Survey on Facial Emotion Recognition and Detection Using Deep Learning

V. Sai Nikhitha[1], K.V. Sree Divya[2], R. Shreya[3]

[1,2,3]*Electronics and Computer Engineering Department, Sreenidhi Institute of Science and Technology Ghatkesar, Hyderabad, Telangana, India*

***Abstract* - In recent times, due to evolution of technologies and the role they play in our day to day lives along with human's interpersonal communication, human and machine or human computer interactions have become very basic. When human's interpersonal communication is concerned, humans are well trained in reading the emotions of others even when the other person is not communicating verbally. But this is not possible with human computer/machine interactions. To make this possible the use of Convolution neural network (CNN) in facial emotion recognition and detection technology plays a vital role.**

***Index Terms* - Communication, Convolution neural network (CNN), Facial emotions, Facial expression recognition.**

I.INTRODUCTION

Facial emotion recognition is a method used for detection and recognition of human emotions.[11] Facial expressions are important in any social setting to display and identify an individual's emotion and also be used to interpret their feelings to establish interpersonal relationships, which are important for any social interaction. Although humans acknowledge facial expressions almost immediately, reliable expression recognition by machine continues to be a challenge. While many algorithms have been proposed, there are still many difficult and challenging problems in face recognition, such as facial expressions, pose variations, illumination variations, facial occlusion, and face rotation. The primary goal is to classify each facial image into one of the seven basic facial emotional categories (anger, fear, happy, sad, neutral, disgust, and surprise) [1]. Facial Expression Recognition is usually performed in four-stages consisting of pre-processing, face detection, feature extraction, and expression classification.[16] Convolutional networks, a subfield of deep learning, can intrinsically learn the key facial features by using only raw pixel data. This convolutional network is also used to handle a variety of problems, including excessive makeup [2], pose, emotion and expression [3]. Therefore making this system reliable to detect and recognise facial emotion even when the person or user is not verbally communicating.

II.DEEP LEARNING

Deep learning is a field which is very accurate for facial emotion detection as it addresses the feature selection issues alongside various learning tasks. The deep learning algorithms are used for facial emotion recognition (FER) as here the process of feature extraction is automated and unique features are extracted automatically.[12] The deep learning algorithms like Convolutional neural networks (CNN) consists of a hierarchical architecture and the feature extraction becomes more accurate as an image passes through the number of layers.[13]

III.WHY CNN

CNN is one of the subfields of Deep learning consisting of various layers hierarchically. CNNs are highly layered structural neural networks with basic layers such as :

- Feature extraction layer
- Classification layer

Convolution layer, pooling layer, softmax layer are these basic layers. CNNs are used for complex feature extraction from the image. It is mainly used for its high accuracy compared to any other algorithm. In other algorithms which consist of pixel vectors, the spatial interaction between the pixels is lost while downsampling. However the CNN algorithm uses the adjacent pixel information to effectively downsample the image by convolution layer followed by prediction layer. CNN also employs hierarchical patch-based convolution procedures, which not only decreases computational cost but also abstracts images on multiple feature levels.

IV.HOW CNN LEARNS

Kernels are often referred to as feature identifiers and are used to identify individual features. As an outcome, the crucial step of training begins. Backpropagation is another name for the training process, which is divided into four distinct phases.
The Forward Pass:
The initial kernels of the first convolution layer are initialized with random values for the first epoch or iteration of the training. As a result, after the first iteration, the output will look like [.1.1.1.1.1.1.1.1.1.1], with no preference for any class due to the kernels lack of specified weights.

The Loss Function:
Because the training includes images and labels, the label for the digit 3 will be, but the output after the first epoch is considerably different, we will measure loss (MSE — Mean Squared Error).
$Etotal=\Sigma\ 1/2(target-output)2$
The goal is to minimize the loss, which is a calculus optimization problem. It entails attempting to lessen the loss by adjusting the weights.
The Backward Pass:
It entails analyzing which weights contributed the most to the loss and devising strategies to reduce the loss. It's calculated using the formula dL/dW, where L stands for loss and W stands for the weights of the matching kernel.
The weight update:
This is where the kernel's weights are adjusted using the equation below.
$w=wi -\eta\ dL/dW$
w = Weight
wi =Initial weight
η=Learning Rate =Initial weight
The programmer determines the Learning Rate in this case. A higher learning rate signifies a bigger number of steps to optimize and a longer time to convolve to an optimal weight.[10]

V.FER STAGES

Some of the basic facial emotion recognition stages are as follows:

