IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

FACIAL EMOTION RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK

[1]Saji Kumar T.V, [2]Bijukumar K, [3]Prasobh P

[1,2,3]Assistant Professor
[1] Dept. of Electronics Engineering, [2,3] Dept. of Electrical Engineering,
[1,2,3] College of Engineering, Chengannur, Kerala, India.

Abstract: Emotions are one of the important factors that enhance the process of human communication. The facial expressions and gestures convey nonverbal communication cues that play in vital role in interpersonal relations. There is significant variability making facial recognition a challenging research area. Features like Histogram of Oriented Gradient (HOG) and Scale Invariant Feature Transform (SIFT) have been considered for pattern recognition. These features are extracted from images according to manual predefined algorithms. In recent years, Machine Learning (ML) and Neural Networks (NNs) have been used for emotion recognition. In this research, a Convolutional Neural Network (CNN) is used to extract features from images to detect emotions. The Python is used for implementation. A CNN model is trained with grayscale images from the MMI Face dataset to classify expressions into five emotions, namely happy, sad, neutral, fear and angry. To improve the accuracy and avoid overfitting of the model, batch normalization and dropout are used. The best model parameters are determined considering the training results.

***Index Terms* – Facial Action Coding System, Convolutional Neural Network, MMI Face dataset, Graphical Processing Units.**

I. INTRODUCTION

Facial expressions play a significant role in social communication since they convey a lot of information about people, such as moods, emotions, and other things. These are created by the movement of muscles in the face that attach to the skin and fascia, creating lines and folds and causing facial features like the mouth, eye, and brows to move. Researchers in computer vision motivated by a wide range of applications have become increasingly interested in designing and implementing automatic systems that recognize facial expressions. Such systems are very needed to substantiate and expedite the analysis of human behavior in a digital environment. Facial Expression Recognition (FER) is an essential tool to detect emotions and has been widely used in many aspects of modern society, such as healthcare, autonomous driving, human-computer interaction, and education. As shown in Fig. 1, a FER system predicts an expression using the facial features. Emotions are represented in different ways, such as Facial Action Coding System (FACS) [1], dimensional affect (*e.g.,* valence and arousal) [2], and categorical facial expressions (*e.g., neutral, happy, sad, surprise, anger, fear,* or *disgust*). Annotating visual channels (video or images) with FACS or dimensional effect is a very rigorous and tedious task and requires specially trained professionals. Consequently, categorical FER models are popular in the field due to their simple interpretation, relative ease of data collection, and wide applicability. Ekman and Friesen [3] argue that the implied meaning of an expression can be labeled with six basic expression categories that are universally the same across different cultures: happy, sad, surprise, anger, fear, and disgust. A neutral expression is subsequently added to this set of prototypical expressions to comprise seven expressions commonly explored in FER research. A FER system that assigns one of these expressions to a facial image uses a categorical model. In this dissertation, we focus on categorical FER systems.

Because of its numerous uses in artificial intelligence, such as human-computer cooperation, data-driven animation, and human-robot communication, detecting emotion from facial expressions will become a pressing requirement. This will also have a wide range of uses, including lie detectors, robotics, and art. Advances in deep learning over recent decades have led to a growing interest in the development of deep learning-based approaches to FER. With modern hardware and storage abundance in recent years, there has been an explosion of research in computer vision tasks using Deep Neural Networks (DNNs). Collecting data is easier than before, and more massive datasets are available for researchers. The computer science community is now able to write complex algorithms to look at the data, analyze the data, and identify patterns. This achievement is possible with powerful Graphical Processing Units (GPUs) with thousands of parallelized computing cores. The current approaches primarily focus on facial investigation keeping background intact and hence building up a lot of unnecessary and misleading features that confuse CNN training process.

