

Hybrid CNN-RF Algorithm for Facial Expression Recognition using Different Datasets

¹*Sarvajeet A. Bhosale and ²Dr. Sangeeta R. Chougule

Submitted: 29/01/2024 Revised: 07/03/2024 Accepted: 15/03/2024

Abstract: Facial emotions have gained a significant importance in the area of computer vision as one can naturally express feeling non-verbally using facial expressions. A human can express a many emotion, like angry, disgust, happy, afraid, sad, neutral and surprised each emotion has a set of components. Using various deep learning techniques, many researchers have been researching in this area for facial recognition. According to existing researches, emotions may vary for trained datasets and also change in image characteristics due to high shutter speed or real time video may occur. These changes may lead to incorrect result in Facial emotion recognition. To overcome this issue, new approach of a hybrid model is proposed for Facial Emotion Recognition (FER). A hybrid model using CNN and RF Classifier with MTCNN is designed to get better performance. The collection of data is the first step in recognising and classifying face expressions in this model. These data are pre-processed for removing the unnecessary data from the raw data and are extracted for features with the MTCNN (Multi-task Cascaded Convolutional Networks) by segmenting and localizing the image. CNN-RF hybrid architecture takes extracted features as input, where CNN architecture is used in the process of feature extraction and RF is used as a classifier. The performance metrics of the proposed architecture varies for different datasets but the overall percentage for accuracy, sensitivity, precision and error of FER are 88.5%, 88%, 82%, 11.5%. Thus, the designed model instantly recognizes the facial expression and classifies the gender in effective manner.

Keywords: Hybrid CNN-RF, FER, MTCNN, Lucy-Richardson, Contrast stretching

1. Introduction

Facial expressions contain emotional information on wider range and it is very useful for interpersonal communication. FER has become a biometric recognition technology due to its distinct features like non-contact, intuition, nature, safety and rapidity [1]. Emotional expression consists of facial expression, voice and language. From this 100% emotion expression, Facial expression has 50% of information, voice information is 40% and remaining is for language and other [2]. Now a days to Facial expressions are studied on larger extent by the researchers as it has gain more importance in AI and human machine interaction. Generally, Static and dynamic features both are considered as an input during the FER system design [3]. The system works in three steps: detection of face, extracting features and classifying expressions.

The main objective of face detection is to check face part is present or not in the captured image and sometime it is also useful in to employ pre-processing techniques, if required. Detected faces contain the facial features which are useful in facial emotion recognition [4]. MTCNN extracts the face frame from image for recognizing the

FER which is recently considered as the dominant multi-stage structure and multi-task structure. Variation of pixel is found to more in face detection, it may vary from few pixels to large pixels than that of object detection [5]. MTCNN uses image pyramid method which may not be more suitable on tiny faces and the faces with higher degree of variations in scale. The basic is in first stage, P-net structure generates candidate window by using shallow CNN. The high discriminative features cannot be extracted from the shallow P-Net, which comes from deep neural network [6]. A loss may occur in the robustness and feature discrimination because of setting the kernel at standard convolution. This is caused due to sampling at same centre and at same size.

FER Classification based on Deep Learning has gained wide attention. Face expressions are key tools for expressing and conveying an individual's emotional reaction and state during daily activities [7]. Using various techniques, many researchers have been researching in this area for facial recognition. Computer Vision field is emerging with the help of deep learning architectures resulting in many research work by adopting big data [8]. CNN reduces the dependency of other pre-processing algorithms [9]. The hybrid approach will also help in increasing the performance rate in FER [10]. Thus, this paper proposes a novel approach of FER. Proposed model has following major aspects:

¹*Department of Electronics Engineering, KITCOE Research Center, ShivajiUniversity, Kolhapur, India.

²Department of Electronics Engineering, KITCOE Research Center, ShivajiUniversity, Kolhapur, India.

¹*Corresponding Author Email: sarvajeet.bhosale@gmail.com

- Real time FER are done by using hybrid learning algorithm.
- Different datasets are used for the FER.
- Image resizing, Lucy-Richardson and contrast stretching techniques are used to pre-process the raw images from the dataset
- MTCNN is used for extracting the face frame features to reduce the classification burdens.
- Facial expression regression is done using hybrid CNN-RF.