- Load and process
- Training
- Image capturing
- CNN

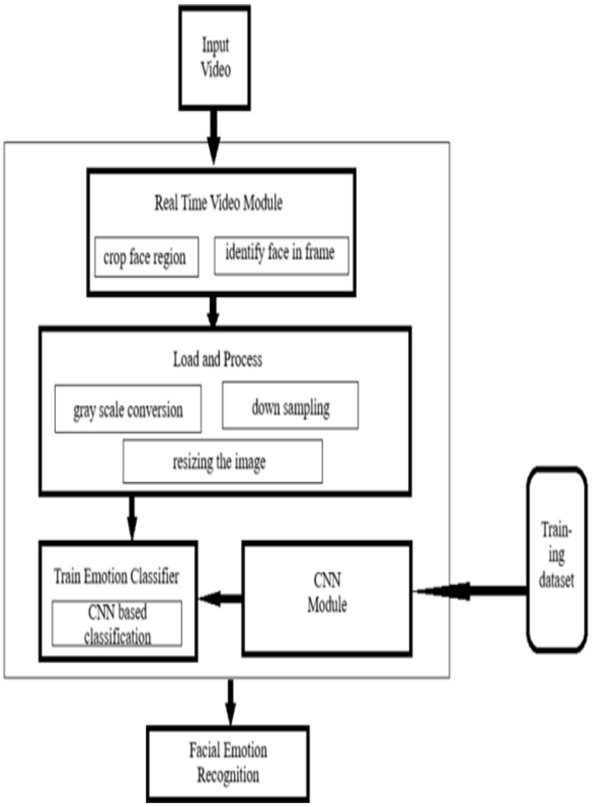

Fig 1. Basic architecture of facial emotion recognition system

1.LOAD AND PROCESS

A dataset is to be considered first (here the dataset can be taken from a number of online sources like (FER 2013). The file to be imported can be of any format like csv, jpeg, png etc. Then the preprocessing takes place, in this step images uploaded are resized and converted into grayscale images. And then these images are passed for training into the training module.[4]
The classifier used for face detection is haar cascade classifier. Haar cascade is an Algorithm for object detection that detects faces in images and real-time videos. The algorithm is given a large number of positive images with faces and a large number of negative photos without faces. Pixels with a value of 1 are darker in the haar feature, while pixels with a value of 0 are lighter. Each of these is in charge of

identifying a specific feature in the image. Any structure in the image with a quick shift in intensities, such as an edge, a line, or any other structure. The goal is to calculate the sum of all image pixels in the darker part of the haar feature, as well as the total of all image pixels in the lighter part of the haar feature. Then figure out how they differ. The haar value will be closer to 1 if the image has an edge dividing dark pixels on the right from light pixels on the left. There is an edge detected if the harr value is closer to 1.[5]

2.TRAINING

After obtaining training data consisting of grayscale images of faces with their proper expression labels, the system learns a set of weights for the network. The image is then normalized in terms of intensity. The normalized images are used to train the Convolutional Network. To ensure that the training performance is unaffected by the order in which the examples are presented, the validation dataset is used to choose the final optimal set of weights from a collection of training completed with samples delivered in different orders. The output of the training stage is a set of weights that offer the best outcome using training data.

3.IMAGE CAPTURING

It is a required step in facial recognition systems for localizing and extracting the face region from the background. OpenCV is a cross-platform library which is mostly used to create real-time computer vision applications. It is primarily concerned with image processing, video recording, and analysis, with features like face detection and object detection. With this library locating a face in a photograph refers to finding the coordinate of the face in the image, whereas localization refers to separating the extent of the face separating the face from surroundings by cropping it. It is an image-based method that learns how to extract faces from the entire image. The Video Capture () function is a function which allows us to connect to various devices such as a camera, a webcam on our computer etc. The video from the camera is captured and converted into grayscale. It accepts either the device index or the name of a video file as an input. A device index is just a number that indicates which camera is being used. Normally, only one camera is connected. As a result, there is an option to pass 0 (or -1). By passing 1,the second camera can be selected, and so on. After that, frame by frame can be captured.[6]

4.CNN

In the CNN module, first the input images are passed into the convolutional layer, the feature maps are built using the output of the convolutional layer and passed on to the next layer that is the max pooling layer, where features like horizontal and diagonal edges are extracted and this is continued several times for the feature extraction and detection of the emotion to extract even more complex features like objects and faces.[7]

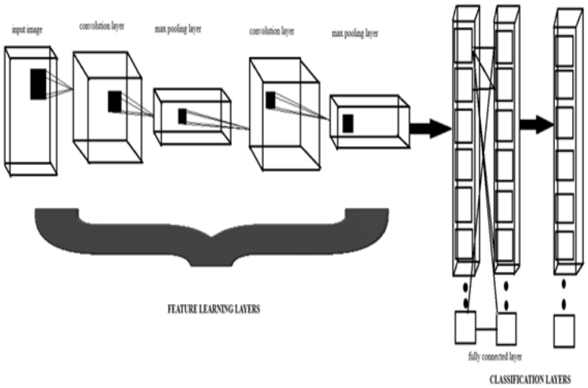

Fig 2: CNN architecture

4.1 FEATURE EXTRACTION LAYERS

CONVOLUTION LAYER

A Convolutional Neural Network's first layer is usually a Convolutional Layer. Convolutional layers perform a convolution operation on the input and send the output to the next layer. A convolution is a process that turns all of the pixels in its receptive area into a single value. For example, when convolution is applied to an image, it will reduce the image size while also merging all of the information in the field into a single pixel. The convolutional layer's final output is a vector.