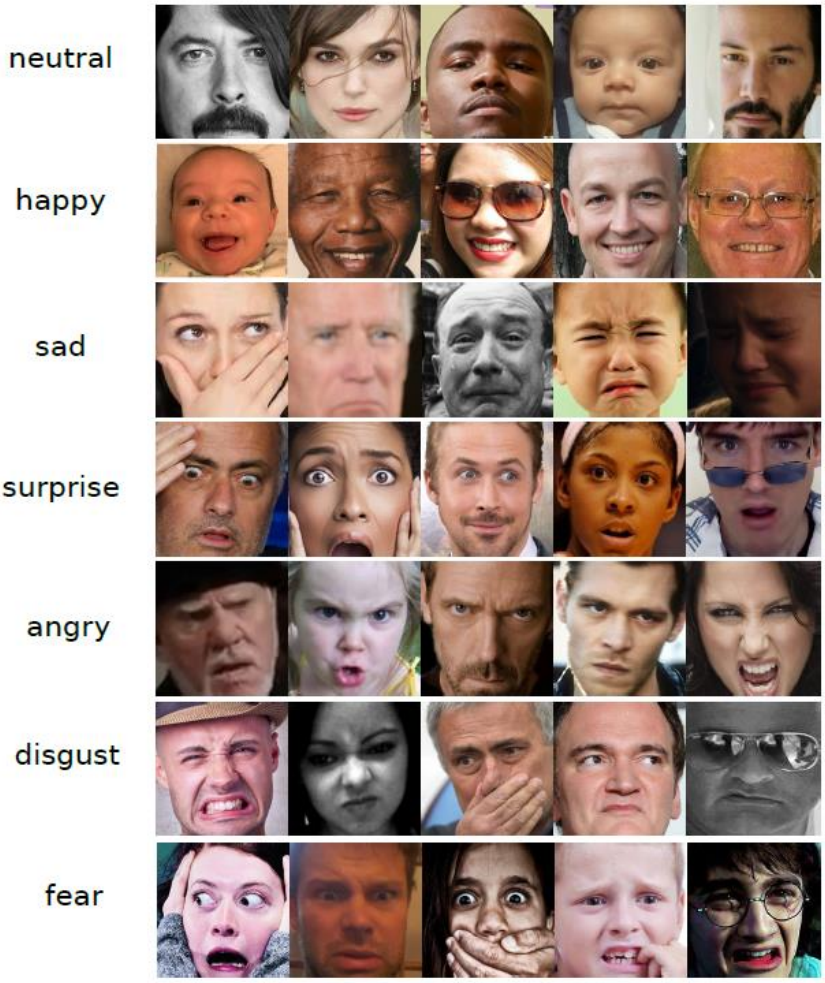


Fig. 1 Example images from a facial expression category

II. LITERATURE REVIEW

A significant amount of research has been conducted to boost FER performance. In this section, we review previous work in the eld from two perspectives: 1. FER using discriminative loss functions, and 2. FER in the wild. Meng et al. developed an Identity-Aware Convolutional Neural Network (IACNN) that simultaneously utilizes both expression-related and identity-related deep features. During the training process, an input image pair is forward-propagated through two identical CNNs with shared parameters to jointly calculate the expression-related and identity-related deep features for both images in the input. The softmax loss function is applied on top of the expression-related deep features to calculate the classi_cation error and optimize the network for learning expression-related deep features. Guo et al. introduce Deep Neural Networks with Relativity Learning (DNNRL) based on the triplet loss to pull the samples with the same expression towards each other and push those with different expressions away from each other in the embedding space. During training, triplets are mined from the dataset including a positive sample, a negative sample, and an anchor. The positive sample shares the same expression with the anchor, and the negative sample has a different expression than the anchor and the positive sample.

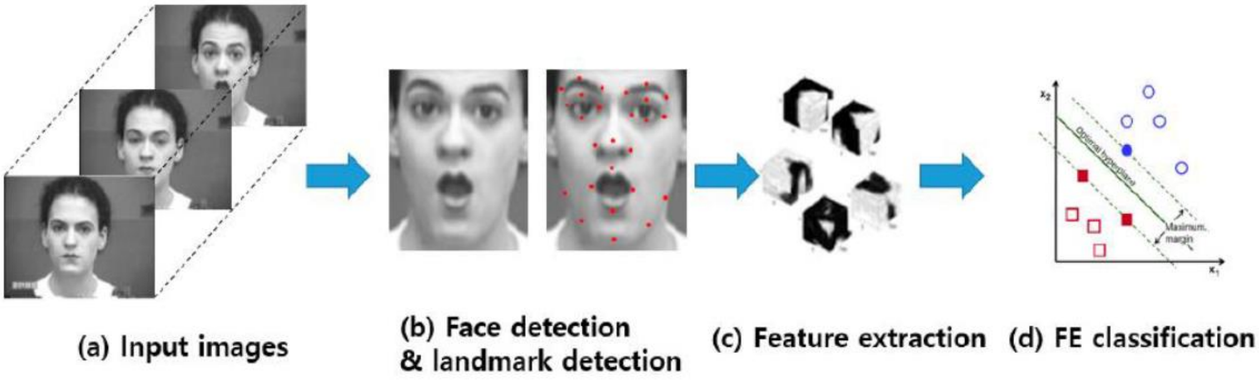
Liu et al. propose (N+M)-tuplet clusters loss function adapted from (N+1)-tuplet loss [4] and Coupled Clusters Loss (CCL) to address the difficulty of anchor selection in triplet loss. Inputs are mined as a set of N positive samples and a set of M negative samples. During training, (N+M)-tuplet clusters loss function forces the samples in the negative set to move away from the center of positive samples and simultaneously clusters the positive samples around their corresponding center to achieve compactness. (N+M)-tuplet clusters loss is evaluated on CK+, MMI, and SFEW.

