Electronics, Communications and Networks
A.J. Tallón-Ballesteros et al. (Eds.)

doi:10.3233/FAIA231252

Improving Facial Emotional Recognition Using Convolution Neural Network with Minimal Layer

Rogerant Tshibangu [a,1] and, Jules R Tapamo [b]
[a] *Electrical Engineering Department, Mangosuthu University of Technology, Durban, South Africa, 4031*
[b] *Discipline of Electrical, Electronic and Computer Engineering, University of Kwazulu-Natal, Durban, South Africa, 4041*

Abstract. Human emotions identification has many applications, including human-computer interaction, illogical analysis, medical diagnosis, data-driven animation, and human-robot interaction. This paper presents a classification model, ConvNet that extracts features from facial images using techniques such as local binary patterns (LBP), convolutional neural networks (CNN), and region-based oriented FAST and rotational BRIEF (ORB). This model converges quickly. Experiment show that ConvNet outperforms existing methods with a precision of 98.13% on the CK+ dataset and 92.05 % on the JAFFE dataset.

Keywords. Face Detection, Facial Emotion Recognition, Convolutional Neural Network, Local Binary Patterns, Deep Learning

1. Introduction

The mental core is another name for the face. Faces can send subtle signals in the form of different facial expressions. If computers can decipher these subtle signals, human-machine interactions can be made safer and more harmonious[1]. People's true feelings are often revealed by their facial expressions. Facial emotion recognition (FER) is an important non-verbal mechanism used in human-machine interface (HMI) systems to better understand a person's personal emotions and intentions [2].

Some of the the still image-based feature extraction techniques that have been intensively used are Active Presence Model (AAM) , Local Binary Pattern (LBP), Haar wavelet transform and Gabor wavelet transform [3] [4]. On the other hand, the dynamics-based approach [5] assumes temporal relevance in the sequence of input facial expressions in attached frames. Common facial expression recognition algorithms include Hidden Markov Models (HMM), Artificial Neural Networks (ANN), AdaBoost and Support

[1]Corresponding Author: Electrical Engineering Department, Mangosuthu University of Technology, Durban, 4031, South Africa; E-mail: tshibangu.rogerant@mut.ac.za

Vector Machines (SVM) [6]. CNN implementations in particular and the most important developments in deep learning have shown promise computer vision.

However, a major challenge associated with using deep learning lies in the requirement of substantial volumes of data to effectively train a model successfully. Although some progress has been made in facial expression recognition by CNN models, some challenges remain, such as long training times and poor recognition rates in complex environments. There are two problems with the deep learning performance of the FER method: (a) a small number of photographs, and (b) images taken in highly organized environments. These concerns influence the development of FER method [7].

The purpose of this work is to identify emotions from an input photo of facial expressions. The authors present a comprehensive analysis of deep learning techniques applied to both static and dynamic facial expression recognition (FER) tasks through 2020. To increase the accuracy of this goal, this research develops an automatic facial recognition recognition (AFER) system using convolutional neural networks (CNNs) [8] . Several current machine learning algorithms are typically used for hand-crafted functions, but have non-persistent equivalents to consistantly interpret the task. The fusion of LBP-ORB and CNN offers interesting and promising opportunities for further exploration and exploitation in various applications. This is because the CNN-based model [9] offers the best solution for current FER-related tasks. Face recognition requires several stages. Facial image capturing, facial image preprocessing, facial feature extraction or learning, facial image registration, and facial image identification.

Most Facial expression recognition extracting geometric features and/or models involve statistical features. This paper aims to develop a CNN based FER model. The presented model can be used to classify human faces live. This paper investigates the following:

- CNNs based model that detects both basic and live facial expressions with the aim to improve the detection performance and produce better results by integrating the local binary pattern function (LBP) and incorporating a fusion of Oriented FAST and Rotated BRIEF (ORB).
- A 'ConvNet' model with four layers is proposed that takes full advantage of the parameters of convolutional neural networks.
- Incorporating advanced testing techniques, including thorough preparation, testing and validation processes, enables consistent results to be achieved over long training periods.
- The scalability of "ConvNet" via the assessment of the performance using various sizes of datasets

This paper starts with a literature review, followed by the proposed architecture, this is then followed by the experiments and results and ending with a conclusion and future work.

