Proceedings of **"National Conference on Recent Trends in Science and Advances in Engineering"**
Organized by Fabtech Technical Campus, College of Engineering & Research, Sangola
International Journal of Innovations in Engineering Research and Technology [IJIERT]
ISSN: 2394-3696, Website: www.ijiert.org, June, 2022

FACIAL EMOTION RECOGNITION USING DEEP CONVOLUTIONAL NETWORKS

Ingale Komal Dhananjay1
Department of Electronic and Telecommunication Engineering
Sveri's college of engineering Pandharpur, Maharashtra, India
Email: komalingale555@gmail.com

Dr.Vibhute A.S2
Department of Electronic and Telecommunication Engineering
Sveri's college of engineering Pandharpur, Maharashtra, India
Email: hossein.khaliliardali@semnan.ac.ir

Mrs. Pol Komal Amit3,Umesh Bhodale
Department of Electronic and Telecommunication Engineering
Fabtech Technical Campus Sangola, Maharashtra, India
Email: komalingale555@gmail.com

***Abstract*—Facial emotion recognition is a developing area that is used in a variety of modern applications such as social robots, neuromarketing, and gaming. Nonverbal communication methods like as facial expressions, eye movement, and gestures are widely employed in various applications of human computer interaction, with facial emotion being the most commonly used since it conveys people's emotional states and sentiments. Emotion recognition is a difficult endeavour since there is no landmark demarcation between the emotions on the face, and there is also a great deal of complexity and unpredictability. Traditional machine learning algorithms employ certain crucial extracted characteristics for modelling the face, therefore it cannot reach high accuracy rates for emotion identification since the features are hand-engineered and rely on existing knowledge. In this work, convolutional neural networks (CNN) were constructed to recognise facial emotion expressions and classify them into seven fundamental categories. CNN calculates features by learning automatically rather than by hand-engineering them. The suggested approach is unique in that it uses facial action units (AUs) of the face, which are first recognised by C NN and then used to recognise the seven fundamental emotion states. The Cohn-Kanade database is used to analyse the proposed model, and the model obtains the best accuracy rate of 97.01 by adding AU, whereas other efforts in the literature employ a direct CNN and reach an accuracy rate of 95.75.**

Keywords—Emotion ,Classification, C onvolutional n eural networks, Action Unit.

I. INTRODUCTION

The study of facial expressions may be traced back to Darwin's studies on the evolution of species, where it first appeared as a kind of nonverbal communication. In his research, he classified facial behaviour into numerous categories. This mode of communication is faster than verbal communication and has provided more benefits to the human species than others. Facial emotion is defined as a person's mental state, goals, and emotional response to an external input [1].

Emotions are a fundamental human characteristic that is required for efficient social interactions. Human communication can be either verbal or nonverbal, with nonverbal communication being the more common [1]. Emotion plays an important part in nonverbal communication since it expresses humans' feelings about the issue, and psychological research has shown that facial emotions are more effective than spoken words in conversation[2].

Emotion recognition is a crossroads of computer science, cognitive science, and psychology, and it may be accomplished by a variety of ways such as body language, speech intonation, and electroencephalography (EEG) [3]. Recognizing emotion from facial expression is the simpler and more practical method. In the interaction context, then, facial emotion recognition is more practical than emotion detection from EEG input since EEG is better suited for clinical applications such as neurofeedback where the participants remain fixed.

In the last two decades, the field of human-computer interactions (HCI) has advanced and played an essential role in the development of computer science by generating a wide range of practical applications that integrate human behaviour with computer equipment [4]. In robotics research, particularly with humanoid robots, it is fascinating to put emotion recognition to the machine to allow for natural communication [5]. There is also interest in using emotion in HCI to carry out an efficient and intelligent interaction or communication between humans and machines that act like humans. Information from facial expressions is dispersed in several areas of the face, and each of them has varied information, such that the lips and eyes have more information than the cheeks and forehead. Several psychology research have revealed that culture and environment may alter the impact of emotion and the way people express themselves. Many of these research have indicated that gender, cultural background, and age have biases in expressing emotion, but there is no strong evidence on the relevance of environment for emotional tendencies [6].

Methods for recognising emotions are classified into two categories: The first group works on static photos, while the second group works on dynamic image sequences. Static techniques ignore temporal information and just employ current picture information, whereas dynamic approaches combine images and temporal information to detect expressed emotion in frame sequences.

Face picture acquisition, feature extraction, and facial emotion expression identification are the three processes in automatic emotion expression recognition. Within-class variations of expression should be minimised in the ideal retrieved features, while between-class variations should be maximised. If the extracted features are unsuitable for the job at hand and lack sufficient information, even the finest classifier may fail to achieve optimal performance.