The further paper is ordered as follows, second section is about Researchers contribution in FER. Brief description of proposed methodology is given in section 3. Discussion on results is done in section 4 and research is concluded in section 5.

2. Literature Survey

Many researches had been performed by various techniques to recognize the expression and classify the gender. Most of the existing techniques are studied and few techniques are reviewed below.

Minaee, S *et al.*, [11] proposed an attentional convolutional network model for FER which focuses on the key features and achieved a large improvement over the previous models. The model used different datasets. The model used visualization technique to find important facial regions for detecting Facial emotions based on the classifier's output which showed that different emotions are sensitive to different parts of the face.

Kim, J. H *et al.*, proposed a hierarchical deep neural network structure to fuse the extracted features with geometric structure. Holistic features are extracted using pre-processed LBP images. This extraction of feature is done using appearance based network. Geometry based features networks are used to determine changes in action units. The proposed model combines the result of two functions using the softmax function and used for consideration of error which is obtained from second highest emotion prediction result [12].

Georgescu, M. I *et al.*, proposed a model based on combined approach of two architectures, CNN and bag-of-visual-words [13]. The CNN model is used to learn the automatic features and hand-crafted features are obtained by bag-of-visual-words model. K nearest neighbour is used for training samples of input image, SVM classifier is used to predict the class of trained samples.

Zhang, S *et al.* proposed deep learning model, for learning affective video features for FER. The model is designed using two separate deep CNN using an video sequences using hybrid algorithms. These two deep learning architectures include a spatial features and

temporal features with individual deep convolution neural networks. Spatial features are employed to interpret facial images, and high-level and temporal information are learned using an optical temporal CNN. The deep belief network is used to incorporate the acquired features. To acquire global video features with a defined length, average pooling is utilised [14].

Fan, Y *et al.*, [15] proposed a model for FER automatically using the facial images. Essential local feature characteristics are recognised using the attention blocks it is seen that the model performance has improved when it was incorporated with deeply supervised framework. To ensemble the intermediate predicted score the characteristics of multiple convolution layers are combined in deeply supervised manner.

From the above reviews various techniques ACN [11] CNN [13], SVM-DBN [14], DSAN [15] are used for Facial expression recognition. Also different approaches have been followed for emotion recognition [16]. In those existing techniques, the facial frames are not correctly extracted for real-time FER with various datasets that leads to inaccurate prediction. Hence the hybrid CNN-RF model with MTCNN is designed for FER classification.

3. Proposed Methodology

Facial gestures and expression plays a vital role in FER which shows the position of muscles on face which contribute in expressing the feelings, opinions and other similar aspects in a non-verbal way. FER and Deep Learning-based Gender Classification have gained wide attention. A human can express a variety of emotions, such as happiness, sadness, disgust, fear, and so on; each emotion has a set of components. In this designed model, the hybrid CNN-RF model is designed with MT-CNN for accurate FER. Figure1 shows the block diagram of proposed model.

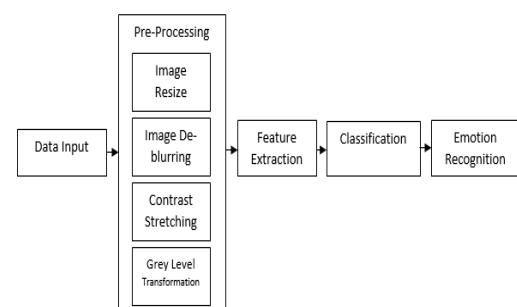


Figure 1: Block diagram of proposed model

First process in FER is to collect the data and sort according to the expression. The collected data is pre-processed with image resizing, Lucy-Richardson algorithm, contrast stretching and grey-scale transformation. The pre-processed data are feature extracted with the MTCNN (Multi-task Cascaded Convolutional Networks) by segmenting and localizing the facial organs such as eyes, nose, mouth, etc. These extracted features are given to the hybrid CNN-RF classifier. Hence the hybrid model works on CNN which is used for feature extraction and the RF is used for classification.