2. Literature review

In this section, we briefly review the literature related to facial emotion recognition, Oriented Fast and Rotated Brief and CNN. A study using LBP (Local Binary Patterns) and

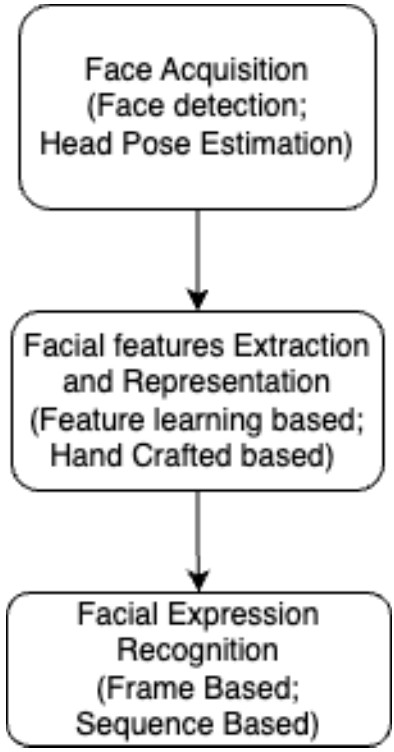

Figure 1. A Basic Framework for Various Field Applications for Automatic Facial Expression Analysis

ORB (Oriented FAST and Rotated BRIEF) for facial expression recognition was presented [10]. Evaluation of the performance of these feature extraction methods compared to SIFT (Scale-Invariant Feature Transform) and HAAR (Haar-like features) shows that the proposed methods have better performance for facial expression recognition.

In the Automatic Facial Expression Analysis (AFEA), its vast potential is evident across diverse sectors such as clinical psychology, neurology, human-computer interactions, and even sophisticated applications like lie detection. As illustrated in Figure 1, the foundational methodology of AFEA can be delineated into three pivotal steps: firstly, the face collection phase, where high-quality facial imagery is gathered; secondly, the crucial phase of facial data extraction and representation, where pertinent facial features are extracted from the raw data; and finally, the recognition stage where these features are processed through advanced algorithms to classify and recognize the facial expression [11]. Notably, within the realm of facial feature extraction, two primary techniques stand out: the geometric or predictive feature-based methods, which zero in on key facial landmarks, and methods rooted in hand-crafted techniques that focus on manually curated features and patterns.

In [12], a study is done on the use of Convolutional Neural Network (CNN) features for facial expression recognition. A deep learning approach is used in this study by training a CNN model on a facial expression dataset and predict emotion. Experiments conducted suggest that deep learning approaches are well-suited to facial expression recog-

nition and can lead to improved performance compared to traditional feature extraction methods.

A conventional DNN architecture is proposed in [13] to learn deep features from raw data. The results show that the proposed method effectively learns deep features and can attain comparable or even higher performance compared to many existing methods.

In [14], the proposed approach trains a CNN model on EEG signals and effective dimension measurement (EFDM) to capture the eigenstructure of EEG data. The model extracts relevant features from the EEG signal and feeds them into a sentiment prediction classifier. Experimental results show that the CNN model combined with EFDM outperforms some existing methods in terms of accuracy and efficiency.

EEG signals and EFDMs has been used to capture the underlying structure of the EEG data, which are then used to train a CNN model for emotion recognition. The CNN model extracts features from the EEG signals and predicts emotions based on the extracted features. The results show that the proposed method based on EEG signals and EFDMs outperforms other methods in terms of accuracy and efficiency [15].

An overview of the existing literature in the field identifying trends, challenges, and future directions are presented in [16]. Convolutional Neural Networks (CNNs), have recently shown promising results in the field.

