ISSN: 0974-5823 Vol. 7 No. 7 July, 2022

International Journal of Mechanical Engineering

Human Facial emotion recognition using Convolutional Neural Network

Yash Kumar
Student
Bhagwan Parshuram Institute of Technology
Delhi, India

Samarth Jain
Student
Bhagwan Parshuram Institute of Technology
Delhi, India

Saksham Arora
Student
Bhagwan Parshuram Institute of Technology
Delhi, India

Tanisha Madan
Assistant Professor
Bhagwan Parshuram Institute of Technology
Delhi, India

I. ABSTRACT

Emotion recognition based on facial expression is a fascinating study area that has been presented and implemented in a variety of fields, including safety, health, and human-machine interactions. Researchers in this subject are interested in developing strategies to understand, code, and extract facial expressions in order to improve computer prediction. With deep learning's astonishing success, several sorts of architectures for this technology are being used to improve performance. nonverbal way of emotional expression, and thus can be used to uncover whether an individual is lying or speaking the truth [3]. In the 20th century, the American psychologists Ekman and Friesen [4] defined six basic emotions (anger, fear, disgust, sadness, surprise and happiness), which are the same across various countries and cultures. This paper focuses on seven essential facial expression classes reported, which are angry, disgust, fear, happy, sad, surprise, and neutral.

Facial emotion recognition has seen increased usage in various industries, namely healthcare, education and gaming industries. As an example, with the recent online examination era due to the global covid-19 pandemic, online classes were being conducted all over the globe. Using facial expression of students, the teacher can adjust their teaching strategy and adjust their educational materials to help foster the education of students. Thus, FER algorithms can be used to increase quality of products and services in various sectors, which is the purpose of this research.

The algorithm presented in this research paper aims to classify facial images into these seven emotion categories. Research shows that facial expression detection can be implemented by following two different approaches, the first one is distinguishing expressions which are classified using an explicit classifier [5], and the second one is classifying images based on the extracted facial highlights [6].

This article follows the first approach. The algorithm consists of three phases: face detection, normalization and emotion recognition classified

The goal of this research is to provide a review of recent work on deep learning-based automatic facial emotion recognition (FER). Our system consists of three phases: face detection using Haar Cascades, normalization and emotion recognition using CNN on FER 2013 database with seven types of expression. The interest of this paper is to understand the working of machine learning models and how we can predict values based on dataset trained.

Keywords— Student facial expression, Emotion recognition, Convolutional neural networks (CNN), Deep learning

II. INTRODUCTION

Facial expressions are vital identifiers for human feelings as the face is the most expressive and into one of the seven emotions taken into consideration.

The rest of this article follows the following structure: Section III reviews some related work, followed by Section IV which discusses the dataset used and section V which discusses the proposed algorithm. The implementation details are presented in the next section VI, followed by the experimental results future scope in the next section VII.

III. RELATED WORK

All humans use facial expressions to communicate their feelings. Many attempts have been made to create an automatic facial expression analysis tool since it has applications in a variety of industries including robotics, medicine, driving assist systems, and lie detectors. [7,8,9]. Many academics are interested in using Face Emotion Recognition to improve the learning environment (FER).

I. Lasri, A. South, and M. Belkacemi [11] proposed a similar method that was successful in accurately identifying expressed emotion.

Savva et al. [12] suggested a web application that analyses the emotions of students engaged in active face-to-face classroom education.

The tool collects live recordings from webcams deployed in schools, then applies machine learning algorithms to classify the data.

The authors of [13] presented a system that recognizes and monitors student emotion and provides real-time feedback in order to improve the e-learning environment and supply more content.

The technology deduces crucial information from students' eye and head movements to comprehend their mood in an e-learning environment.

Ayvaz et al. [14] created a Facial Emotion Recognition System (FERS) that recognizes students' emotional states and motivation in videoconference-based e-learning.

The system employs four machine learning algorithms (SVM, KNN, Random Forest, and Classification & Regression Trees), with the KNN and SVM algorithms achieving the highest accuracy rates. Kim et al. [15] created a system that may offer real-time recommendations to the teacher in order to improve the memorability and quality of their lecture by allowing the teacher to change their non-verbal behavior such as body language and facial expressions in real-time.

