Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 11, November-2022

A Comparative Study of CNN Techniques and Datasets Regarding Facial Emotion Recognition

Gopal Upadhye[1], Onkar Dhengale[2*], Dhruv Goel[3], Vrushabh Bhanjewal[4], Aditya Lokhande[5]
Dept of Computer Engineering,
Pimpri-Chinchwad College of Engineering, Nigdi[1,2,3,4,5].

***Abstract*:- Facial expression analysis has so many uses in artificial intelligence, including computer and human interaction such as their collaboration and communication, data-driven animation, etc., the ability to detect emotion from facial expression has become urgently necessary. [1] This paper's goal is to do an analysis of recent work on automatic facial emotion recognition using deep learning. We focus on the architecture and databases employed, the contributions that were dealt with, and we illustrate the progress made by contrasting the suggested procedures and the outcomes attained. The purpose of this study is to assist and direct scholars by reviewing previous efforts and offering suggestions for how to enhance this topic.**

Keywords:- ***Convolutional Neural Network, Face Detection, Feature Extraction, Deep Learning, HCI, Dataset, Haar-Cascade***

1. INTRODUCTION

The expression of emotions by a person through the movement of their facial muscles is referred to as facial expression. It reveals details regarding that person's mental health. Mood is a mental state. It is the respective person's internal reaction to what is happening to him outside of himself. [01] Numerous studies have demonstrated that the deformation of facial features, such as the eyes, brows, and lips, is what causes facial expressions.[02] The concept of building a duplicate of the human mind has always interested people. Additionally, capacity is the most crucial requirement for these systems, to grasp a person's feelings and respond to the situation appropriately. The real objective is to close the gap between a machine-like robot and humans, increasing its dependability.[3] Facial Emotion Recognition (FER) is widely used in real-world situations since it offers vital insights unique to that person. As they are working in various systems, huge amount of research has been carried out on them. FER can be utilized for a variety of purposes, such as medical operations like patient monitoring, to gauge an interviewee's emotional state. Emotional expression is a powerful form of communication and the cornerstone of interpersonal harmony, cooperation, and understanding. In the advancement of computer vision and artificial intelligence, the research on human sentiment in video and picture has emerged as a topic of interest in the domains of machine learning and pattern recognition. Future HCI will be more intelligent, and quick. Computers will be able to perceive, record, and discern human emotions as well as changes in emotion. Based on this, they will be able to produce effective and intelligent replies, thus giving machines "brains." Make it possible for machines to comprehend emotions in order to better serve human needs. In recent years, the field of emotion identification has started to use deep learning techniques because of the tremendous advancements in computer processing power and network architecture design. The research demonstrates that the CNN model has produced excellent results in face detection and facial expression recognition. The research also demonstrates that as the network layer depth increases, the network's ability to detect targets and facial expressions improves, as is the case with AlexNet [04] has 5 convolution layers. People now have increasingly complex needs for emotion detection as a result of affective computing's ongoing growth, and the classic discrete emotion recognition paradigm is unable to capture the full range of human emotions. As a result, dimensional emotion recognition has attracted a lot of attention lately. Emotions that have multiple dimensions are referred to as multi-dimensional emotions. [05] At some point in the emotional spectrum, human emotions can be measured, revealing complicated categories of human emotions like happiness, fear, wonder, and more. Dimensional emotional recognition offers more adequate and effective assistance for the research fields of affective computing, facial expression recognition, and emotional perception since dimensional emotional space encompasses all human emotions. [05] Facial images and videos can be captured anywhere thanks to the ubiquity and small size of the cameras hence this gives an upper hand in recognizing facial emotions.

2. RELATED WORK

The works that are connected to the suggested system is covered in this section. All of these techniques have selected unique strategies to address various issues and advance previous work in the field of emotion recognition. Different feature extraction and pre-processing approaches are included in this section even though this research only focuses on the CNN model as we are comparing the change in accuracy when these are combined with the CNN model. Rabie Helaly [06] presents an interesting approach of sentiment recognition system. The suggested approach is put into practise using the Raspberry Pi 4 embedded system. To accomplish that, they used the Xception CNN model to identify their emotional system. The embedded system's classifiers classify the registered facial photos into seven different facial expressions using them as input. The FER 2013 data collection is utilised for training. In terms of graphics processing unit accuracy, the proposed model

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 11, November-2022

provides 94% (GPU). Following implementation by Mohamed Ali Hajjaji, et al [06] on the embedded system, the accuracy on the Raspberry Pi 4 is 89% in accordance with his restriction on GPU performances.