In [17], a new approach is introduced to detect facial emotions in videos using deep neural networks that combine convolutional neural networks (CNN) and long short-term memory (LSTM) networks. The aim was to improve the performance of facial emotion detection in videos by using a deep neural network. The proposed deep neural network consists of two main components: a CNN for feature extraction and an LSTM for sequence learning. The CNN is trained to extract features from each video frame, while the LSTM is trained to learn the temporal relationships between the features extracted by the CNN. The results show that the proposed deep neural network outperforms existing methods with respect to accuracy, precision, and recall.

Hans and Smitha [18] proposed using deep CNNs to extract features from facial images for facial emotion recognition. The method consists of training a deep CNN to extract features from facial images and then using these features to classify the emotions in the images. The results indicate that the proposed method surpasses existing methods in terms of accuracy, precision, and recall. The authors further conduct an ablation study to exhibit the individual contributions of each component in the proposed method towards its overall performance.

It can be observed that facial recognition analysis has several drawbacks. For example, lack of model use and familiarity, inability to capture emotions and behavior in complex situations, brevity of participants seeking greater accuracy, and lack of ability to recognize effectiveness. Ongoing research includes impact identification and usability testing integration, but additional analysis is needed to assess feature applicability and potential user experience testing failures.

3. Proposed architecture and methods

The model shown in Figure 2 contains four convolutional layers and two fully connected layers. The vector obtained from each filter after convolution is the feature vector [19]. The convolutional feature map is created by fusing the LBP feature map with the feature

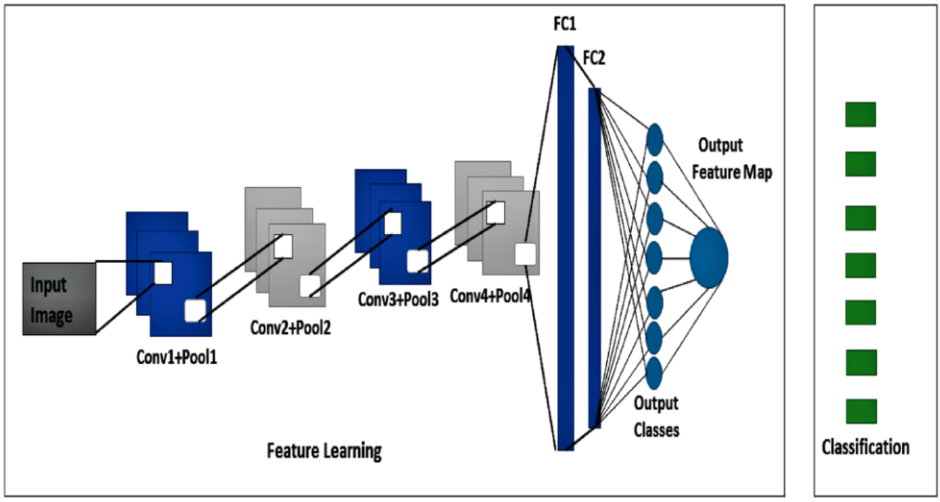

Figure 2. Proposed architecture for CNN model [19]

vector map. Convolutional layers have weights that need to be learned, while pooling layers use fixed functions to transform activations. ReLU (Rectified Linear Unit) is used to introduce nonlinearity into the network architecture while preserving the received field of the convolutional layers. The efficiency of convolutional layers is achieved by grinding. Furthermore, the model's output is derived by evaluating the training set using the loss function, and the learning parameters, including kernels and weights, are iteratively fine-tuned by backpropagation to optimize the loss. This investigation includes including or excluding specific transforms during the training process. A max pooling layer or a convolutional layer. Unique and valuable model outputs are produced by applying specific kernels and weight configurations. These outputs are subject to pooling operations. This process involves nonlinear downsampling in the spatial dimension. Pooling effectively reduces the spatial size of the representation, minimizing parameters and computations to prevent overfitting. The final step transforms the pooled feature map, which was originally a 2D structure, into a 1D flattened vector. The result is a pooled feature map that has been flattened.