Data collection

Data collection is the primary process of collecting necessary information in the form of image which is to be predicted. The data is collected from camera and trained with the help of proposed model to detect different emotions.

Pre-processing:

The collected data is in the form of raw data and cannot be used directly with the model; the raw input will not produce the accurate result. Thus processing the raw data increases the model's accuracy.

Image Resizing:

Image resizing is the process in which the size of image is changed and it is reduced in pixel format. The proposed design model uses the pixel format with 256x256 pixels. All three datasets are resized in same order.

Lucy-Richardson algorithm:

Richardson and Lucy derived deconvolution method using bayes theorem, commonly used in Medical imaging [17]. The Lucy Richardson (LR) algorithm is used for image restoration using non-linear approach based on the maximum possibility formulation. In LR algorithm picture is represented using toxic statistic. The below mentioned iteration converges to Maximizing the likelihood function [17].

$$\left[\hat{f}_{k+1}(x, y) = \hat{f}_k(x, y) \left[h(-x, -y) * \frac{g(x, y)}{h(x, y) * \hat{f}_k(x, y)} \right] \right] \quad (1)$$

The amount of iteration cannot be predicted using this procedure. Stable solution can be obtained with the small PSF matrix. Few number of iteration will result in image softening and stable solution is obtained. If number of iterations is increased, computational process will slow down increasing noise. The equation for RL algorithm is given as mentioned below [17],

$$\left[f^{n+1} = f^n H * \left(\frac{g}{Hf^n} \right) \right] \quad (2)$$

Where f^{n+1} is the new estimate, f^n is previous estimate, H is the blur filter, H^* is the adjoint of H, n is the number of the step in the iteration and g is the blurred image [17].

Contrast Stretching:

Contrast stretching is used to extend the range of intensity values in a picture in order to increase contrast. The objective is to span a desired range of values, usually the entire pixel value range that the image type permits.

Grey-scale transformation:

Grey level transformation is used directly to process pixels hence most of processing techniques convert images in to grey levels. This grey level transformation uses 256 levels. The plot of histogram gives horizontal range from 0 to 255 and the vertical axis gives the number of pixels in the image. Greyscale conversion equation is given as below.

$$GF(x, y) = R * 0.299 + G * 0.587 + B * 0.114 \quad (3)$$

Feature Extraction:

In Feature extraction, process of reducing the dimensionality of the data and condensing the raw input data into a manageable group is carried out.

MTCNN:

The MTCNN generally works on two aspects as scaling and sorting. In scaling, according to the inputs to network lays the image is scaled to different sizes. Then based on the direction and accuracy three independent layers are sorted. Rough to fine principal is used while sorting images from poor to good. Hence a cascaded structure is obtained from the three convolution neural networks which provide multi task detection [19]. Three tasks such as detection of faces, boundary regression and point positioning of facial features can be done using MTCNN as these three tasks require different training labels and also require different loss functions.

MTCNN algorithm processes image in three cascaded network stages. Firstly the process of data pre-processing is carried out. Changes in face sizes are predicted by the algorithm obtained from the face pyramid then the scaling factor is used to scale original image and as per the input data from cascaded network image pyramid is constructed.

Stage one deal with the realization of face frame. Full CNN PNet uses all images from image pyramid to obtain face frame regression and face frame. The position of candidate face frame is obtained using face frame regression. Ten NMS (Non-maximum suppression) algorithm is used for merging candidate face frame having large overlap rate. Figure 2 illustrates the network structure of PNet.

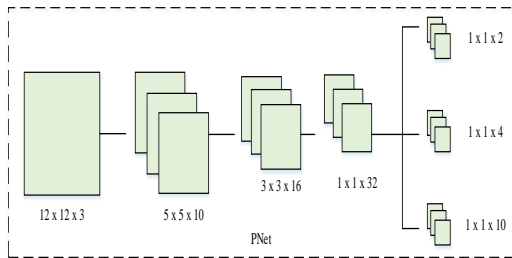


Figure.2 Structure of PNet

In second stage, RNet is used with candidate face frame input created from first stage. The network of RNet is more complicated than the PNet network. Fine tuning of face frames is done by removing the incorrect face frames; face frame regression vector is used for this fine tuning. To reduce the face frames NMS algorithm is used. Figure.3 shows the network structure of RNet.