POOLING LAYER

Pooling is a technique for reducing the width and height of an image. Pooling layers gives a method for down sampling feature maps by summing the existence of features in patches of the feature map. This is accomplished by using a filter (of a size 2x2 dimension) and a stride of the same length. It then applies it to the input volume and returns the highest number in each subregion that the filter convolves around. This fulfils two functions. The first is that the number of parameters or weights is reduced by 75%, lowering the computation cost. The second benefit is that it will prevent overfitting. When a model is so tailored to the training examples that it is unable to generalize successfully to the validation and test sets, this is referred to as overfitting.[8]

4.2 CLASSIFICATION LAYER

SOFTMAX LAYER

Softmax is a mathematical function that transforms a number vector into a probability vector. The softmax function converts a vector of K real values to a vector of K real values that add to one. The softmax translates input values that are positive, negative, zero, or higher than one into values between 0 and 1, allowing them to be understood as probabilities. If one of the inputs is a small value or negative, the softmax converts it to a low probability, and if an input is large or positive, the softmax converts it into a large probability, but it will always remain between 0 and 1.[9]

VI.LIMITATIONS

The major limitation for using deep learning in the facial emotion recognition system is that the accuracy will be purely dependent on the size of the raw data available for its training (for example: FER 2013 where the dataset is of size 35,000 images). Hence the dataset collection is the major problem.

VII.CONCLUSION

CNNs are capable of learning face traits and enhancing facial emotion detection, according to the findings. This means that by using only raw data, convolutional networks may learn the key facial traits on their own. As a result, this research is capable of detecting and recognizing facial emotion. The accuracy of detection can always be improved by extending the raw data limits, that is the accuracy of facial emotion detection here is always proportional to the amount of raw data. When the accuracy of the system is high, this system can be used in various fields like Entertainment industry for capturing the true feedback to produce desirable content, medical field to know patient's mental status, security purposes as in lie detectors. [14,15] It can also be used to interpret especially abled people's feelings who are not able to communicate verbally.

VIII.ACKNOWLEDGEMENT

We sincerely thank and express deep gratitude to my supervisor Dr. C. Sunil Kumar, Professor in Dept. of ECM, SNIST, for guidance, encouragement, and valuable suggestions throughout the course of our project. Without his constant guidance and continuous supervision, this thesis could not have been completed.

REFERENCES

[1] Matsumoto D (1992) More evidence for the universality of a contempt expression. Motiv Emot 16(4):363

[2] Sajid M, Ali N, Dar SH, Iqbal Ratyal N, Butt AR, Zafar B, Shafique T, Baig MJA, Riaz I, Baig S (2018) Data augmentation-assisted makeup-invariant face recognition. Math Probl Eng 2018:1–10

[3] Ratyal N, Taj I, Bajwa U, Sajid M (2018) Pose and expression invariant alignment based multi-view 3D face recognition. KSII Trans Internet Inf Syst 12:10

[4] K, renu. "loading custom image dataset for deep learning models." towardsdatascience.com, 2020, https://towardsdatascience.com/loading-custom-image-dataset-for-deep-learning-models-part-1-d64fa7aaeca6. Accessed 31st may 2021

[5] Khan, Tanwir. "Computer vision- Detecting objects using haar cascade classifier." towards data science, 2019, https://towardsdatascience.com/computer-vision-detecting-objects-using-haar-cascade-classifier-4585472829a9. Accessed 28 May 2021.

[6] n.p, n.p. "opencv." opencv.org, https://docs.opencv.org/master/dd/d43/tutorial_py_video_display.html. Accessed 28th May 2021.

[7] M, Manav. "Convolution Neural networks (CNN)." Analytics Vidhya, 2021, https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn. Accessed 27 May 2021.

[8] Deshpande, adit. "Understanding cnn." github,n.y,https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks-Part-2/.Accessed 27th may 2021.

[9] Wood, Thomas. “Softmax function.” DeepAI, ny, https://deepai.org/machine-learning-glossary-and-terms/softmax-layer. Accessed 20 May 2021

[10] Chatterjee, Chandra Churh. “Basics of the classic CNN.” towards data science, 2019, https://towardsdatascience.com/basics-of-the-classic-cnn-a3dce1225add. Accessed 20 May 2021.

[11] Mehendale, N. Facial emotion recognition using convolutional neural networks (FERC). SN Appl. Sci. 2, 446 (2020).

[12] C. M. M. Refat and N. Z. Azlan, "Deep Learning Methods for Facial Expression Recognition," 2019 7th International Conference on Mechatronics Engineering (ICOM), 2019, pp. 1-6, doi: 10.1109/ICOM47790.2019.8952056.

[13] G. E. Sakr, M. Mokbel, A. Darwich, M. N. Khneisser and A. Hadi, "Comparing deep learning and support vector machines for autonomous waste sorting," 2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), 2016, pp. 207-212, doi: 10.1109/IMCET.2016.7777453.

[14] 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

[15] 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

[16] Michael Revina, W.R. Sam Emmanuel, A Survey on Human Face Expression Recognition Techniques, Journal of King Saud University - Computer and Information Sciences,2018