DOI: 10.21917/ijsc.2022.0397

EFFICIENTNET FOR HUMAN FER USING TRANSFER LEARNING

Rajesh Singh[1], Himanshu Sharma[2], Naval Kishore Mehta[3], Anil Vohra[4] and Sanjay Singh[5]
[1,3,4,5]*Department of Electronic Science, Kurukshetra University, India*
[2]*CSIR-Central Electronics Engineering Research Institute, Pilani, India*

Abstract

Automatic facial expression recognition (FER) remained a challenging problem in computer vision. Recognition of human facial expression is difficult for machine learning techniques since there is a variation in emotional expression from person to person. With the advancement in deep learning and the easy availability of digital data, this process has become more accessible. We proposed an efficient facial expression recognition model based EfficientNet as backbone architecture and trained the proposed model using the transfer learning technique. In this work, we have trained the network on publicly available emotion datasets (RAF-DB, FER-2013, CK+). We also used two ways to compare our trained model: inner and cross-data comparisons. In an internal comparison, the model achieved an accuracy of 81.68 % on DFEW and 71.02 % on FER-2013. In a cross-data comparison, the model trained on RAF-DB and tested on CK+ achieved 78.59%, while the model trained on RAF-DB and tested on FER-2013 achieved 56.10% accuracy. Finally, we generated an t-SEN distribution of our model on both datasets to demonstrate the model's inter-class discriminatory power.

Keywords:

FER, Deep Convolution Neural Network, EfficientNet, Transfer Learning

1. INTRODUCTION

Emotions play an essential role in human communication, whether verbal or nonverbal. Body language, speech modulations, and facial expressions are important components in communicating emotions, and facial expression analysis has been a well-proven and well-researched field. Ekman et al. [1] studied human facial emotions in depth and discovered seven primary expressions: neutral, happy, sad, fear, anger, surprise, and disgust. Emotion may also be defined as a complicated experience involving several senses that reveal a person's current state of mind while observing proceedings around. Facial expression recognition has recently been a burning issue in psychology, psychiatry, and mental health research [2]. The facial expression may be used to rapidly and accurately determine emotions.

Additionally, utilizing facial expressions to identify emotions has a lot of advantages, including the fact that it is non-invasive and very affordable. Automatic identification of human emotions from facial expressions benefits smart living, health care systems, smart e-learning systems, and human-computer interface (HCI) systems [3],[5]. As a result, the research community has started concentrating on facial emotion recognition (FER) owing to its vast future prospects.

Inspired by the success of Deep Neural Networks (DNNs) in other fields of study, particularly convolutional neural networks (CNNs), they have an intrinsic feature extraction process from images which are attributed to the fundamental reason for their superior performance compared to other approaches [6],[7]. Although much work has been done using CNN to address FER, the underlying problem remains that the slight difference in facial expression attributes to distinct emotional states, which complicates the classification process, as well as the lack of a sufficiently high-resolution FER database with facial action coding system (FACS) [8],[9]. In real life, the target face may be captured from an unexpected angle. Therefore, a more practical FER system is required to recognize emotion from both frontal and side viewpoints [10],[11].

In this study, we have proposed the Modified EfficientNet (MoEffNet) FER pipeline, which includes a modified EfficientNet-B0 architecture that uses the transfer learning (TL) approach. TL uses the knowledge gained by the already trained model to solve the other related problems [12]. The proposed MoEffNet built over EfficientNet-B0 backbone architecture is evaluated on the well-known publicly available datasets, i.e. FER-2013 [13], RAF-DB [14], and CK+ [15] facial image datasets.

The FER-2013 and RAF-DB datasets are more challenging to evaluate due to various photos with diverse profile views. The proposed approach, which utilizes EfficientNet-B0 as its backbone architecture, has proven highly accurate on any dataset with any pre-trained model. Furthermore, the trained model meets the requirements for commercial applications in the real world.