Proposed modification of CNN. The fine-tune method replaces the fully-connected layers of a pre-trained model with a new set of fully-connected layers for training on a given dataset and fine-tunes it. Kernels based on convolutional layers are fine-tuned in whole or in part using backpropagation. The suggested architecture in this study consists of four convolutional layers and two fully connected layers. For this task, only detailed high-level functional blocks and fully connected layers to be considered as classifiers should be trained. In contrast, in humans, he only has 7 emotions, so the author resets Softmax's ranking from 1000 ranks to 7 levels.

Proposed CNN pipeline. In this configuration, the architecture contains four layers of convolutions with pooling followed by two fully connected layers. Each convolutional fully connected layer in all four networks is equipped with ReLU, batch normalization, and dropout layers. After four convolutional layers, an ultra-dense layer is used in com-

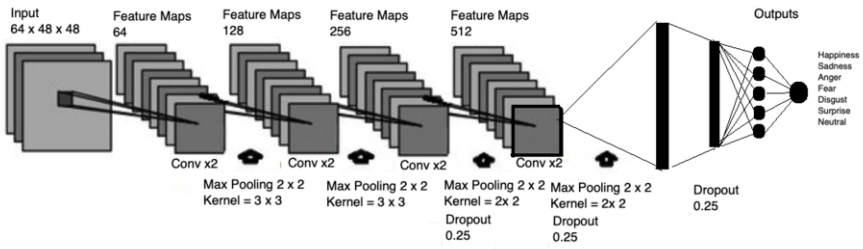

Figure 3. System diagram of the proposed CNN architecture

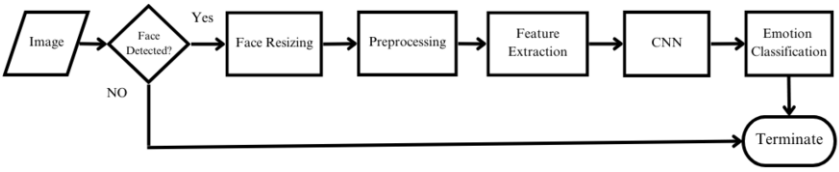

Figure 4. Flowchart diagram of the proposed method for emotion grouping

bination with two fully connected layers. The diagram in Figure 3 shows the complete pipeline of the proposed CNN model.

Preprocessing: Data records are then preprocessed into a suitable format. This step involves performing various preprocessing operations such as resizing, normalization, and other transformations to ensure that the dataset is in a standardized, optimal form for analysis. After preprocessing the dataset, a generalized algorithm is applied to obtain effective results. These algorithms use preprocessed datasets to extract meaningful insights and patterns related to facial emotion recognition.

Feature extraction: This step involves applying face detection algorithms to identify facial regions in images and extract facial features or landmarks. After feature extraction, the CNN algorithm is used for emotion classification. A trained CNN model is loaded, and the extracted features are fed into the model to predict emotion class labels. The facial emotion detection mechanism is implemented once all necessary functions and procedures have been defined.

Facial expression recognition: Features are then extracted for each real-time image and used for sentiment classification using a CNN algorithm. Predicted emotions are printed or saved for further analysis.

Face detection: Face detection is performed using a method proposed by Viola and Jones [20]. A classifier that finds objects in photos and videos that contain many positive and negative images learns a cascade function. Moreover, visual object recognition using Haar cascades has been shown to be effective and accurate. Hair features identify three