Ansamma In addition to the CNN model, John et al. [07] provide an innovative way for enhancing real-time emotion identification. This method uses additional feature extraction techniques to improve training accuracy. Performance study was conducted using datasets like FER2013 and JAFFE. The first module uses a webcam to capture real-time footage and local binary patterns to identify faces (LBP) The following module performs feature selection for emotion recognition and pre-processing. Neural Convolutional Network The proposed structure consists of an input layer, two fully linked layers for classification, two layers of pooling, and four layers of convolution. The JAFFE and FER2013 datasets demonstrate the exceptional performance that the suggested approach is capable of. The results showed precision of 91.2% and 74.4%.

The difficulties of Emotion Recognition Datasets were studied by Sabrina Begaj et al [07], and we also experimented with various CNN settings and designs. ICV MEFED has been selected by Ali Osman Topal et al [07] as the primary dataset. There are three primary phases that are taken while using the deep learning approach: deep feature learning, deep feature classification, and pre-processing The first CNN network that was put into use had two fully connected layers, one dropout, four max pooling layers, and four convolutional layers. As can be seen in our Confusion Matrix, the system has thus far performed best in detecting pleased expressions and poorest in detecting contempt. With the FER2013 dataset, CNN reports an accuracy of 91.62%.

The author. [05] Shuang Liu used Convolutional Neural Network along with Keras deep learning network. Keras was beneficial to them as it is user friendly and modular. The main parameters of the convolutional layer consist of size of convolutional kernel, step size of convolutional translation and mode of padding. Pooling layer is a down sampling process, it is also called as down sampling layer (subsampling layer). Developers call the integrated network layer module which reduces the development cost and the code is easy to debug. The CNN designed in this paper has nine layers, 1 input layer and 1 output layer. Four convolutional layers and pooling layers are used.

The author proposed a method using CNN along with data augmentation where the image is sent for pre-processing if the face is detected. The preprocessing is done using Cascade Classifier. After that the data augmentation is done using the ImageDataGenerator function present in KerasAPI.

Facial detection, feature extraction, and facial expression categorization are the three primary elements of a standard Facial Emotion Recognition (FER) system. During the pre-processing phase, also known as face detection, face areas are found. The aim of facial feature extraction is to find the most accurate representation of facial images for recognition. According to Phavish Babajee[08], our approach is built on a geometric feature-based method that extracts data using an edge detection framework.

3. DATASETS

A collection of photos or video clips with various emotional facial expressions is called a facial expression database. For the development of expression recognition systems, well-annotated media content of facial movement is crucial for the training, testing, and confirmation of algorithms. Either a sustained scale or separate sentiment labels can be used for the annotation of emotions.

There are various datasets present today, the Japanese Female Facial Expressions formerly known as JAFFE [9] consists of seven basic emotions which are neutral, sadness, surprise, happiness, anger, fear, and disgust. It consists of total 326 statically posed greyscale images of resolution 256*256.

Another interesting dataset is Extended Cohn-Kanade (CK+) [10] dataset having eight different emotion 593 images having resolution 640*490 along with greyscale property.

FER-2013 [11] is a mostly used dataset which contains total 28000 training and 3500 validation and testing data each. The images present in this dataset varies in various factors like age, pose, etc. The FER-2013 dataset was made public during this competition. Aaron Courville and Pierre Luc Carrier developed FER-2013. It is a component of a continuous, bigger project. The dataset was developed by searching for photographs of faces that correspond to the list of 184 emotion-related phrases, such as “calm,” “frustrated,” etc., using the Google image search API.

There are [12] 755,370 photos in the multi-PIE database, covering 337 different topics. Minimum and maximum subject attendance ranged from 203 to 249 every session of 337 subjects, 264 had at least two recordings made, and 129 showed up in all four sessions. Most of the subjects were male. most of the participants were European Americans, 35% were middle-east Asians, 3% were African Americans, and 2% were others. The subjects were 27.9 years old on average which gives better aspects of different images.

AffectNet [14] is the name of the largest database of categorical and dimensional models of affect currently available (Affect from the Internet). Three search engines were used to query keywords associated with emotions to create the database, which was subsequently annotated by experts. The Affective-MIT Facial Expression Dataset (AM-FED) database contains 242 facial videos (160K frames) of individuals watching Super Bowl commercials on their webcams. The recording environment's lighting and contrast were chosen at random. Each frame's 14 FACS action units, head movements, and automatically identified landmarks were recorded in the database. AM-FED is an excellent tool for studying AUs in the real world. However, there is not a lot of variances in head posture.