The rest of this paper is structured as follows. Related work is discussed in Section 2. Section 3 describes the specifics of our proposed model for emotion recognition. The datasets used in the experiments are summarised in Section 4. Section 5 discusses the results, followed by the conclusion in Section 6.

2. RELATED WORK

FER remained a challenging problem in the last few decades. Several techniques have been investigated to solve the FER problem. In literature, the FER methods are prominently categorized into two main methods namely machine learning-based methods and Deep learning-based methods. The deep learning-based FER approaches are briefly explained below.

Even though machine learning-based approaches have been used extensively in the field of FER, their performance is still insufficient for real-world applications. The fundamental reason is that all classic FER approaches only address frontal views and perform poorly when applied to other perspectives. This constraint was solved, and performance was enhanced using deep learning-based FER approaches.

Zhao et al. [16] used a deep belief network for unsupervised feature learning and a neural network for emotion classification. Ding et al. [17] proposed a novel architecture called FaceNet2ExpNet. Another hybrid architecture with a transfer learning approach is proposed by [18], in which AlexNet is used with a support vector machine algorithm.

FER uses the Multi-Region Ensemble CNN (MRE-CNN) framework to collect global and local features from various

human face sub-regions [19]. Zhao et al. [20] presented a feature selection approach within AlexNet that can extract and filter facial features automatically. Wang et al. [21] introduced the Region Attention Network (RAN) to capture the relevance of face regions for occlusion and position variations in an adaptable manner. Self-Cure Network (SCN) reduces sample uncertainty in two ways: a self-attention mechanism that uses ranking regularisation to weight each sample in training and a rigorous relabelling technique that alters the labels of these samples in the lowest-ranked group [22]. Wen et al. [23] presented a probabilistic fusion strategy on an ensemble of CNN. Breuer et al. [24] employed CNN understanding techniques to study the link between the properties used by these computational networks, the FACS, and Action Units (AU). Cai et al. [25] new Probabilistic Attribute Tree-CNN directly addresses high intra-class variability.

Deep learning-based algorithms now take frontal photos into account to make the process easier. Most studies have even eliminated faces from the experiment using a range of profile view photos [10].

3. PROPOSED WORK

This section describes the basic working principle of convolution neural networks (CNNs) and the modified architecture of the EfficientNet-B0 backbone network. We will also describe the advantage of using EfficientNet-B0 for transfer learning.

3.1 CONVOLUTION NEURAL NETWORKS (CNN)

An artificial neural network (ANNs) is a set of algorithms that mimics the way the human brain performs complex tasks, i.e. recognizing human beings. These algorithms are used to find the relationships in a set of data. The traditional artificial neural network algorithms do not perform well for complex problems, i.e. image classification, video classification, pattern recognition, etc. Convolution neural network belongs to the family of artificial neural networks, which uses the mathematic operation Convolution [26] [27] to find the relationship between given data points. CNN outperforms the earlier state-of-the-art methods and achieves the highest accuracy in these applications.

CNN, in general, contains four layers, named convolution layer, pooling layer, dropout, and fully connected layer. These primary CNN layers extract the features from the input data. The CNNs can self-learn the effective features from the data, and these features of interest are represented by a small patch of convolution filters present in the convolution layers. These filters transform the data based on the filter values. The output of the convolution layer [28] is modelled by Eq.(1).

$$G[p,q] = \sum_{m}\sum_{n} h[m,n]i[p-m,q-n] \quad (1)$$

where i is the input image, h is the convolution filter, and (p,q) denotes the size of the output matrix.