In [16], the authors proposed a model for detecting emotions in a virtual learning environment based on facial emotion detection using the Haar Cascades approach [17] to identify mouth and eyes on the JAFF database.

Chiou et al. exploited wireless sensor network technology to construct an intelligent classroom management system that allows teachers to quickly change instruction modes to avoid wasting time in [18].

IV. DATASET USED

To train our CNN architecture, we used the FER2013 database. The data consists of 48x48 pixel greyscale images of faces. The faces have been automatically registered so that the face is centered and occupies about the same amount of space in each image.

The task is to categorize each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set consists of 28,709 examples and the public test set consists of 3,589 examples.

Figure 1: 7 universal emotions

Emotion Label	Emotion	Number of Images	Training Images	Test Images
0	Angry	4953	3995	958
1	Disgust	547	436	111
2	Fear	5121	4097	1024
3	Happy	8989	7215	1774
4	Sad	6077	4830	1247
5	Surprise	4002	3171	831
6	Neutral	6198	4965	1233

Table 1: Image classification in train dataset

communicative part of a human being [1]. The facial expression is roughly in 55% of cases [2] a

V. **PROPOSED ALGORITHM**

The proposed work uses the deep learning convolutional neural network (CNN) for human facial emotion recognition, which to extract features from the input image and identify the human emotional state (e.g., anger, contempt, disgust, fear, happiness, sadness, and surprise).

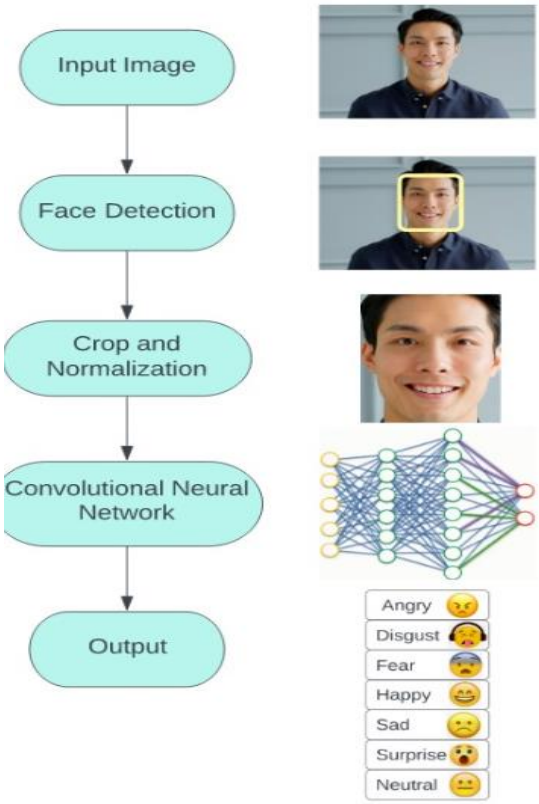
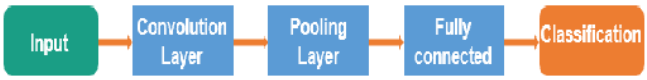


Figure 2: Algorithm overview

We used the CNN architecture. A Convolutional Neural Network (CNN) is a deep artificial neural network that can identify visual patterns from input image with minimal pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered [19]. The important unit inside a CNN layer is a neuron. They are connected, in order that the output of neurons at a layer becomes the input of neurons at the next layer. In order to compute the partial

derivatives of the cost function the backpropagation algorithm is used. The term convolution refers to the use of a filter or kernel on the input image to produce a feature map. In fact, CNN model contains 3 types of layers as shown:



Convolution Layer: is the first layer to extract features from an input image. The primary purpose of Convolution in case of a ConvNet is to extract features from the input image. Convolution preserves the spatial relationship between pixels by learning image features using small squares of input data [20]. It performs a dot product between two matrices, where one is the image, and the other is a kernel. The convolution formula is represented in Equation 1: net (t, f) = (x ∗ w)[t, f] = ∑m ∑n x[m, n]w[t − m, f − n] (1) Where net(t, f) is the output in the next layer, x is the input image, w is the filter matrix and ∗ is the convolution operation.