There are two ways to feature extraction for emotion recognition: Approaches based on geometric features and methods based on appearance [7]. The first approach considers the placement and form of components of the face such as the eyes, mouth, brows, and nose, whereas the second method considers specific areas or the entire face.

Because separating expressions' feature space is a complex challenge, computer expression recognition remains a difficult undertaking. Some issues may arise as a result of this; for example, extracted features from two faces with equal expressions may differ, yet extracted features from one face with two expressions may be equal, or some expressions, such as "fear" and "sad," may be extremely similar.

The suggested method's key contribution is to model emotion expression on static photos in order to recognise seven (happy, surprise, angry, disgust, fear, sad, and neutral) face emotional states. To that purpose, photos were first altered by pre-processing techniques to have the exact part of the face, and then a deep CNN was trained to categorise the emotions. The suggested approach is unique in that it uses AUs [8] such as "Lip stretcher" to recognise emotions. Because there are semantic links between distinct AUs, the performance of emotion detection would improve when AUs were generated using CNN's automatically learned features.

II. Related Work

Recent techniques to facial emotion recognition that have achieved a high degree of accuracy will be explored in this part. Many emotion identification systems are based on hand-engineered features such as histograms of oriented gradients (HOG) features, scale invariant feature transform (SIFT) descriptors, Gabor filters, or local binary patterns (LBP).

A variety of algorithms for emotion recognition categorise the facial picture into fundamental emotions such as happiness, rage, and sorrow [9] [10] [11], while others attempt to identify AUs on the face in order to characterise objective facial expression characteristics [12].

[13] proposed a novel boosted deep belief network (BDBN) in order to performing three steps of feature learning, feature selection, and classifier construction iteratively. The authors claim that their proposed method based on BDBN is ableto learn a set of features which can help to characterize facial appearance for emotion expression. Experiments were conducted on the CK+ dataset and JAFFE dataset, and tested the six basic emotion. The proposed approach yields an accuracy rate of 97.70% in the CK+ database. [9] proposed a novel approach based on transformations of given image intensity into a 3D spaces, in order to be invariant to monotonic transformations. Their experiments were performed on two databases of static images static facial expression recognition sub-challenge (SFEW) and emotion recognition in the wild challenge (EmotiW 2015), with a noteworthy 40% improvement in performance. [10] created an emotion identification system known as extreme sparse learning (ESL) that can learn a dictionary of basis functions and non-linear models. The proposed network uses an extreme learning machine (ELM) and sparse representation to improve classification accuracy in noisy and poor pictures. The studies were carried out on the CK+ dataset and yielded an accuracy of 95.33 percent. [14] investigated emotion identification by examining facial features such as the lips and eyes and used principal component analysis (PCA) and neural networks. The experiments were carried out on the JAFFE and FEEDTUM datasets. [15] suggested a technique for emotion expression categorization based on transfer learning of existing convolutional neural networks with fully connected layers and pretrained layers. Experiments on the CK+ and JAFFE datasets yielded training accuracy of 90.7 percent and test accuracy of 57.1 percent. [11] suggested a method for recognising facial expressions that makes use of picture pre-processing methods and CNN. They retrieved certain aspects of emotions on the face using picture pre-processing algorithms. The studies were carried out on three public datasets, CK+, JAFFE, and BU-3DFE, with a success rate of 96.76 percent on the CK+ dataset. [16] presented a CNN-based technique that is independent of any hand-engineered characteristics. Their network structure is divided into four sections, the first of which is for image pre-processing and the others for feature extraction. Following feature extraction, a fully linked layer in CNN classifies seven expressions. Their suggested structure has 15 layers and obtained 99.6 percent and 98.63 percent accuracy on the CK+ and NMI datasets, respectively.

RECOGNITION OF EMOTION IN DEEP NETWORK

In the current study, a deep convolutional network recognises seven states of facial expression while concurrently performing three phases of feature learning, selection, and classification. Training neural networks with more than two layers was a difficult task in the previous decade, but with the advancement of GPUs, it is now possible to train neural networks with more than one layer. Deep neural networks have three alternating types of layers: convolutional, subsampling, and fully connected.

Convolutional Neural Network

CNN has been proved to be particularly successful for learning features and modelling high levels of abstraction since 1998 [17]. CNN is made up of six layers: convolutional, subsampling, rectified linear unit (ReLU), fully connected, output, and softmax [18].