Cai et al. improve on center loss by adding an extra objective function to achieve intra-class compactness and inter-class separation simultaneously. The modified center loss called Island loss is defined as the summation of the center loss and the pairwise cosine distance between the class centers in the embedding space. During training, the cosine distance is maximized to separate the centers learned by center loss angularly. Island loss is evaluated on CK+, MMI, and Oulu-CASIA[5].

Li et al. introduce separate loss to address large intra-class variation and inter-class similarity. Separate loss is a cosine version of center loss and island loss. It consists of two parts: 1. Intra-class loss, and 2. Inter-class loss. The intra-class loss is the normalized cosine similarity between a sample's deep feature representation, and the interclass loss is the normalized cosine similarity between the centers in the embedding space. During training, the intra-class loss is minimized, and the inter-class loss is maximized [6]. Since both loss functions are based on the normalized cosine similarity metric, they are considered to be commensurate.

III. OVERVIEW OF EXISTING METHODOLOGY

Existing facial expression recognition system typically has four steps. The first is to detect a face in an image and draw a rectangle around it and the next step is to detect landmarks in this face region. The third step is extracting spatial and temporal features from the facial components. The final step is to use a Feature Extraction (FE) classifier and produce the recognition results using the extracted features [7]. Fig.2 shows the FER procedure for an input image where a face region and facial landmarks are detected. Facial landmarks are visually salient points such as the end of a nose, and the ends of eyebrows and the mouth as shown in Fig. 3. The pairwise positions of two landmark points or the local texture of a landmark are used as features. Table 1 gives the definitions of 64 primary and secondary landmarks [8]. The spatial and temporal features are extracted from the face and the expression is determined based on one of the facial categories using pattern classifiers.



(a) Input images (b) Face detection & landmark detection (c) Feature extraction (d) FE classification