In the next step, the information from the convolution layer output $G[p,q]$ is sent to subsequent pooling layers. Without any loss of data, the pooling layer reduces the size of the input data. For the classification task, the 2-dimensional data from the pooling layer passes through the flatten layer. The flatten layer converts the 2-D data to 1-D data. This 1-D data is then fed to the basic classifiers (sigmoid or softmax). Backpropagation algorithms were used to propagate errors back into the network, and these are also used to adjust the weights. The main objective of the backpropagation algorithm is to reduce the error (loss) function. The update of weight is done using Eq.(2).

$$W_i = W_i + \Delta W_i \quad (2)$$

Where randomly initialized initial network weights are denoted by W_i and ΔW_i denotes the error in the weights. The calculation of error is done using Eq.(3).

$$\Delta W_i = n\frac{dE}{dW_i}X_i \quad (3)$$

where the initial learning rate is denoted by n, the error function is represented by E, and the input image is denoted by X_i

3.1.1 *EfficientNet-B0 Architecture as DCNN for the Proposed Method:*

The traditional convolution neural networks (CNNs) (i.e., VGG16, and ResNet) architecture are developed at a fixed resource cost, and when more resource is made available these architectures are scaled up to achieve improved accuracy.

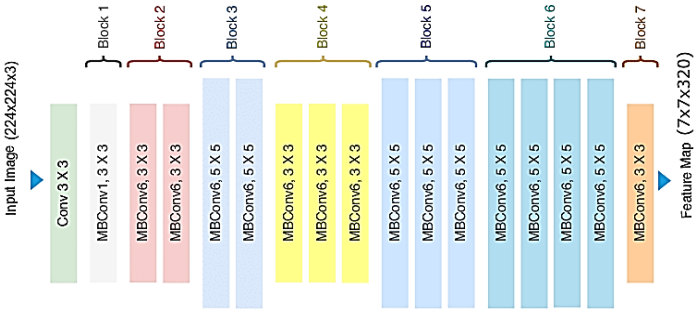


Fig.1. The architecture for baseline EfficientNet-B0 is used in our proposed FER pipeline

The EfficientNet uses the compound model scaling method to provide the balance in all dimensions, i.e., width, depth, and image resolution. Grid search is used to find the best relationship between different scaling dimensions and a given fixed resource constraint. Scaling coefficients for each dimension are computed, and these computed coefficients are then applied to scale up the baseline network to the target model size.

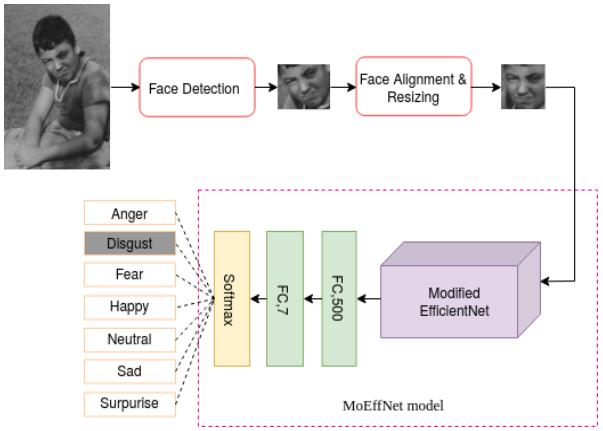


Fig.2. Block diagram representation of proposed FER pipeline. FC fully connected layer

The model scaling completely relies on the baseline network in [29] author uses the shown baseline network and named it EfficientNet-B0. The baseline architecture showed in Fig.1. In comparison with other existing CNNs on ImageNet, all versions of EfficientNet performed well in terms of accuracy and model size. The EfficientNet model archives both higher accuracy and better efficiency over existing CNNs, with reduced FLOPS and parameter size. By the Fig.the EffIcientNet-B7 archives a state-of-the-art 84.4% top-1 and 97.1% top-5 accuracy on ImageNet. The model is 8.4x smaller and runs 6.2x faster on the CPU interface than the widely used ResNet-50.

The complete FER pipeline is illustrated in Fig.2. We use the Dlib frontal face detector to recognize faces and perform face alignment on the data. The detected faces are resized to 48×48 using bicubic interpolation. For conducting this proposed experiment, we have used all the different versions of EfficientNet B0-B7. For the experiment, we made all the layers non-trainable, which means that the network weights do not improve during the training. We have trained all 8 known versions of EfficientNet on the ImageNet dataset, and for using the backbone architecture for 7 class emotion classifiers, we removed the input layer and classification layer from the backbone architecture. Finally, we selected EfficientNet-B0 in our proposed FER pipeline shown in Fig.1 because of its superior performance.