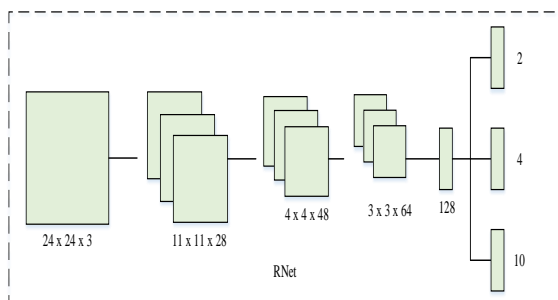


Figure.3 Structure of RNet

The third stage is ONet network structure. It is similar to the second stage i.e. as in second stage third stage also uses previous stage output as input. It is used to extract position coordinate information of obtained facial features. The ONet structure is more complex than that of RNet structure, but more accuracy is gained using this network. Figure.4 shows the network structure of ONet.

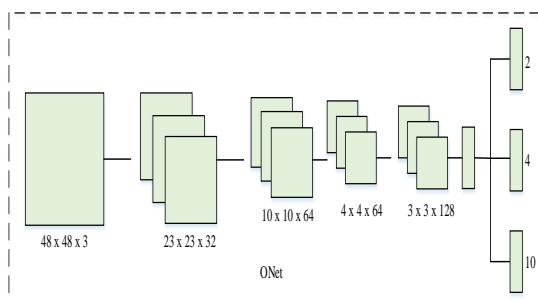


Figure.4 Structure of ONet

These combinations of three network structures will forms the network structure of MTCNN. From this MTCNN extracts accurate face frame, the extracted face frames are then given to the classifier for emotion prediction.

Classifier:

A classifier sorts data into labelled classes using some algorithms. In this designed model, hybrid CNN-RF model is used recognize facial expression.

Hybrid CNN-RF

While machine learning can only be used for classification, deep learning algorithms can be utilised for both feature extraction and classification. For machine learning, a separate feature extractor is therefore needed. The suggested design model uses RF for classification and CNN for feature extraction [20]. This is achieved by replacing the fully connected layer of the CNN to RF for classification. Thus, the combination of CNN and RF produces the hybrid CNN-RF model. Figure.5 illustrates the structure of hybrid CNN-RF

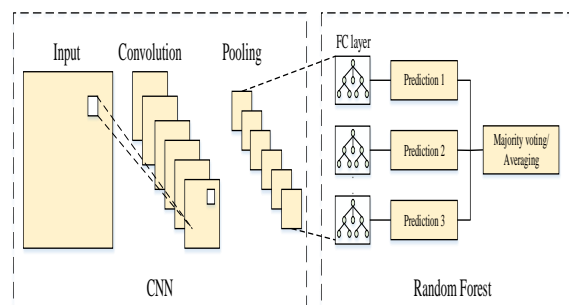


Figure.5 Structure of Hybrid CNN-RF

4. Results And Discussion:

The proposed CNN-RF model is designed for FER using three datasets such as CKPLUS, KMU-FED and KDEF. Data collection is the first process for FER. These collected datasets are pre-processed by resizing, deconvolution and contrast stretching. MTCNN is used for segmenting and localizing the facial organs from the image obtained by pre-processing. By using these segmented features, the hybrid CNN-RF classifies the facial expression. The collected output from different datasets are evaluated and compared.

Dataset

Dataset is used for training the classifier to predict the required output. In this designed model, four different datasets are used for comparison whether the recognition and gender classification of the classifier changes for different datasets. The CKPLUS datasets consist of 981 png files, KMU-FED consist of 1200 jpg files and KDEF consist of 490 jpg files for different facial recognition.