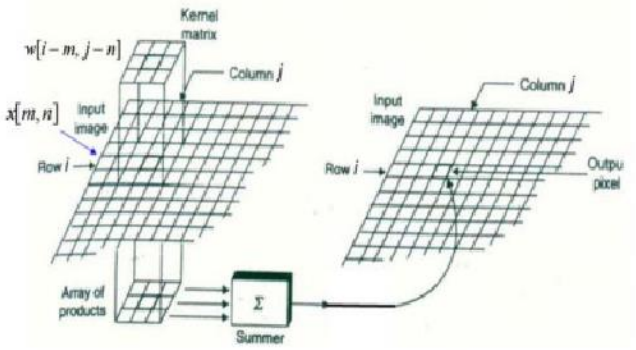


Figure 3: working of convolution Layer

Pooling Layer: reduces the dimensionality of each feature map but retains the most important information [20]. Pooling can be of different types: Max Pooling, Average Pooling and Sum Pooling. The function of Pooling is to progressively reduce the spatial size of the input representation and to make the network invariant to small transformations, distortions and translations in the input image [20]. In our work, we took the maximum of the block as the single output to pooling layer as shown

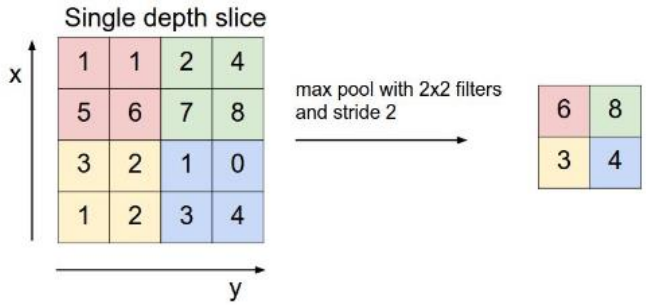


Figure 4: working of pooling layer

Fully connected layer: it is a traditional Multi-Layer Perceptron that uses an activation function in the output layer. The term "Fully Connected" implies that every neuron in the previous layer is connected to every neuron on the next layer. The purpose of the Fully Connected layer is to use the output of the convolutional and pooling layers for classifying the input image into various classes based on the training dataset. So, the Convolution and Pooling layers act as Feature Extractors from the input image while Fully Connected layer acts as a classifier [20].

Our proposed algo is as follows:

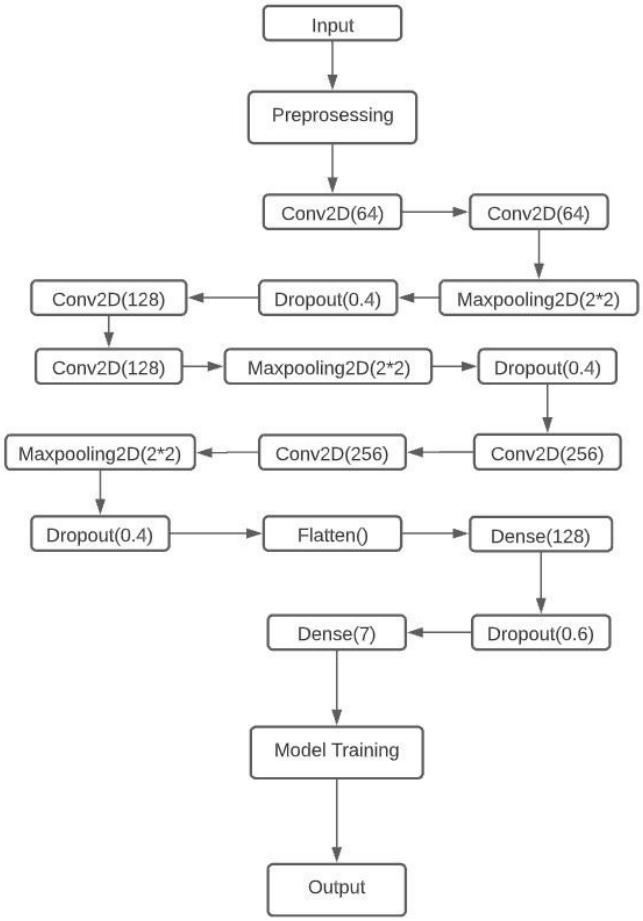


Figure 2: Proposed algorithm overview

We used the CNN architecture. The input image will be of 48*48*1 pixel greyscale face images. The images are centered and occupied equal amount of space. We are using 6 convolution layers with the first layer of the CNN is a sequential layer after which a convolutional layer with filter 64, kernel size (5, 5), 'elu' activation, 'same' padding and 'he_normal' kernel activation with 48*48*1 input shape size. The next layer is a batch normalization layer which allows every layer of the network to do learning more independently. The fourth layer is also a convolutional with filter 64, kernel size (5, 5), 'elu' activation, 'same' padding and 'he_normal' kernel activation.