Convolutional layer: Convolutional layers are determinedby number of generated maps and kernel's size. The kernelis moved over the valid area of the given image (perform a convolution) for generating the map. If f_k be a filter with a kernel size $n \times m$ and is supposed to applied into the given image x, output of the layers can be calculates as follows:

$$C(x_{u,v}) = \sum_{i=-\frac{n}{2}}^{\frac{n}{2}} \sum_{i=-\frac{m}{2}}^{\frac{m}{2}} f_k(i, j) x_{u-i,v-j} \qquad (1)$$

Where each CNN neuron has $n \times m$ number of input connections.

Sub-sampling layers: Sub-sampling layers in CNN lower the map size of the preceding layer in order to improve kernel invariance. Sub-sampling is classified into two types: average pooling and maximal pooling [18]. The maximum function in the Max-Pooling reduces the input value at xi. If m is the kernel size, the output of max-pooling may be determined as follows:

$$M(x_i) = Max\{x_{i+k,i+l} \mid |k| \leq \frac{m}{2}, |l| \leq \frac{m}{2} k, k \in \mathbb{N}\} \quad (2)$$

Rectified linear unit: A rectified linear unit is a activation function which it simply thresholded at zero and can be calculated as follows:

$$R(x) = max(0, x) \quad (3)$$

ReLU has an advantage over the tanh/sigmoid function in that it may be implemented using simple thresholding at zero, whereas the tanh/sigmoid function requires costly calculations such as exponentials. ReLU also prevents gradient error and accelerates stochastic gradient descent convergence as compared to tanh/sigmoid functions.

Fully connected layer: Fully connected layers are identical to neurons in generic neural networks in that all neurons in the preceding layer are fully linked. If x is given a size of k and the number of neurons in the fully connected layer is given a value of l, the layer may be determined as follows:

$$F(x) = \sigma(W * x) \quad (4)$$

Where σ is activation function.

Output Layer: The output layer represents the class of the input picture, and its size is proportional to the number of classes. Output vector x yields the following class:

$$C(x) = \{i \mid \exists i \; \forall j \quad i : x_j \leq x_i\} \quad (5)$$

Softmax layer: The error of the network is propagated back through a softmax layer. If N be the size of the input vector, a mapping can be calculates by softmax such that: $S(x) : \mathbb{R} \rightarrow [0, 1]^N$, and each components of the softmax layer is calculated as follows:

$$S(x)_j = \frac{x^{x_i}}{\sum_{i=0}^{N} e^{x_i}} \quad (6)$$

Where $1 \leq j \leq N$.

CNN learning works by determining the optimal synapses' weights of neurons. Unlike typical neural networks, which use created features as input, CNN uses raw pictures as input [19]. During the training phase, the network is fed training data, which includes grayscale pictures with labels.,

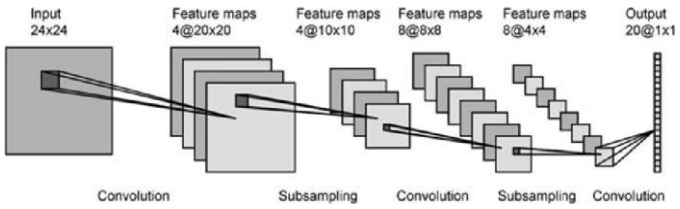

Fig. 1: An example of CNN architecture [20].

TABLE I: Seven basic emotions based on AU codes.

Emotion	Action Unites
Anger	4+5+7+23
Contempt	R12A+R14A
Disgust	9+15+16
Fear	1+2+4+5+7+20+26
Happiness	6+12
Sadness	1+4+15
Surprise	1+2+5B+26

In order to identify the best set of weights, a collection of photos is segregated for validation. During the test phase, the network receives a grayscale picture, and the network's output is the anticipated class of the provided image. Figure 1 is a CNN architectural sample. As input, the network gets a grayscale picture 24 24 and produces a label for each class. The target in the picture is the class with the greatest value. The CNN design consists of two convolutional layers, two pooling (subsampling) layers, and one fully connected layer, all of which generate the softmax loss function and scores.

A. Proposed Architecture

In the proposed method, a psychological framework called Facial Action Coding System (FACS) [8] is used to increase the accuracy of the recognising system, and instead of giving an image to the system and classifying seven emotions, it has been used coding of facial movements by AUs for classifier output, and determining final emotion state by combination of AUs. FACS can describe emotion expression by their appearanceon the face using AU which is based on anatomical movement of face muscles [8]. AU for facial expression includes 46
atomic component of facial movement and each emotions consist some AUs. FACS can be used in many approached for measuring and describing facial behaviors, and is based on actions of observable face muscles from the anatomical aspect or AUs. It also can estimate intensity of expressed emotion which can bed used in study of complex facial behavior.