Fig. 2 FER procedure for an image

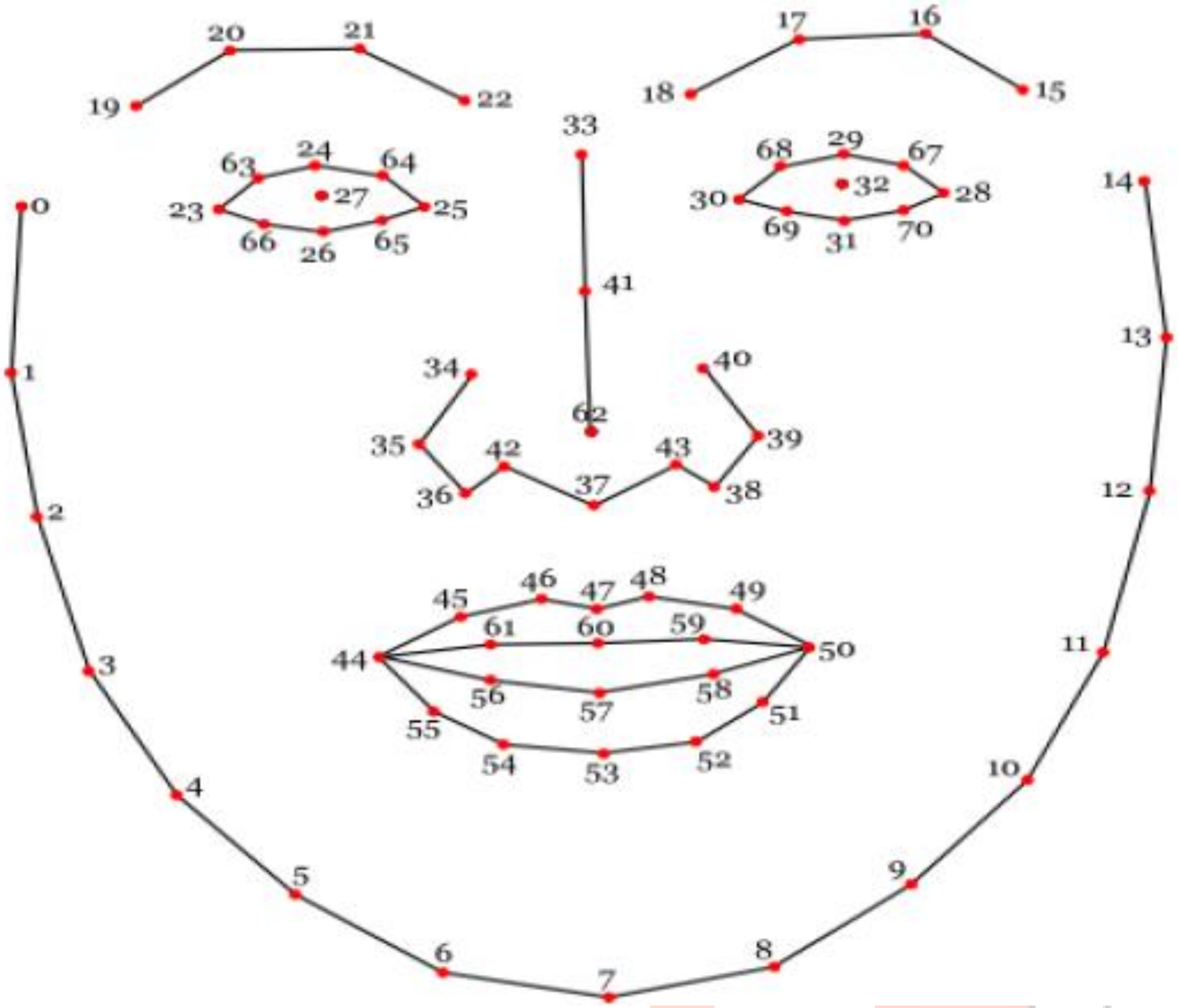


Fig. 3 Facial landmarks to be extracted from a face.

Table 1 Definitions of 64 primary and secondary landmarks.

Primary landmarks		Secondary landmarks	
Number	Definition	Number	Definition
16	Left eyebrow outer corner	1	Left temple
19	Left eyebrow inner corner	8	Chin tip
22	Right eyebrow inner corner	2-7,9-14	Cheek contours
25	Right eyebrow outer corner	15	Right temple
28	Left eye outer corner	16-19	Left eyebrow contours
30	Left eye inner corner	22-25	Right eyebrow corners
32	Right eye inner corner	29,33	Upper eyelid centers
34	Right eye outer corner	31,35	Lower eyelid centers
41	Nose tip	36,37	Nose saddles
46	Left mouth corner	40,42	Nose peaks (nostrils)
52	Right mouth corner	38-40,42-45	Nose contours
63,64	Eye centers	47-51,53-62	Mouth contours

Deep learning based facial expression recognition approaches greatly reduce the dependence on face-physics based models and other preprocessing techniques by enabling end to end learning directly from the input images [9]. Among DL models, Convolutional Neural Networks (CNNs) are the most popular. With a CNN, an input image is filtered through convolution layers to produce a feature map. This map is then input to fully connected layers, and the facial expression is recognized as belonging to a class based on the output of the FE classifier.

IV. PROPOSED METHODOLOGY

The proposed method is based on a two-level CNN framework. The first level recommended is background removal 10], used to extract emotions from an image, as shown in Fig. 4. Here, the conventional CNN network module is used to extract primary expressional vector (EV). The expressional vector (EV) is generated by tracking down relevant facial points of importance. EV is directly related to changes in expression. The EV is obtained using a basic perceptronunit applied on a background-removed face image. In the proposed FERC model, we also have a non-convolutional perceptron layer as the last stage. Each of the convolutional layers receives the input data (or image), transforms it, and then outputs it to the next level. This transformation is convolution operation. All the convolutional layers used are capable of pattern detection. Within each convolutional layer, four filters were used. The input image fed to the first-part CNN (used for background removal) generally consists of shapes, edges, textures, and objects along with the face. The edge detector, circle detector, and corner detector filters are used at the start of the convolutional layer 1. Once the face has been detected, the second-part CNN filter catches facial features, such as eyes, ears, lips, nose, and cheeks. The edge detection filters used in this layer are shown in Fig. 5. The second-part CNN consists of layers with 3 × 3 kernel matrix, e.g., [0.25, 0.17, 0.9; 0.89, 0.36, 0.63; 0.7, 0.24, 0.82]. These numbers are selected between 0 and 1 initially. These numbers are optimized for EV detection, based on the ground truth we had, in the supervisory training dataset. Here, we used minimum error decoding to optimize filter values. Once the filter is tuned by supervisory learning, it is then applied to the background-removed face (i.e., on the output image of the first-part CNN), for detection of different facial parts (e.g., eye, lips. nose, ears, etc.) To generate the EV matrix, in all 24 various facial features are extracted. The EV feature vector is nothing but values of normalized Euclidian distance between each face part. Fig. 6 shows convolution operation on an input image. Fig. 7 Illustrates the general pipeline for FER using a CNN model.