The next layer is also a Batch normalization layer. Followed by a maxpooling layer with 2*2 pool size. After these layers, we have one dropout layer with a dropout value of 0.4 that means 40% of neutrons are ignored and this is to avoid over-fitting problem. The next layer is also a convolutional layer with filter 128, kernel size (3, 3), 'elu' activation, 'same' padding and 'he_normal' kernel activation. Followed by a batch normalization layer. The previous convolutional layer is repeated again with filter 128, kernel size (3, 3), 'elu' activation, 'same' padding and 'he_normal' kernel activation. Followed by a batch normalization layer again. The next layer is a maxpooling layer with 2*2 pool size. After these layers, one dropout layer is used with the dropout value of 0.4. The next layer is also a convolutional layer with filter 256, kernel size (3, 3), 'elu' activation, 'same' padding and 'he_normal' kernel activation. Followed by a batch normalization layer. The last two layers are repeated one more time. A maxpooling layer is used with a pool size of 2*2, if you notice for each layer the filter count is increasing, because the initial layer represents the high-level features of the image, and the deeper layers will represent more detailed features and so they usually have number of filters. After six convolutional layers, we have one more dropout layer with the dropout value of 0.4. And once the image passes through the convolution layers, it has to be flattened again to be fed into fully connected layers (it's called a dense layer). We have 2 dense layers and the first one is having 128 neurons and elu activation, this is also arbitrary, and we can have the neuron count as per our choice. Followed by another batch

normalization layer and a dropout layer with the value of 0.6. The second dense layer will have only 7 neurons as we have only seven classes to classify, usually, the number of neurons in the output layer will be equal to the number of classes in our problem. This layer will use SoftMax activation. SoftMax activation will calculate the probabilities of each target class over all possible target classes and the sum of all the probabilities will always be 1. The input will be classified into any of the target class based on the higher probability value in SoftMax. We are using Adam optimizer with "categorical-crossentropy" as loss function and learning rate of 0.0001 and decay would be e^6. We train our model with 55 epochs after which we stopped due to overfitting (for every epoch the model will adjust its parameter value to minimize the loss), there are 28709 training steps and 7178 validation steps.

VI. **RESULTS**

To analyze the performance of the algorithm, extended a test set from FER 2013 was used. The results at 55 epochs for different activators were as shown:

Figure 3: Model predicting a happy face

Figure 4: Model predicting an angry face

Activation	Accuracy
Elu	83.1%
Relu	82.1%
Selu	82.4%
SoftPlus	82.8%

Table 2: Accuracy achieved using different activators

VII. CONCLUSION AND FUTURE SCOPE

In this paper, the aim was to classify facial expressions into one of seven emotions by using various models on the FER2013 dataset. Model was trained using multiple layers and was implemented using CNN algorithm. The effects of different hyper parameters on the final model were then investigated. The final accuracy of 0.82+ was achieved using the Adam optimizer. It should also be noted that a nearly state-of-the-art accuracy was achieved with the use of a single dataset as opposed to a combination of many datasets. While it is true that other related works have managed to obtain higher accuracies, they have used a combination of different datasets and large models in order to increase their overall accuracy. Given that only the FER-2013 dataset was used in this case without the use of other datasets, an accuracy of 0.82 is admirable as it demonstrates the efficiency of the model. In other words, the model demonstrated has used significantly less data for training and a deep but simple architecture to attain near-state-of-the-art results. At the same time, it also has its shortcomings. While the model did attain near-state-of-the-art results, it also means that it did not achieve state-of-the-art. Additionally, the relatively lower amount of data for emotions such as "disgust" make the model have difficulty predicting it. This however does illuminate a path for future work. If provided with more training data while still retaining the same network structure, the efficiency of the proposed system will be enhanced considerably. In the future, the model can be further improved by adding more layers to the model, increasing the training dataset and implementing real time emotion recognition using a video camera support. This also implies that with some work, the model could very well be deployed into real-life applications for effective utilization in domains such as in healthcare, marketing and the video game industry.