There are several codes for AUs which each of them is based on facial muscles, and can be used for study of facial emotion expressions. A list of AUs and action descriptors can be found in [21], that by using them seven basic emotion can be created. Table I shown seven basic emotion which based on combination of some AUs.

To recognise each AU from face expression, customised high precision features are required for optimum performance. To that goal, a CNN is proposed as a detection method.

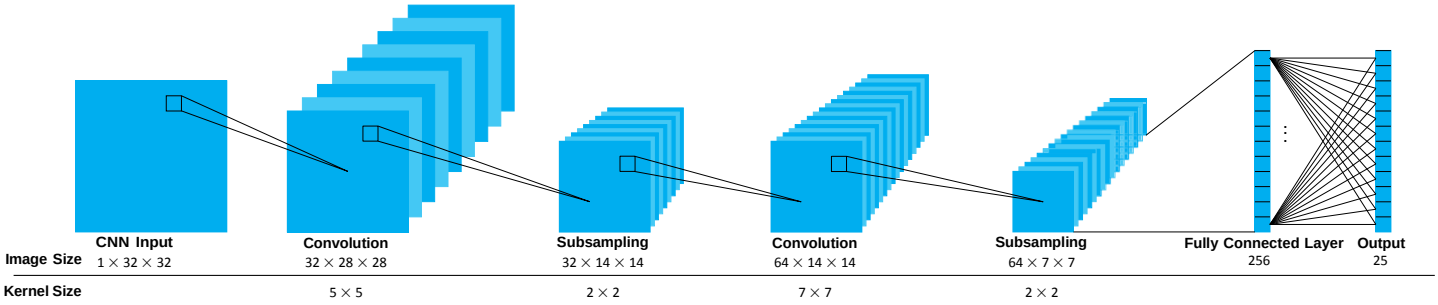

Fig. 2: Architecture of proposed CNN for detecting AUs.

AUs on the face without any hand involvement, and after detecting 25 AUs from given image, by using activation of a set of AUs the expressed emotion can be recognized.
The suggested CNN architecture for identifying AUs is divided into four components, as illustrated in fig2. The network inputs are grayscale pictures of size 32 32, and the network consists of two convolutional layers with 32 and 64 filters, with filter sizes of 5 5 and 7 7. A Rectified Linear Unit (ReLU) activation function is applied after each convolutional layer. Following convolutional layers, max pooling layers and a fully connected layer with 256 hidden units are added. Finally, in order to classify AUs, a softmax layer with 25 neurons that indicate the number of AUs and is fully coupled to the preceding layer is examine The synaptic weights between neurons are determined using the stochastic gradient descent method.d. The first layer extracts fundamental visual elements such as corners, aligned edges, and the contour of the mouth, eyes, and brow. The synaptic weights among the neurons are calculated by stochastic gradient descent approach. Presence or absence of each AUs denoted by scored in the network output so that low value indicate absence of AUs, while a high value indicated presence of AUs.

III. Experiments, Results and Evaluation

The proposed system was trained and tested on Extended Cohn-Kanade (CK+) dataset [22] which is a publicly available and used for facial expression studies. The dataset includes 123 subjects which they were asked to perform a series of emotion expressions, and the images were captured from front of the subjects. Dataset images are grayscale with size 640 by 480 and 8-bit precision, and for each image there is a descriptor file which contain labels for denoting AUs that presented in each images. The dataset includes images for the expressions: neutral, happy, sad, surprise, fear, anger, disgust and contempt. In order to perform a fair comparison with other proposed methods in the literature, contempt expression is omitted. Figure 3 shown some examples of images in the CK+ dataset.

In order to improve classification accuracy, certain pre-processing techniques were performed to the photos in order to retrieve the precise expressed emotion on the face. Image cropping, rotation correction, downsampling, spatial normalisation, and intensity normalisation are among the pre-processing techniques. On the Ubuntu 16.04 operating system, the pre-processing procedures were written in C++ and OpenCV, while all additional experiments were implemented in the NVIDIA CUDA framework. For this study, an Intel Core i5 3.3 Ghz processor and an NVIDA GeForce GTX 730 with 4GB RAM and 1152 CUDA Cores were used.

To train the network, the cross-validation approach was applied to all pictures in the dataset with batch sizes of 128 and 250 epochs. Furthermore, a confusion matrix is created to demonstrate the behaviour of various emotion classes. Figure 4 depicts a visual representation of the confusion matrix. As seen in Figure 4, there is a link between the angry label and the neutral label, indicating that there are some photos with real labels furious that the system predicted as neutral. The network outperforms other labels in predicting the happy label, indicating that learning the happy feature is simpler than learning other emotions.