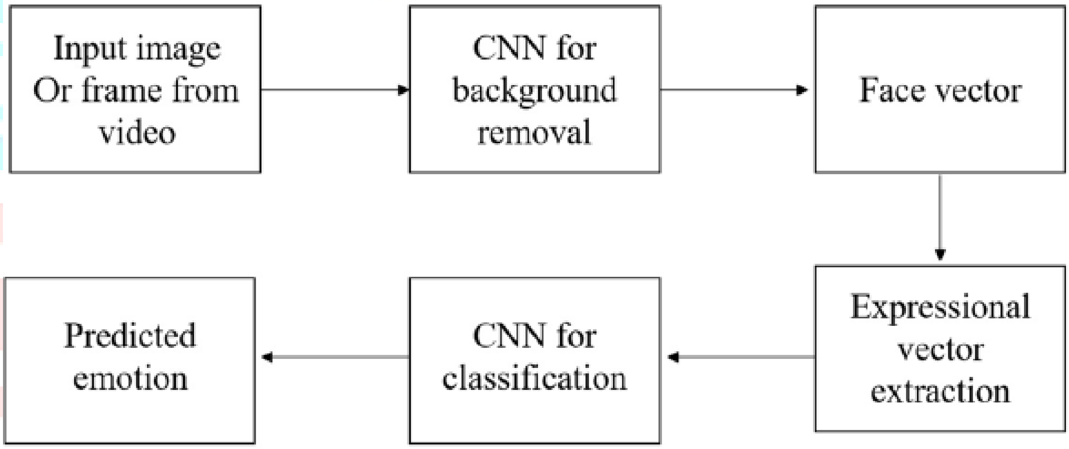
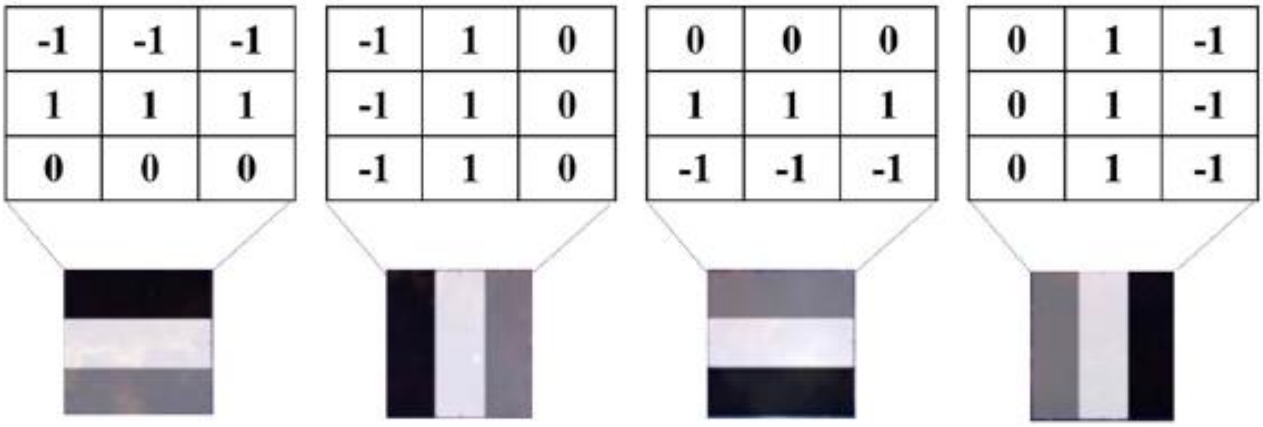

Input image Or frame from video
CNN for background removal
Face vector
Expressional vector extraction
CNN for classification
Predicted emotion

Fig. 4 Facial Expression System using CNN

Once the input image is obtained, skin tone detection algorithm is applied to extract human body parts from the image. This skin tone-detected output image is a binary image and used as the feature, for the first layer of background removal CNN (also referred to as the first-part CNN in this manuscript). This skin tone detection depends on the type of input image. If the image is the colored image, then YCbCr color threshold can be used.



-1	-1	-1
1	1	1
0	0	0

-1	1	0
-1	1	0
-1	1	0

0	0	0
1	1	1
-1	-1	-1

0	1	-1
0	1	-1
0	1	-1

Fig. 5 Vertical and horizontal edge detector filter matrix used at layer 1 of background removal CNN

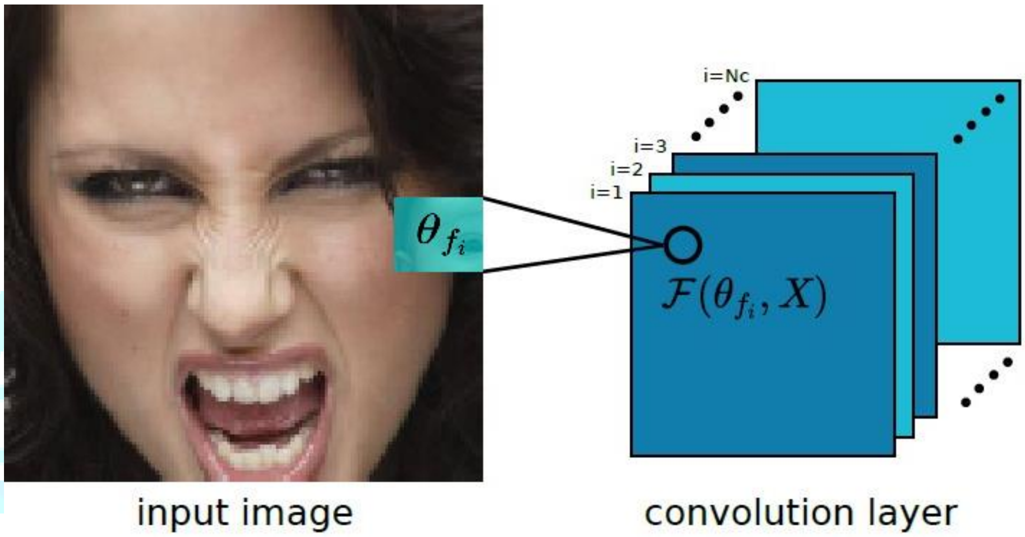


Fig. 6 Convolution operation on an input image.