4. DATASET

In this work, the proposed model was trained and evaluated on three openly available datasets FER2013 [13], the extended Cohn–Kanade [15], and Real-world Affective Faces (RAF-DB) [14]. Before diving into the results, we briefly overview these databases.

4.1 FACIAL EXPRESSION RECOGNITION 2013 (FER2013)

The Facial Expression Recognition 2013 (FER-2013) database is a challenging but openly available database containing facial images of seven basic human emotions (happy, neutral, sad, disgust, fear, sadness, and surprise). This database was introduced at the ICML 2013 challenge in representation learning. The database contains 35,887 wild settings images of size 48x48. This dataset provides the train (28,709) and test (3589) images separately. This database also contains various variations, i.e., low-contrast images and face occlusion. Sample images from the FER dataset are shown in Fig.3.

4.2 EXTENDED COHN-KANADE

The extended Cohn–Kanade (also known as CK+) is a publicly available dataset. The CK+ facial expression database contains 593 videos of size 640x480. The video covers both posed and non-posed (spontaneous) sequences recorded from 123. The research uses the last frame of each video for experimental purposes. Sample images from this dataset are shown in Fig.3.

4.3 REAL-WORLD AFFECTIVE FACES (RAF-DB)

RAF-DB [14] is an openly available facial expression image dataset. The images in RAF-DB were collected randomly from the internet. The dataset contains a total of 29672 facial images of size 100×100. The RAF-DB provides the most diverse data sampling, which makes this dataset and thus provides all the real-world challenges in FER. This dataset was also annotated using the crowdsourcing method. The developer of the RAF-DB database provides separate training and testing data. Training data contains 12271 images, and testing data consists of 3068 images. Some sample images from it are shown in Fig.3.

Fig.3. Sample Images of datasets (RAF-DB, FER-2013, and CK+) utilized to test the proposed model

5. EXPERIMENTS AND RESULTS

5.1 IMPLEMENTATION DETAILS

We have trained a two-layer neural network on the features extracted by EfficientNet to classify 7 expressions, as shown in Fig.3. We optimized our networks using SGD with momentum set to 0.9 and an initial learning rate of 0.001, with a decay of 10% after every 10 epochs. The above experiments were carried out on the Keras framework with the TensorFlow backend. The training took place on an NVIDIA-DGX1 machine. The machine is equipped with a dual 20-core Intel Xeon E5-2698 v4 2.2 GHz processor and 512 GB of internal RAM, as well as an NVIDIA Tesla V100 GPU with 32 GB of RAM for GPU calculation.

5.2 PERFORMANCE EVALUATION

We test the classification ability of the modified EfficientNET-B0 network on the wild datasets, that is, RAF-DB and FER-2013. The result comparisons of RAF-DB and FER-2013 datasets are reported in Table.1 and Table.2, respectively. The confusion matrix of the proposed model for FER-2013 and RAF-DF is provided in Fig.4. The confusion matrix in Fig.4 shows the classification performance of the trained model. On both datasets, our model achieved a competitive and generalized performance.

We have also performed the cross-dataset comparison of our MoEffNet, where the fine-tuning of the EffIcientNet-B0 was trained on RAF-DB, and the model was tested on the CK+ dataset and FER-2013. The results of the cross-dataset are reported in Table.3 and Table.4, respectively.

Table.1. Trained model comparison on RAF-DB dataset

Method	Backbone	Accuracy
MRE-CNN [19]	VGG16	82.63%
FSN [20]	AlexNet	81.10%
RAN [21]	ResNet-18	86.90%
SCN [22]	ResNet-18	87.03%
Ad-Corre [30]	Xception	86.96%

MoEffNet (ours)	EffIcientNet-B0	81.68%

Table.2. Trained model comparison on FER-2013 dataset.