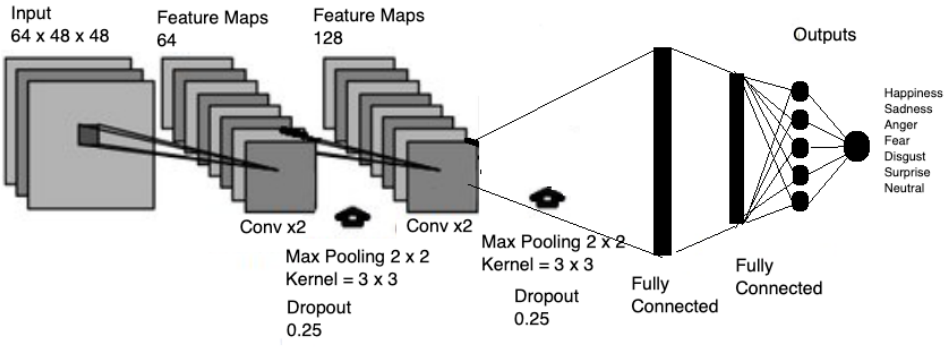

Figure 5. Improved design of the proposed CNN model

dark areas on the forehead as part of facial recognition. For example B. Eyebrows. Using the Haar cascade classifier, unwanted background information is efficiently removed from images, allowing accurate identification of facial regions. This process has been implemented by adapting the Haar cascade from OpenCV.

A combination approach: Before feature fusion, normalization is performed to improve the recognition rate. In this study, the LBP features are rescaled to the range 0-1, and the LBP and ORB features are normalized using the following formulas:

$$X = \frac{x}{max(x)} \tag{1}$$

here x represents the feature value. Since no single feature type can cover all visual features, feature fusion is often used to combine multiple features. This article uses the Z-score approach to fuse the LBP and ORB descriptors.

$$\sigma = \frac{1}{J}\sum_{i=1}^{J}(f_i - \mu)^2 \tag{2}$$

$$\mu = \frac{1}{J}\frac{\sum_{i=1}^{J} f_j}{j} \tag{3}$$

$$\hat{f}_j = M\frac{(x_j - \mu)}{\sigma + C} \tag{4}$$

here fj is an LBP or ORB feature, and $\hat{\text{f}}$j is the fusion feature data. In the following trials, M is a factor multiplied by $\hat{\text{f}}$j, and M is 100.

Design Improvements and contributions:

To improve system design and performance, the authors made some changes from the original architecture by removing two convolutional layers, as shown in Figure 5. This adapts the network structure to consist of an input layer followed by his two convolutional layers and an additional pooling layer.

In addition, as already mentioned, the model contains several important layers to improve its functionality. These include a ReLU layer, a batch normalization layer, and a dropout layer. A ReLU layer introduces nonlinearity into the network, allowing the

model to capture complex patterns and relationships in the data. A batch normalization layer normalizes the intermediate feature maps and helps stabilize and speed up the training process. Finally, the dropout layer randomly zeros out some of the input units during training to prevent overfitting and improve the model's ability to generalize.

By incorporating these layers, the modified network architecture benefits from nonlinearity, normalization, and regularization techniques that help improve model performance and robustness in facial emotion detection tasks.

4. Experiments and results

4.1. Datasets:

The experiments performed in this study uses FER2013, JAFFE and the Extended Cohn-Kanade (CK+) datasets. Note, however, that the resolution of the FER2013 dataset is limited to 48×48 pixels. This property helps evaluate the performance under low-light conditions. Datasets contain specific subsets for validation and testing purposes. The JAFFE and FER2013 datasets consist only of grayscale images, whereas the CK+ dataset offers the possibility to use both RGB and grayscale images.

The FER2013 dataset consists of 35,887 images, divided into 28,709 shots of trains and 3589 validations. The dataset also contains 3589 for the final test.