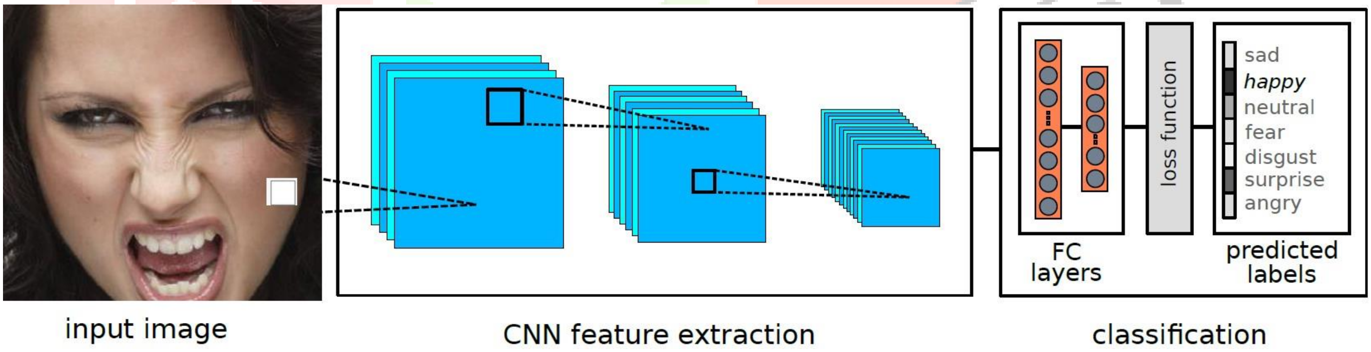


Fig. 7 Illustration of the general pipeline for FER using a CNN model

V. IMPLEMENTATION

ResNet-18, a standard deep residual network from the family of ResNets used in this implementation due to its close to state-of-the-art performance on canonical visual recognition tasks while offering fewer parameters. Hence, the proposed models are trained faster compared to deeper networks. The CNN layer details are shown in Table 2.

Table 2: Convolutional network layer details.

layer name	output size	layer detail
conv1	112×112	$conv\{7 \times 7, 64, 2\}$
maxpool1	56×56	$3 \times 3, stride = 2$
conv2	56×56	$\begin{bmatrix} conv\{3 \times 3, 64\} \\ conv\{3 \times 3, 64\} \end{bmatrix} \times 2$
conv3	28×28	$\begin{bmatrix} conv\{3 \times 3, 128\} \\ conv\{3 \times 3, 128\} \end{bmatrix} \times 2$
conv4	14×14	$\begin{bmatrix} conv\{3 \times 3, 256\} \\ conv\{3 \times 3, 256\} \end{bmatrix} \times 2$
conv5	7×7	$\begin{bmatrix} conv\{3 \times 3, 512\} \\ conv\{3 \times 3, 512\} \end{bmatrix} \times 2$
pooling layer	1×1	average pool, K-neuron fully-connected layer

5.1 Dataset

A facial expression database is a collection of images or video clips with facial expressions of a range of emotions. MMI Face Database has been developed to address most (if not all) of the issues mentioned above. It contains more than 1500 samples of both static images and image sequences of faces in frontal and in profile view displaying various facial expressions of emotion, single AU activation, and multiple AU activation. It has been developed as a web-based direct manipulation application, allowing easy access and easy search of the available images 116. Examples of static frontal-view images of facial expressions in the MMI Facial Expression Database are shown in Fig.7.

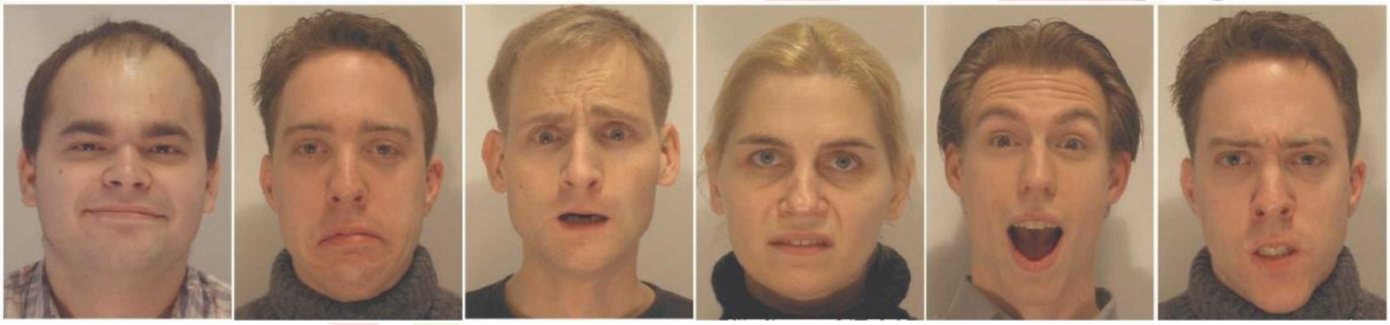


Fig 7: Examples of static frontal-view images of facial expressions in the MMI Facial Expression Database

Table 3 Facial expression database image categories.