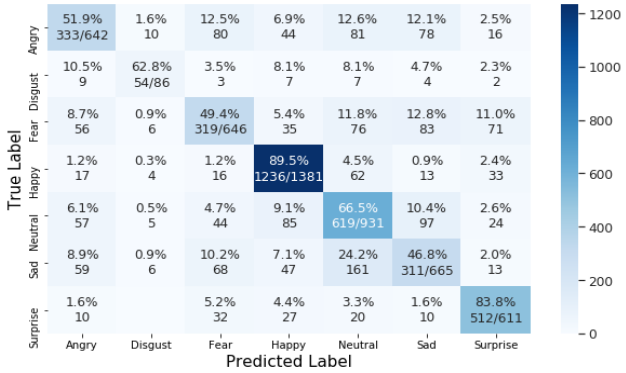
Method	Backbone	Accuracy
ECNN [23]	VGG16	69.96%
Ron et al. [24]	AlexNet	72.10%
Pat-ResNet [25]	ResNet-18	72.00%
Pat-VGGNet [25]	VGG-16	72.16%
LHC-Net [31]	ResNet34v2	74.42%
MoEffNet (ours)	EffIcientNet-B0	71.02%

Table.3. Trained model cross dataset comparison on CK+

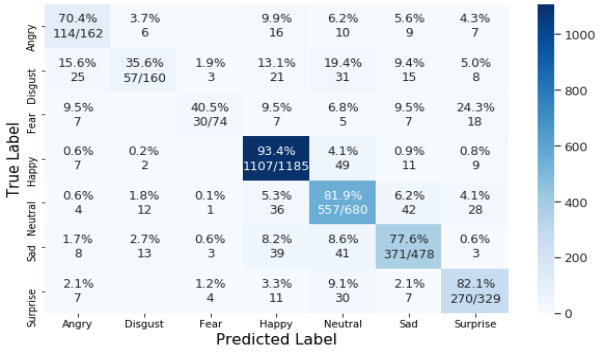
Method	Source	Target	Accuracy
CNN-Li [32]	RAF-DB	CK+	78.00%
MoEffNet (ours)	RAF-DB	CK+	78.59%

Table.4. Trained model cross dataset comparison on FER-2013

Method	Source	Target	Accuracy
CNN-Li [32]	RAF-DB	FER-2013	55.38%
MoEffNet (ours)	RAF-DB	FER-2013	56.10%



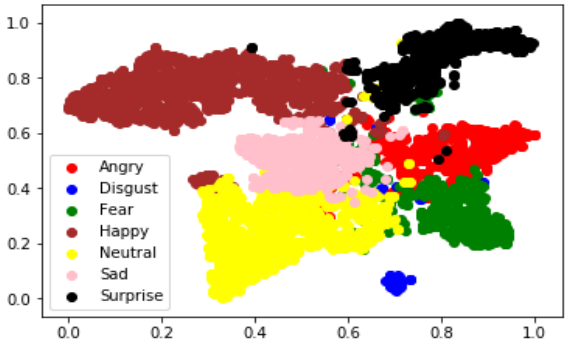
(a)



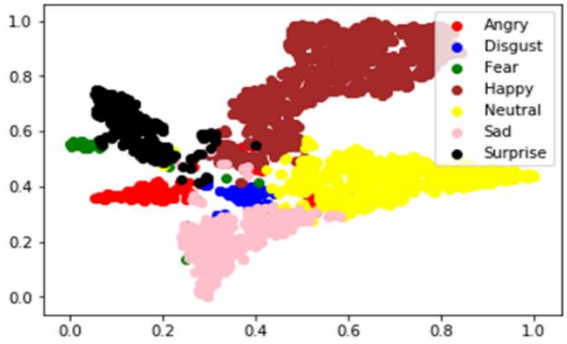
(b)