4. RECOGNIZING EMOTION IN DEEP LEARNING

Seven facial emotion states are found in the current study using a deep neural network that simultaneously performs the three phases of features extraction, selection, and classification. In the previous few years, training networks with more than two layers was a complicated task, but with the advancement of Graphics Processors, it has become easier to manipulate neural networks with more than one

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 11, November-2022

layer. Convolutional, sub-sampling, and fully connected layers comprise the three interchanging types of layers found in deep neural networks.

4.1 Convolutional Neural Network

A CNN architecture is made up of a stack of unique layers that uses a variational function to transform the input volume into the output volume. The use of a few specific types of layers is widespread.

Convolutional layer is the first layer to extract features from the input image. In order to retain the connection between pixels, the convolutional algorithm will first learn image properties from the small input data squares. Applying filters to the convolution of an image allows better performance. When the image is too huge, the pooling layer's next purpose is to lower the number of parameters. The technique of spatial pooling, also known as subsampling or down sampling, decreases the number of dimensions on each map while preserving crucial data. There are three different types of spatial pooling: maximum, average, and total. A fully connected layer, which functions like a neural network, is created by flattening the matrix into vectors and feeding it into the layer. This layer has connection to all the previous neurons with added bias which are activation functions. Most common activation functions are sigmoid functions and ReLU. In sigmoid function the value ranges between 0 and 1. The ReLU consists of the rectifier-using units. Fixing all the negative values in the feature map to zero is a step in the operation. This is the most prevalent activation function in deep learning models; it is defined as the argument's positive component; if the rectifier takes any input that is - negative, it will go back to zero. And remains as it is in case of positive input. In the neural network's last layer, known as the output layer, relevant predictions are made. A neural network has a single output layer that generates the desired outcome. Prior to deriving the result, it applies its own range of weights and biases.

4.2 Image Pre-processing

The primary objective is to reduce the amount of input image's unnecessary information and improve the ability to recognize the key qualities. The elements of the image that were acquired with the camera were not necessary for identifying facial expressions. The entire neck, the hair, etc., are not necessary. These undesirable details were thus eliminated. If not, the recognition system will have to cope with additional data, which will make it more difficult and ineffective. This undesirable information is taken out of the raw image during pre-processing. Cropping, scaling, and intensity normalisation are a few of the pre-processing stage. The most crucial areas of the face for detecting emotion are those around the mouth and eyes. The cropped image is then further shrunk to guarantee that the pixel file's data size corresponds to CNN's input size.

About illumination and lighting conditions of the object, image brightness and contrast fluctuate [15]. These changes make feature sets and the detection procedure more complex. An intensity normalisation was used to minimise these problems. The suggested technique employs MinMax normalisation after the original image has undergone a linear change.

The process of artificially derived fresh data from previously collected training data is known as data augmentation. Techniques including cropping, cushioning, flipping, rotating, and resizing come under data augmentation. It strengthens the model's performance and addresses problems like overfitting and a lack of data.[16] Mollahosseini et al 275k images were divided, thereafter dividing a dataset into 40% for testing, 20% for training, and validation. The validation set was skipped in this instance to whenever the hyperparameters were tweaked, it was in support of retraining the entire model. While this took more effort and using more processing power, it gave a larger training set in the end. The labels were encoded using a single-hot method rather than using numbers from 0 to 6 to characterize emotions while the actual testing was conducted.

4.3 System Architecture

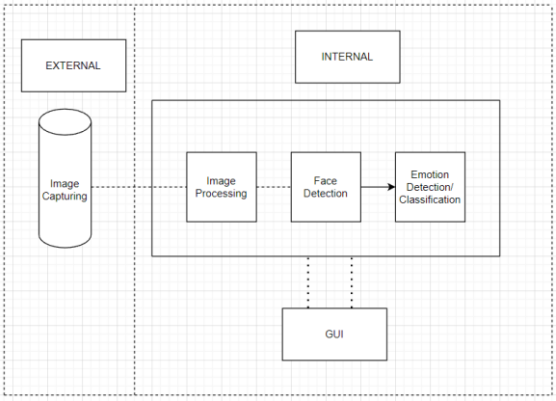

Figure 1 system architecture

The webcam on a computer or an external webcam is used to capture the human face. The face is removed from that live steam, and all other undesirable components are disregarded. We chose the Viola-Jones algorithm [13] for this work to achieve this efficiency and correctness.