Table.1 Comparison of samples for original and feature extracted images using ckplus dataset





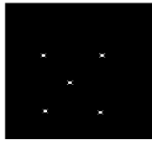















CKPLUS dataset				
Original image	Image resizing	Lucy-Richardson	Contrast stretching	MTCNN
				
				
				
				

Table.2 Comparison of samples for original and feature extracted images using KMU-FED dataset
















KMU-FED dataset				
Original image	Image resizing	Lucy-Richardson	Contrast stretching	MTCNN
				
				
				



Table.3 Comparison of samples for original and feature extracted images using KDEF dataset

KDEF dataset				
Original image	Image resizing	Lucy-Richardson	Contrast stretching	MTCNN

Table 1, 2 and 3 illustrates the samples of original and extracted images using different datasets. Extraction techniques used for this designed model are image resizing that resizes the image according to same order. Lucy-Richardson is used to de-blur the image. Contrast stretching enhances the contrast of the image and MTCNN is used to segment and localize the facial organs.

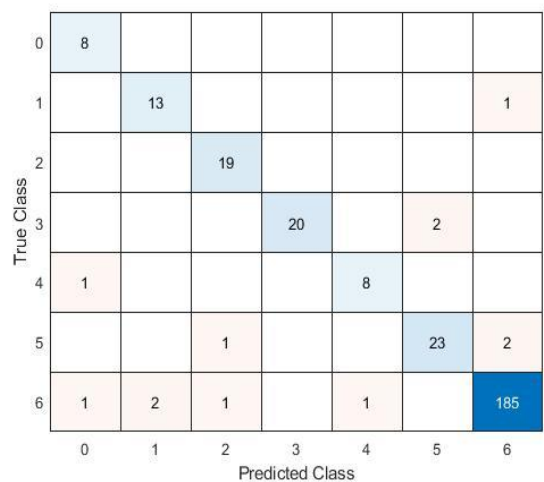


Figure 2: Confusion matrix for CKPLUS dataset

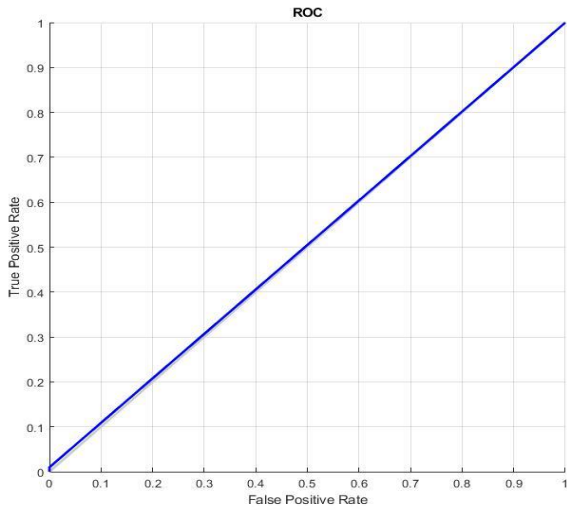


Figure 3: ROC from CKPLUS dataset

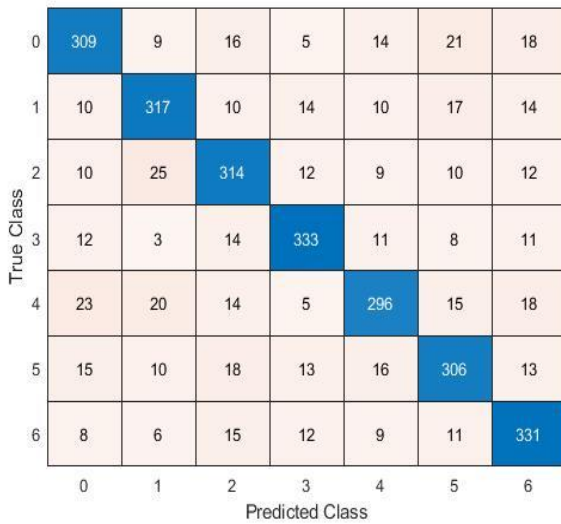


Figure 4: Confusion matrix from KDEF dataset

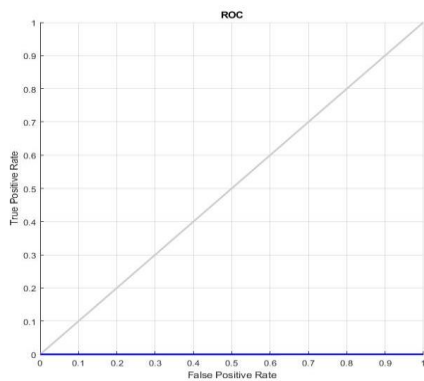


Figure 5: ROC obtained from KDEF dataset

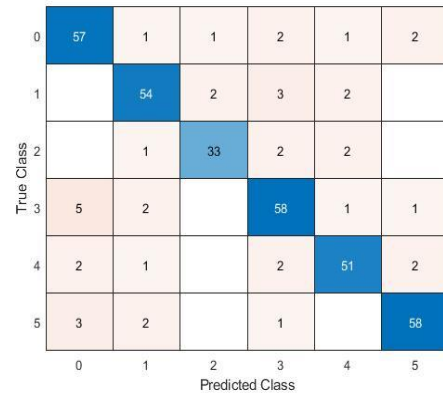


Figure 6: Confusion matrix from KMU-FED dataset

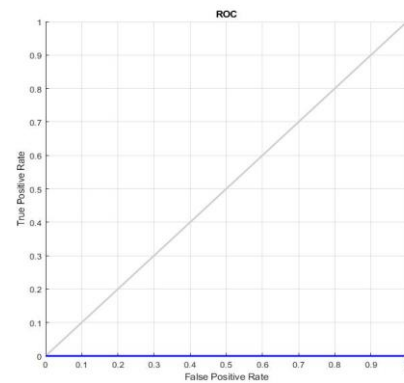


Figure 7: ROC from KMU-FED dataset

Figure 2, 4, 6 shows the confusion matrix. Confusion matrix is computed for three different datasets, representing the rows with predicted class and columns with the true class. The confusion matrix for different datasets explains the comparison of actual and predicted classes. The classes chosen for FER from 0 to 6 are afraid, angry, disgust, happy, sad, surprised and neutral. Figure 3, 5 and 7 shows the ROC (Receiver Operating Characteristic Curve) for FER using CKPLUS, KDEF, KMU-FED. ROC shows performance of a proposed model at thresholds. Two metrics, such as the true positive rate and the false positive rate, are plotted on this curve at various categorization thresholds. A lower categorization threshold will result in a greater number of positive classifications. As a result, true positives and false positives will be increased.