Fig.4. Confusion matrix of MoEffNet on (a) FER-2013 dataset (b) RAF-DB dataset. The darker the color higher the accuracy

On the FER2013 and RAF-DB datasets, the t-SEN (Stochastic Neighbor Embedding) approach is used to visualize the model's learned in-depth features. The t-SEN function converted the high-dimensional features into nearby embedded low-dimensional points. As a result, Fig.5 shows clusters of various emotional expressions for both datasets. The model learns from the silent areas of the facial regions, which play an essential role in FER. The discriminative power of the model is proportional to the inter-class separability in the t-SEN plot.



(a)



(b)

Fig.5. Models t-SEN distribution on (a) FER-2013 test data (b) RAF-DB test data.

Fig.6. Samples of misclassified disgust expressions

The Fig.4 shows that the model gets the most confused between Fear and Sad emotions, with 12.8 % in FER-2013. Whereas in RAF-DB model is most confused between Fear and Surprise emotion with 24.3%. Fig.6 highlights numerous circumstances in which the RAF-DB model fails to distinguish Disgust expressions. The poor performance of the model can be attributed by two factors. First, because there are fewer samples to train for a given emotion, for example, disgust is one of the minority sample counts in the RAF-DB dataset, followed by sad emotion. Second, a variety of factors, such as face occlusion,

random brightness, non-frontal faces, and so on, may have a negative impact on model performance.

6. CONCLUSION

We have proposed an efficient convolution network that uses EffIcientNet as its backbone architecture. The backbone architecture was pre-trained on the ImageNet dataset. We have retrained our proposed deep convolution network and generalized it well on openly available databases, i.e. RAF-DB, FER-2013, CK+. The trained model performs well in the inner database and cross-database comparison with the other state-of-the-art CNNs.

In the future, this work can be extended to video sequences where these modified models were used for the efficient findings of the changes in the emotional state of humans.

REFERENCES

[1] P. Ekman, "Universal Facial Expressions of Emotion", *Mental Illness Journal*, Vol. 8, pp. 151-158, 1970.

[2] P.S. Suchitra and S. Tripathi, "Real-Time Emotion Recognition from Facial Images using Raspberry Pi II", *Proceedings of International Conference on Signal Processing and Integrated Networks*, pp. 666-670, 2016.

[3] A. Fernandez-Caballero, R. Zangroniz and J.M. Latorre, "A Smart Environment Architecture for Emotion Detection and Regulation", *Biomed Research International*, Vol. 2016, pp. 55-73, 2016.

[4] U. Thonse and S.K. Sharma, "PSVN Facial Emotion Recognition, Socio-Occupational Functioning and Expressed Emotions in Schizophrenia versus Bipolar Disorder", *Psychiatry Research*, Vol. 264, pp. 354-360, 2018.

[5] N.K. Mehta and S. Singh, "Three-Dimensional DenseNet Self-Attention Neural Network for Automatic Detection of Student's Engagement", *Applied Intelligence*, Vol. 18, pp.1-21, 2022.

[6] M.Z. Alom and V.K. Asari, "A State-of-the-Art Survey on Deep Learning Theory and Architectures", *Electronics*, Vol. 8, pp. 292-298, 2019.

[7] M. Sahu and R. Dash, "A Survey on Deep Learning: Convolution Neural Network (CNN)", *Proceedings of International Conference on Smart Innovation, Systems and Technologies*, pp. 317-325, 2021.

[8] X. Zhao and S. Zhang, "Facial Expression Recognition via Deep Learning", *IETE Technical Review*, Vol. 32, pp. 347-355, 2015.

[9] M.A. Akhand and T. Shimamura, "Facial Emotion Recognition using Transfer Learning in the Deep CNN", *Electronics*, Vol. 10, No. 9, pp. 1036-1045, 2021

[10] C.F. Liew and T. Yairi, "Facial Expression Recognition and Analysis: A Comparison Study of Feature Descriptors", *IPSJ Transactions on Computer Vision and Applications*, Vol. 7, pp. 104-120, 2015.