4.2. Results

Table 1. Accuracy comparison with related network

Algorithm	***Accuracy range***
Alexnet	55-88
VGG	65-68
GoogleNet	82-88
Resnet	72-74
FER (our proposed)	85-98

Evaluation of the performance of the presented facial emotion detection system will involve three metrics: Precision, specificity, and sensitivity. These are calculated using true positive (TP), true negative (TN), false positive (FP), and false negative (FN) given below:

$$Accuracy = \frac{TN + TP}{TP + FP + FN + TN} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

Table 1 presents a brief comparison of the proposed approach and other relevant studies. The table clearly shows that the CNN approach outperforms alternative tech-

Figure 6. Training and validation loss across epochs

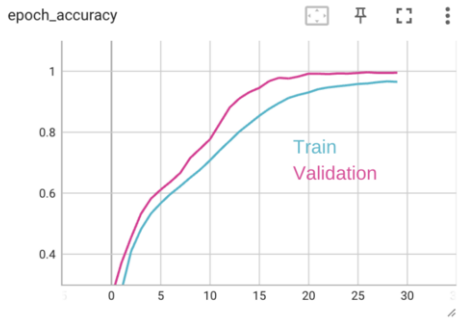

Figure 7. Training and validation accuracy across epochs .

niques and approaches to human emotion detection and that our proposed model shows superior results.

Experiments with FER2013 produced a confusion matrix shown in Table 2. It can be seen that the accuracy is high. Bold numbers on the diagonal indicate the number of samples well classified. The numbers on each side of the diagonal line also indicate the number of incorrectly listed images. Since these numbers are smaller than those on the diagonal, we can conclude that the algorithm performed well.

Table 2 shows the performance achieved by the CNN model trained on the FER2013 dataset. It clearly shows that the accuracies of both the training and validation improve significantly as the number of epochs increases.

Figure 6 shows corresponding training and validation loss. This means that training loss decreases and validation loss decrease as epochs increase. Furthermore, we expect the validation loss to decrease as time increases and therefore, this model is well-suited for training results.

Validation accuracy must be equal to or slightly lower than preparation accuracy to yield an improved system. In Figure 7, as the epochs increased, finally introducing some convolutional layers. As a result, the network is larger and more diverse, and the training accuracy was marginally greater than the validation accuracy. The validation loss still outweighs the lack of preparation. In Figure 6, as the number of epochs increases, the training and validation loss decreases. Additionally, we expect the validation loss to decrease as time increase. At higher epochs, we would expect a lower percentage

Table 2. Confusion matrix

	Anger	*Disgust*	*Fear*	*Happiness*	*Sadness*	*Surprise*	*Neutrality*
Anger	**91.2**	0.3	0.3	0	4.3	3.9	2.6
Disgust	5.4	**89.6**	3.1	2.7	1.1	0	2.5
Fear	0.25	2	**88.5**	3.3	5.8	1.17	0.8
Happiness	1.5	0	2.1	**95.9**	0	0	0.5
Sadness	1.1	5.4	4.1	0	**88.8**	0	0.7
Surprise	0.25	3.1	0.64	0	0.54	**94.5**	1.1
Neutrality	2.11	3.9	3.28	0.31	0.78	0.45	**92.4**

of validation loss compared to the training loss observed in the final stage of mapping. Therefore, this model is suitable for achieving good training results.

To improve the results we removed two convolution layers to obtain the proposed design, as stated earlier. This new design improved the training accuracy to over 98 percent. Figure 6 and Figure 7 show the trend of the results achieved. As mentioned earlier, The results show that removing convolutions from the model leads to improvement.

The improvement in accuracy is shown in Figure 6 and Figure 7. These numbers indicate accuracy values over the course of training epochs or iterations. Figure 6 shows the training accuracy, showing the evolution of the model's performance on the training data. We can observe a steady improvement in accuracy as training progresses, indicating that the model can effectively learn from the training data.

Fig 7, on the other hand, shows the validation accuracy, which provides insight into the generalizability of the model. Validation accuracy describes the model's performance on unknown data and is important in evaluating how well the model generalizes to new examples. In this figure, we can observe a consistent improvement in validation accuracy through training iterations. This indicates that the modified model succeeded in improving its ability to generalize unseen data. The results shown in Figure 6, and Figure 7 collectively support the conclusion that deconvolution significantly improved both training and validation accuracy. These results highlight the potential benefits of changing the model architecture and serve as a basis for further investigation and optimization.