VIII. REFERENCES

[1] R. G. Harper, A. N. Wiens, and J. D. Matarazzo, Nonverbal communication: the state of the art. New York: Wiley, 1978 Engineering and Education," IEEE Access, vol. 6, p. 5308-5331, 2018

[2] Mehrabian A (2017) Nonverbal communication. Routledge, London

[3] Bartlett M, Littlewort G, Vural E, Lee K, Cetin M, Ercil A, Movellan J (2008) Data mining spontaneous facial behavior with automatic expression coding. In: Esposito A, Bourbakis NG, Avouris N, Hatzilygeroudis I (eds) Verbal and nonverbal features of human–human and human–machine interaction. Springer, Berlin, pp 1–20

[4] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," Journal of Personality and Social Psychology, vol. 17, no 2,
p. 124-129, 1971

[5] Gizatdinova Y, Surakka V (2007) Automatic detection of facial landmarks from AU-coded expressive facial images. In: 14th International conference on image analysis and processing (ICIAP). IEEE, pp 419–424

[6] Liu Y, Li Y, Ma X, Song R (2017) Facial expression recognition with fusion features extracted from salient facial areas. Sensors 17(4):712

[7] Ali N, Zafar B, Riaz F, Dar SH, Ratyal NI, Bajwa KB, Iqbal MK, Sajid M (2018) A hybrid geometric spatial image representation for scene classification. PLoS ONE 3(9): e0203339

[8] Ali N, Zafar B, Iqbal MK, Sajid M, Younis MY, Dar SH, Mahmood MT, Lee IH (2019) Modeling global geometric spatial information for rotation invariant classification of satellite images. PLoS ONE 14:7

[9] Ali N, Bajwa KB, Sablatnig R, Chatzichristofis SA, Iqbal Z, Rashid M, Habib HA (2016) A novel image retrieval based on visual words integration of SIFT and SURF. PLoS ONE 11(6): e0157428

[10] I. Lasri, A. R. Solh and M. E. Belkacemi, "Facial Emotion Recognition of Students using Convolutional Neural Network," *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, 2019, pp. 1-6, doi:10.1109/ICDS47004.2019.8942386.

[11] Pravallika, D. S., Sai, V. M., Sahithi, D. L., & Rishitha, K. (2020). "Face Expression Recognition by Hybrid Local Binary Pattern with Haar Cascade Method". Solid State Technology, 63(6), 12919-12927.

[12] A. Savva, V. Stylianou, K. Kyriacou, and F. Domenach, "Recognizing student facial expressions: A web

application," in 2018 IEEE Global Engineering Education Conference (EDUCON), Tenerife, 2018, p. 1459-1462

[13] Krithika L.B and Lakshmi Priya GG, "Student Emotion Recognition System (SERS) for e-learning Improvement Based on Learner Concentration Metric," Procedia Computer Science, vol. 85, p. 767-776, 2016

[14] U. Ayvaz, H. Gürüler, and M. O. Devrim, "USE OF FACIAL EMOTION RECOGNITION IN E-LEARNING SYSTEMS," Information Technologies and Learning Tools, vol. 60, no 4, p. 95, sept. 2017.

[15] Y. Kim, T. Soyata, and R. F. Behnagh, "Towards Emotionally Aware AI Smart Classroom: Current Issues and Directions for

[16] P. Viola and M. Jones, "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, Kauai, HI, USA, 2001, vol. 1, p. I-511-I-518.

[17] D. Yang, A. Alsadoon, P. W. C. Prasad, A. K. Singh, and A. Elchouemi, "An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment," Procedia Computer Science, vol. 125, p. 2-10, 2018

[18] C.-K. Chiou and J. C. R. Tseng, "An intelligent classroom management system based on wireless sensor networks," in 2015 8th International Conference on Ubi-Media Computing (UMEDIA), Colombo, Sri Lanka, 2015, p. 44-48

[19] aionlinecourse.com/tutorial/machine-learning/convolution-neural-network

[20] ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/