[11] H. Alshamsi and H. Meng, "Stacked Deep Convolutional Auto-Encoders for Emotion Recognition from Facial Expressions", *Proceedings of International Conference on Neural Networks*, pp. 1586-1593, 2017.

[12] M. Oquab and J. Sivic, "Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1717-1724, 2014.

[13] I.J. Goodfellow, "Challenges in Representation Learning: A Report on Three Machine Learning Contests", *Proceedings of International Conference on Neural Networks*, pp. 117-124, 2013.

[14] 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, 2018.

[15] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 94-101, 2010.

[16] X. Zhao and S. Zhang, "Facial Expression Recognition via Deep Learning", *IETE Technical Review*, Vol. 32, pp. 347-355, 2015.

[17] H. Ding, S.K. Zhou and R. Chellappa, "FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition", *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 118-126, 2017.

[18] S. Shaees and H. Aldabbas, "Facial Emotion Recognition using Transfer Learning", *Proceedings of International Conference on Computing and Information Technology*, pp. 1-5, 2020.

[19] Y. Fan, J.C. Lam and V.O. Li, "Multi-Region Ensemble Convolutional Neural Network for Facial Expression Recognition", *Proceedings of International Conference on Artificial Neural Networks*, pp. 84-94, 2018.

[20] S. Zhao, H. Cai, H. Liu, J. Zhang and S. Chen, "Feature Selection Mechanism in CNNs for Facial Expression Recognition", *Proceedings of International Conference on Computer Vision*, pp. 317-328, 2018.

[21] K. Wang and Y. Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition", *IEEE Transactions on Image Processing*, Vol. 29, pp. 4057-4069, 2020.

[22] K. Wang, "Suppressing Uncertainties for Large-Scale Facial Expression Recognition", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6897-6906, 2020.

[23] G. Wen, Z. Hou, H. Li, D. Li, L. Jiang and E. Xun, "Ensemble of Deep Neural Networks with Probability-Based Fusion for Facial Expression Recognition", *Cognitive Computing*, Vol. 9, pp. 597-610, 2017.

[24] R. Breuer and R. Kimmel, "A Deep Learning Perspective on the Origin of Facial Expressions", *Proceedings of International Conference on Computer Vision*, pp. 1-4, 2017.

[25] J. Cai and Z. Li, "Probabilistic Attribute Tree in Convolutional Neural Networks for Facial Expression Recognition", *Proceedings of International Conference on Computer Vision*, pp. 1-9, 2017.

[26] C.J.L. Flores, A.E.G. Cutipa and R.L. Enciso, "Application of Convolutional Neural Networks for Static Hand Gestures Recognition under Different Invariant Features", *Proceedings of International Conference on Electronics, Electrical Engineering and Computing*, pp. 1-4, 2017.

[27] I. Rocco, R. Arandjelovic and J. Sivic, "Convolutional Neural Network Architecture for Geometric Matching", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 39-48, 2017.

[28] L.C. Yan and B. Yoshua, "Deep Learning", *Nature*, Vol. 521, pp. 436-444, 2015.

[29] M. Tan and Quoc V. Le, "Efficientnet: Improving Accuracy and Efficiency through AutoML and Model Scaling", *Proceedings of International Conference on Computer Vision*, pp. 1-9, 2019.

[30] A.P. Fard and M.H. Mahoor, "Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild", *IEEE Access*, Vol. 10, pp. 26756-26768, 2022.

[31] R. Pecoraro and S. Gallo, "Local Multi-Head Channel Self-Attention for Facial Expression Recognition", *Proceedings of International Conference on Computer Vision*, pp. 1-14, 2021.

[32] S. Li and W. Deng, "A Deeper Look at Facial Expression Dataset Bias", *IEEE Transactions on Affective Computing*, Vol. 13, No. 2, pp. 881-893, 2020.