Table 4: Performance Parameters for FER

Parameters	CK+	KDEF	KMU-FED
Accuracy	0.9619	0.9183	0.9117
Error	0.0381	0.0817	0.0883
Sensitivity	0.9588	0.9080	0.9339

Specificity	0.9684	0.9280	0.9000
Precision	0.9841	0.9226	0.8309
False Positive Rate	0.0316	0.0720	0.1000
F1_score	0.9713	0.9152	0.8794

Table 4 shows the performance parameters for different datasets in detecting facial emotions.

Conclusion

FER classification is done by using the hybrid CNN-RF model with different datasets. The suggested model classifies data using feature extraction and pre-processing methods. The raw image from the dataset is pre-processed with image resizing, Lucy-Richardson and contrast stretching techniques. These pre-processed images are extracted for facial frame with MTCNN. Three different datasets are used for training the hybrid CNN-RF to compare whether the FER is having any incorrect recognition or gender classification. The accuracy of the proposed for accuracy, sensitivity, precision and error of FER classification are 88.5%, 88%, 82%, 11.5% which are greater compared to the exiting techniques with various datasets. In future, the real time FER can be done more accurate and precise with the modified CNN for extracting the facial features from different dataset.

Declaration Of Interests:

Funding

On Behalf of all authors the corresponding author states that they did not receive any funds for this project.

Conflicts Of Interest

The authors declare that we have no conflict of interest.

Competing Interests

The authors declare that we have no competing interest.

Data Availability Statement

All the data is collected from the simulation reports of the software and tools used by the authors. Authors are working on implementing the same using real world data with appropriate permissions.

Ethics Approval

No ethics approval is required.

Consent To Participate

Not Applicable

Consent For Publication

Not Applicable

Human And Animal Ethics:

Not Applicable.

Code Availability:

Not Applicable.