Table II compares average accuracy for all emotion classes for the proposed technique and other state-of-the-art methods in the literature for which they also utilised the CK+ dataset for evaluation.
TABLE II: Comparison of emotion recognition algorithms for seven expressed emotions on the CK+ dataset.

Method	Description	Accuracy
[23]	SVM + Gabor filters + LBP	88.90
[24]	Local Directional Number Pattern (LDN)	89.30
[25]	LBP + SVM	91.40
[11]	Normalization+ DL	95.75
Proposed	Normalization + Action Units + DL	97.01

CONCLUSION

This research introduced an emotion identification system using a unique technique for recognising action units (AUs), which is a psychological framework coding of facial motions. A CNN is created for optimum feature extraction, detection of AUs, and detection of seven stated emotions. The experimental findings shown that deep CNNs can learn face expression features and improve facial emotion identification accuracy.

REFERENCES

[1] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[2] John N Bassili. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of personality and social psychology*, 37(11):2049, 1979.

[3] Mostafa Mohammadpour, S M R Hashemi, and Negin Houshmand. Classification of EEG-based Emotion for BCI Applications. In *The 7th joint Conference on Artificial Intelligence & Robotics and the 9th RoboCup IranOpen International Symposium*, pages 127–131. IEEE, 2017.

Fig. 3: Some example of images in CK+ dataset. Subjects are in emotional states neutral, happy, surprise and disgust [22].

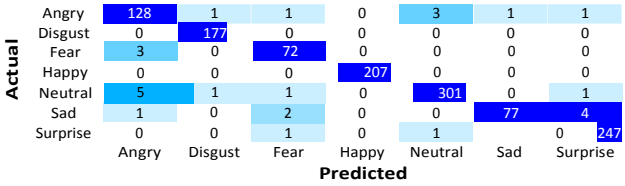

Fig. 4: Confusion matrix of seven expressed emotion on the CK+ database.

[4] N Fragopanagos and John G Taylor. Emotion recognition in human–computer interaction. *Neural Networks*, 18(4):389–405, 2005.

[5] Li Zhang, Ming Jiang, Dewan Farid, and M Alamgir Hossain. Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Systems with Applications*, 40(13):5160–5168, 2013.

[6] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203, 2002.

[7] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2009.

[8] Paul Ekman and Wallace Friesen. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto: Consulting Psychologists*, 1978.

[9] Gil Levi and Tal Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 503–510. ACM, 2015.

[10] Seyedehsamaneh Shojaeilangari, Wei-Yun Yau, Karthik Nandakumar, Jun Li, and Eam Khwang Teoh. Robust representation and recognition of facial emotions using extreme sparse learning. *IEEE Transactions on Image Processing*, 24(7):2140–2152, 2015.

[11] Andre´ Teixeira Lopes, Edilson de Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos. Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. *Pattern Recognition*, 61:610–628, 2017.

[12] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

[13] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[14] Damir Filko and Goran Martinović. Emotion recognition system by a neural network based facial expression analysis. *automatika*, 54(2):263–272, 2013.

[15] Dan Duncan, Gautam Shine, and Chris English. Facial Emotion Recognition in Real Time.

[16] Peter Burkert, Felix Trier, Muhammad Zeshan Afzal, Andreas Dengel, and Marcus Liwicki. Dexpression: Deep convolutional neural network for expression recognition. *arXiv preprint arXiv:1509.05371*, 2015.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] Michael A Nielsen. Neural networks and deep learning, 2015.

[19] A Deshpande. A Beginner's Guide To Understanding Convolutional Neural Networks. *Retrieved March*, 31:2017, 2016.

[20] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, and Gang Wang. Recent advances in convolutional neural networks. *arXiv preprint arXiv:1512.07108*, 2015.

[21] Wallace V Friesen and Paul Ekman. EMFACS-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2(36):1, 1983.

[22] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[23] Thiago H H Zavaschi, Alceu S Britto, Luiz E S Oliveira, and Alessandro L Koerich. Fusion of feature sets and classifiers for facial expression recognition. *Expert Systems with Applications*, 40(2):646–655, 2013.

[24] Adin Ramirez Rivera, Jorge Rojas Castillo, and Oksam Oksam Chae. Local directional number pattern for face analysis: Face and expression recognition. *IEEE transactions on image processing*, 22(5):1740–1752, 2013.

[25] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.