, Y. Xia, J. Wang, H. Yu and R.-C. Chen, Real-time facial affective computing on mobile devices, *Sensors*, vol.20, no.3, 2020.

[4] A. V. Savchenko, L. V. Savchenko and I. Makarov, Classifying emotions and engagement in online learning based on a single facial expression recognition neural network, *IEEE Transactions on Affective Computing*, vol.13, no.4, pp.2132-2143, 2022.

[5] X. Wang, X. Hao and K. Wang, Facial expression recognition based on multi-branch adaptive squeeze and excitation residual network, *International Journal of Innovative Computing, Information and Control*, vol.17, no.3, pp.735-751, 2021.

[6] W. Cheng, H. Hsiao, Y. Hong and D. Wang, Masked face recognition based on FaceNet and genetic algorithm, *International Journal of Applied Science and Engineering*, vol.20, no.3, 2023.

[7] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, Palo Alto, CA, USA, 1978.

[8] B. Martinez, M. F. Valstar, B. Jiang and M. Pantic, Automatic analysis of facial actions: A survey, *IEEE Transactions on Affective Computing*, vol.10, no.3, pp.325-347, 2019.

[9] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider and J. Kissler, Mapping the emotional face. How individual face parts contribute to successful emotion recognition, *PLoS ONE*, vol.12, no.5, e0177239, 2017.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, An image is worth 16×16 words: Transformers for image recognition at scale, *The 9th International Conference on Learning Representations (ICLR 2021)*, Virtual Event, Austria, 2021.

[11] L. Lo, H. Xie, H.-H. Shuai and W.-H. Cheng, Facial chirality: From visual self-reflection to robust facial feature learning, *IEEE Transactions on Multimedia*, vol.24, pp.4275-4284, 2022.

[12] R. Khoeun, P. Chophuk and K. Chinnasarn, Emotion recognition for partial faces using a feature vector technique, *Sensors*, vol.22, no.12, 2022.

[13] R. Melaugh, N. Siddique, S. Coleman and P. Yogarajah, Facial expression recognition on partial facial sections, *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp.193-197, 2019.

[14] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk and A. van Knippenberg, Presentation and validation of the Radboud Faces Database, *Cognition & Emotion*, vol.24, no.8, pp.1377-1388, 2010.

[15] S. Li, W. Deng and J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2584-2593, 2017.

[16] S. Li and W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, *IEEE Transactions on Image Processing*, vol.28, no.1, pp.356-370, 2019.

[17] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv Preprint*, arXiv: 1409.1556, 2014.

[18] M. Tan and Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, *Proc. of the 36th International Conference on Machine Learning*, vol.97, pp.6105-6114, 2019.

[19] A. V. Savchenko, Facial expression and attributes recognition based on multi-task learning of lightweight neural networks, *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, Subotica, Serbia, pp.119-124, 2021.

[20] W. Wu, H. Peng and S. Yu, YuNet: A tiny millisecond-level face detector, *Machine Intelligence Research*, vol.20, pp.656-665, 2023.

[21] J. Kwon, J. Kim, H. Park and I. K. Choi, ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks, *The 38th International Conference on Machine Learning*, vol.139, pp.5905-5914, 2021.