Category	Train database	Test database
Angry	14040	27
Disgust	12220	24
Fearful	7800	15
Happy	14300	28
Sad	8970	15
Surprised	8580	16
Scorn	5460	12
Neutral	8970	15

5.2 Tools used

5.2.1 Python:

Python was the language of selection for this project. This was a straightforward call for many reasons. Python as a language has a vast community behind it. Any problems which may be faced is simply resolved with a visit to Stack Overflow. Python is among the foremost standard language on the positioning that makes it very likely there will be straight answer to any question. Python has an abundance of powerful tools prepared for scientific computing Packages like NumPy, Pandas and SciPy area unit freely available and well documented. Packages like these will dramatically scale back, and change the code required to write a given program.This makes iteration fast. 3. Python as a language is forgiving and permits for program that appear as if pseudo code. This can be helpful once pseudo code given in tutorial papers must be enforced and tested. Using python this step is sometimes fairly trivial. However, Python is not without its errors. The language is dynamically written and packages are area unit infamous for Duck writing. This may be frustrating once a package technique returns one thing that, for instance, looks like an array instead of being an actual array. Plus the actual fact that standard Python documentation does not clearly state the return type of a method, can lead to a lot of trials and error testing that will not otherwise happen in a powerfully written language.

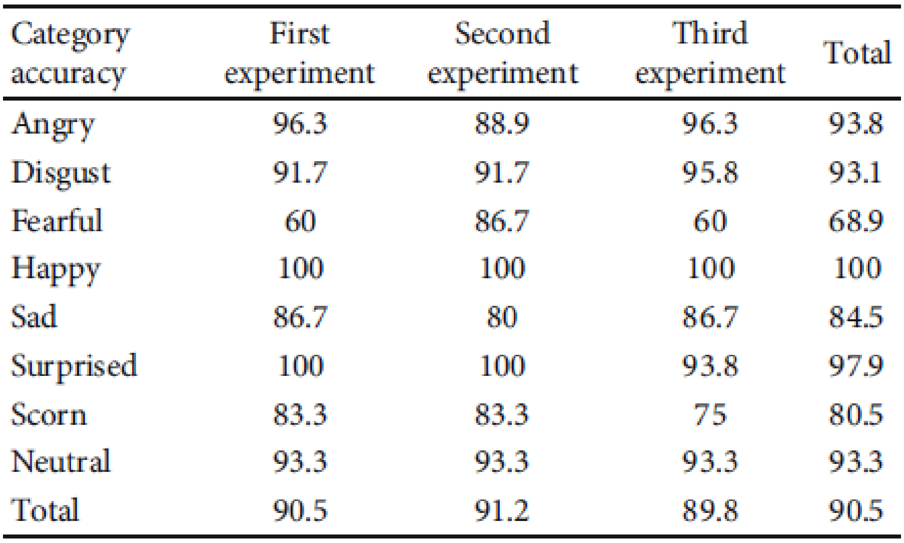
5.2.2 Jupiter Notebook:

The Jupyter Notebook is an open-source web application that enables you to make and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more. Noise Removal and Sharpening: Unwanted data of element are remove using filter and image Can be sharpen and black and white gray scale image is used as a input. Erosion and Dilation: It is applied to binary image, but there are many versions so that can be work on grayscale images. The basic effect of the operator on a binary image is eroding away to the boundaries of regions for ground pixels. Negation: A negative is an image, usually it used on a strip or sheet of transparent plastic film, in negation the lightest areas of the photographed subject appear darkest and the darkest areas appear lightest.

VI. RESULTS AND DISCUSSION

To find the optimal situation of the facial expression recognition system, we 10 fold validation in the CNN model. The average recognition accuracy and speed obtained as a result are shown in Table 3. The experiment is the average of three experiments.

Table 4: The recognition accuracy of the improved CNN model.