Haar-cascade is fundamentally a classifier that is used to distinguish the object from the source for which it has been produced. The positive image is prepared by layering it over several negative images to create the Haar-cascade. The training is frequently carried out on a server and on several levels. Using best imagery and increasing the number of steps for which the classifier is prepared are necessary for better results. The available, preconfigured Haar- cascade can also be used. The Haar wavelet method is used by the Haar-cascade classifier to break down the pixels in the image into squares. The "integral image" concepts are used in this to identify the distinct "highlights." In order to produce an efficient classifier result, Haar-cascades uses the Ada-help learning computation, which selects a small number of important features of a large set.

The major area under the eyes on the face is used to identify emotions. The mouth and brows are considered. Also, the division of region-splitting is referred to as the mouth and the eyebrows. [17] Based on Xception and Convolution Neural Network (CNN), Authors have used mini-Xception that makes it simple to emphasis on key features like the face and draws significant gains over past research. Authors experimented with our concept by building a real-time

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 11, November-2022

vision system that uses their suggested mini-Xception architecture to simultaneously perform the tasks of face identification and emotion categorization. Based on the output of the classifier, they continue to employ a visualisation approach that is prepared to recognise significant facial regions while distinguishing different emotions. Authors used the FER-2013 dataset for experimental study, and the findings show that the suggested technique can effectively complete all tasks such as emotion detection and classification using seven distinct emotions.

5. CONCLUSION AND FUTURE WORK

The study highlights a contemporary FER research, enabling us to learn about the most recent advancements in this field. In order to have and attain an accurate detection of human emotions, we have discussed various CNN architecture lately proposed by various researchers, who also presented various datasets from the actual world, we have also given a traditional architecture for this system.

Facial recognition technology has a highly promising future. This technology is predicted to grow at a remarkable rate and produce significant income in the years to come. This technology would close the gaps in the widely used password technology. Robots that use facial recognition technology may eventually make an appearance. They may be useful in finishing jobs that are unrealistic or challenging for people to do using this technology.

6. REFERENCES

[1] T. U. Ahmed, S. Hossain, M. S. Hossain, R. ul Islam and K. Andersson, "Facial Expression Recognition using Convolutional Neural Network with Data Augmentation," 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2019, pp. 336-341.

[2] R. Guteri, A. Chetouani, H. Tabia and N. Khlifa, "Real time emotion recognition in video stream, using B-CNN and F-CNN," 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 2020, pp. 1-6

[3] A. John, A. MC, A. S. Ajayan, S. Sanoop and V. R. Kumar, "Real-Time Facial Emotion Recognition System With Improved Pre-processing and Feature Extraction," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1328-1333.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. COMMUNICATION OF THE ACM. 2017. (pp. 84-90).

[5] S. Liu, D. Li, Q. Gao and Y. Song, "Facial Emotion Recognition Based on CNN," 2020 Chinese Automation Congress (CAC), 2020, pp. 398-403.

[6] R. Helaly, M. A. Hajjaji, F. M'Sahli and A. Mtibaa, "Deep Convolution Neural Network Implementation for Emotion Recognition System," 2020 20th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), 2020, pp. 261-265.

[7] S. Begaj, A. O. Topal and M. Ali, "Emotion Recognition Based on Facial Expressions Using Convolutional Neural Network (CNN)," 2020 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA), 2020, pp. 58-63.

[8] P. Babajee, G. Suddul, S. Armoogum and R. Foogooa, "Identifying Human Emotions from Facial Expressions with Deep Learning," 2020 Zooming Innovation in Consumer Technologies Conference (ZINC), 2020, pp. 36-39.

[9] M. Lyons, M. Kamachi, and J. Gyoba, "The Japanese Female Facial Expression (JAFFE) Dataset." Zenodo, Apr. 1998.

[10] P. Lucey, et.al, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94-101.

[11] I. J. Goodfellow et al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," in Neural Information Processing, 2013, pp. 117–124.

[12] R. Gross, I. Matthews, J. Cohn, T. Kanade and S. Baker, "Multi-PIE," 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, 2008, pp. 1-8.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001, pp. I-I.

[14] A. Mollahosseini, B. Hassani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," CoRR, vol. abs/1708.03985, 2017, [Online]. Available: http://arxiv.org/abs/1708.03985

[15] PATRO, S GOPAL & Sahu, Dr-Kishore Kumar. (2015). Normalization: A Preprocessing Stage. IARJSET. 10.17148/IARJSET.2015.2305.

[16] A. Mollahosseini, D. Chan and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1-10.

[17] S. A. Fatima, A. Kumar, and S. S. Raoof, "Real Time Emotion Detection of Humans Using Mini-Xception Algorithm," IOP Conference Series: Materials Science and Engineering, vol. 1042, no. 1, p. 012027, Jan. 2021.