Reference:

- [1] Li, X., Yang, Z., & Wu, H. (2020). Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks. *IEEE Access*, 8, 174922-174930.
- [2] Wang, K., Peng, X., Yang, J., Meng, D., & Qiao, Y. (2020). Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29, 4057-4069.
- [3] Fan, Y., Lam, J. C., & Li, V. O. (2018, October). Multi-region ensemble convolutional neural network for facial expression recognition. In *International Conference on Artificial Neural Networks* (pp. 84-94). Springer, Cham.
- [4] Li, J., Jin, K., Zhou, D., Kubota, N., & Ju, Z. (2020). Attention mechanism-based CNN for facial expression recognition. *Neurocomputing*, 411, 340-350.
- [5] Vo, T. H., Lee, G. S., Yang, H. J., & Kim, S. H. (2020). Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, 8, 131988-132001.
- [6] Chen, Y., Wang, J., Chen, S., Shi, Z., & Cai, J. (2019, December). Facial motion prior networks for facial expression recognition. In *2019 IEEE Visual Communications and Image Processing (VCIP)* (pp. 1-4). IEEE.
- [7] Li, M., Xu, H., Huang, X., Song, Z., Liu, X., & Li, X. (2018). Facial expression recognition with identity and emotion joint learning. *IEEE Transactions on affective computing*, 12(2), 544-550.
- [8] Porcu, S., Floris, A., & Atzori, L. (2020). Evaluation of data augmentation techniques for facial expression recognition systems. *Electronics*, 9(11), 1892.
- [9] Zhang, H., Huang, B., & Tian, G. (2020). Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. *Pattern Recognition Letters*, 131, 128-134.
- [10] Li, H., & Xu, H. (2020). Deep reinforcement learning for robust emotional classification in facial

- expression recognition. *Knowledge-Based Systems*, 204, 106172.
- [11] Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9), 3046.
- [12] Kim, J. H., Kim, B. G., Roy, P. P., & Jeong, D. M. (2019). Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. *IEEE access*, 7, 41273-41285.
- [13] Georgescu, M. I., Ionescu, R. T., & Popescu, M. (2019). Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, 7, 64827-64836.
- [14] Zhang, S., Pan, X., Cui, Y., Zhao, X., & Liu, L. (2019). Learning affective video features for facial expression recognition via hybrid deep learning. *IEEE Access*, 7, 32297-32304.
- [15] Fan, Y., Li, V., & Lam, J. C. (2020). Facial expression recognition with deeply-supervised attention network. *IEEE transactions on affective computing*.
- [16] Sarvajeet A. Bhosale and Dr. Sangeeta R.Chougule(2023). A Review on Face Emotion Recognition using EEG Features and Facial Features. 2023 1st International Conference on Cognitive Computing and Engineering Education (ICCCEE), MIT ADT University Pune, India. Apr 27-29, 2023.
- [17] Inampudi, S., Vani, S., & TB, R. (2019). Image Restoration using Non-Blind Deconvolution Approach–A Comparison. *International Journal of Electronics and Communication Engineering and Technology*, 10(1).
- [18] Asokan, A., Popescu, D. E., Anitha, J., & Hemanth, D. J. (2020). Bat algorithm based non-linear contrast stretching for satellite image enhancement. *geosciences*, 10(2), 78.
- [19] Xie, Y., Wang, H., & Guo, S. (2020). Research on MTCNN Face Recognition System in Low Computing Power Scenarios. *Journal of Internet Technology*, 21(5), 1463-1475.
- [20] Sun, Y., Zhang, H., Zhao, T., Zou, Z., Shen, B., & Yang, L. (2020). A new convolutional neural network with random forest method for hydrogen sensor fault diagnosis. *IEEE Access*, 8, 85421-85430.
- [21] <https://www.kaggle.com/shawon10/ckplus>
- [22] <https://www.kaggle.com/tom99763/kdef-512x512-super-resolution-colored>
- [23] C. Li-Fen and Y. Yu-Shiuan, "Taiwanese facial expression image database," Brain Mapping laboratory, Institute of Brain Science, National Yang-Ming University, Taipei, Taiwan, Tech. Rep., 2007.
- [24] KMU-FED. Accessed: Feb. 13, 2020. [Online]. Available: <http://cvpr.kmu.ac.kr/KMU-FED.html>