5. Conclusion and Future work

The proposed model aimed to integrate the features of LBP, ORB, and CNN to enable accurate facial expression recognition. The model showed excellent accuracy when applied to both the CK+ and FER2013 datasets, outperforming several recent methods. This highlights the effectiveness and potential of the proposed approach.

Although the proposed model shows promising results, there is still room for improvement. Further research should focus on optimizing models that enable a more natural and differentiated approach to facial expression recognition.

The ultimate goal is to develop a robust and reliable model that can accurately analyze facial expressions in real-world scenarios, thus providing a more complete understanding of human emotional states.

References

[1] Ekman, R. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS) (Oxford University Press, 1997).

[2] Nwosu, L.,Wang, H., Lu, J., Unwala, I., Yang, X., Zhang, T. Deep convolutional neural network for facial expression recognition using facial parts. In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/ PiCom/DataCom/Cyber SciTech) 1318–1321 (IEEE, 2017).

[3] Talegaonkar, I., Joshi, K., Valunj, S., Kohok, R., Kulkarni, A. Real-time facial expression recognition using deep learning. Available at SSRN 3421486 (2019).

[4] Kumar, N. and Bhargava, D. A scheme of features fusion for facial expression analysis: A facial action recognition. J. Stat. Manag. Syst. 20(4), 693–701 (2017).

[5] Zhao, G. and Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. 29(6), 915–928 (2007).

[6] Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N., Yan, S. Peak-piloted deep network for facial expression recognition. In European Conference on Computer Vision 425–442 (Springer, 2016).

[7] Ahmadinia, M. et al. Energy-efficient and multi-stage clustering algorithm in wireless sensor networks using cellular learning automata. IETE J. Res. 59(6), 774–782 (2013).

[8] Zhang, H., Jolfaei, A. and Alazab, M. A face emotion recognition method using convolutional neural network and image edge computing. IEEE Access 7, 159081–159089 (2019).

[9] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H. et al. Challenges in representation learning: A report on three machine learning contests. In International Conference on Neural Information Processing 117–124 (Springer, 2013).

[10] Niu, B., Gao, Z. and Guo, B. Facial expression recognition with LBP and ORB features. Comput. Intell. Neurosci. 2021, 1–10 (2021).

[11] Yu, Z. and Zhang, C. Image-based static facial expression recognition with multiple deep network learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction 435–442 (2015).

[12] González-Lozoya, S. M. et al. Recognition of facial expressions based on cnn features. Multimed. Tools Appl. 79, 1–21 (2020).

[13] Christy, A., Vaithyasubramanian, S., Jesudoss, A. and Praveena, M. A. Multimodal speech emotion recognition and classification using convolutional neural network techniques. Int. J. Speech Technol. 23, 381–388 (2020).

[14] Niu, H. et al. Deep feature learnt by conventional deep neural network. Comput. Electr. Eng. 84, 106656 (2020).

[15] Wang, F. et al. Emotion recognition with convolutional neural network and EEG-based EFDMs. Neuropsychologia 1(146), 107506 (2020).

[16] Wang, F. et al. Emotion recognition with convolutional neural network and eeg-based efdms. Neuropsychologia 146, 107506 (2020).

[17] Nonis, F., Dagnes, N., Marcolin, F. and Vezzetti, E. 3d approaches and challenges in facial expression recognition algorithms—A literature review. Appl. Sci. 9(18), 3904 (2019).

[18] Hans, A. S. A. and Smitha, R. A CNN-LSTM based deep neural networks for facial emotion detection in videos. Int. J. Adv. Signal Image Sci. 7(1), 11–20 (2021).

[19] Debnath T, Reza MM, Rahman A, Beheshti A, Band SS, Alinejad-Rokny H. Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity. Sci Rep. 2022 Apr 28.

[20] Viola, P. Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 (Vol. 1, pp. I-I). IEEE.