Category accuracy	First experiment	Second experiment	Third experiment	Total
Angry	96.3	88.9	96.3	93.8
Disgust	91.7	91.7	95.8	93.1
Fearful	60	86.7	60	68.9
Happy	100	100	100	100
Sad	86.7	80	86.7	84.5
Surprised	100	100	93.8	97.9
Scorn	83.3	83.3	75	80.5
Neutral	93.3	93.3	93.3	93.3
Total	90.5	91.2	89.8	90.5

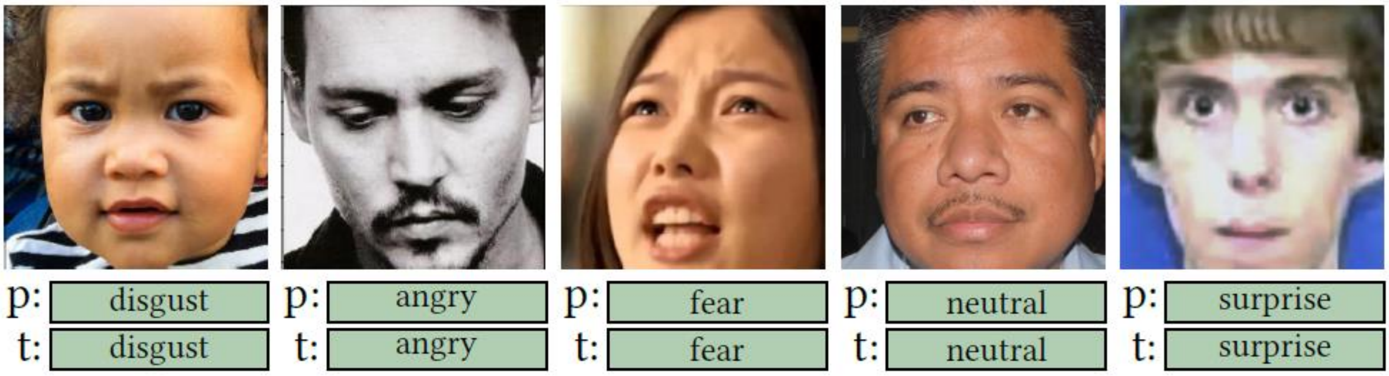


Fig. 8: Predicted (p) and true label (t) images

The expression recognition performance of the proposed methodology is shown in Table 5.

Table 5: Comparison of recognition accuracy and time

Model	LeNet-5	Without decision level
Accuracy (%)	74.6	87.9
Time (s)	0.59	0.31

VII . CONCLUSION

Aiming at the expression recognition of low-pixel face images, the article proposes an improved CNN expression recognition method. The article increases the nonlinearity of the network model by adding a convolutional layer. The result shows that LeNet-5 has the accuracy of 74.6% and without decisional level the accuracy is 87.9%.

REFERENCES

[1] X. Zhou, “Video expression recognition method based on spatiotemporal recurrent neural network and feature fusion,” Journal of Information Processing Systems, vol. 17, no. 2, pp. 337–351, 2021.

[2] X. Liu, Y. Xia, H. Yu, J. Dong, M. Jian, and T. D. Pham, “Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 10, pp. 2325–2332, 2020.

[3] J. K. Park and D. J. Kang, “Unified convolutional neural network for direct facial keypoints detection,” The Visual Computer, vol. 35, no. 11, pp. 1615–1626, 2019.

[4] A. Satapathy and L. J. Livingston, “A lite convolutional neural network built on permuted Xceptio-inception and Xceptioreduction modules for texture based facial liveness recognition,” Multimedia Tools and Applications, vol. 80, no. 7, pp. 10441–10472, 202.

[5] N. Mehendale, “Facial emotion recognition using convolutional neural networks (FERC),” SN Applied Sciences, vol. 2, no. 3, pp. 1–8, 2020.

[6] K. S. Yoon and J. Y. Choi, “Compressed ensemble of deep convolutional neural networks with global and local facial features for improved face recognition,” Journal of Korea Multimedia Society, vol. 23, no. 8, pp. 1019–1029, 2020.

[7] H. Liao, G. Wen, Y. Hu, and C. Wang, “Convolutional herbal prescription building method from multi-scale facial features,” Multimedia Tools and Applications, vol. 78, no. 24, pp. 35665–35688, 2019.

[8] H. Adachi, K. Oiwa, and A. Nozawa, “Drowsiness level modelling based on facial skin temperature distribution using a convolutional neural network,” IEEJ Transactions on Electrical and Electronic Engineering, vol. 14, no. 6, pp. 870–876, 2019.

[9] F. Kong, “Facial expression recognition method based on deep convolutional neural network combined with improved LBP features,” Personal and Ubiquitous Computing, vol. 23, no. 3- 4, pp. 531–539, 2019.,

[10] G. Yolcu, I. Oztel, S. Kazan et al., “Facial expression recognition for monitoring neurological disorders based on convolutional neural network,” Multimedia Tools and Applications, vol. 78, no. 22, pp. 31581–31603, 2019.

[11] M. Z. Lifkooee, Ö. M. Soysal, and K. Sekeroglu, “Video mining for facial action unit classification using statistical spatial–temporal feature image and LoG deep convolutional neural network,” Machine Vision and Applications, vol. 30, no. 1, pp. 41–57, 2019.