A Review of Various Deep Learning Models and Datasets for Emotion Recognition

Nayana N Panicker
M.Tech, CSE Dept.
Mangalam College of Engineering
Ettumanoor, Kottayam, India

Dr. K. John Peter
Prof., CSE Dept.
Mangalam College of Engineering
Ettumanoor, Kottayam, India

***Abstract*—Emotion recognition based on facial expressions is an attractive research area, and it has applications in different areas such as safety, health, customer service, and human-machine interfaces. Due to the exceptional success of this technology, numerous types of deep learning architecture are being explored to gain higher performance. The main objective of this paper is to conduct a review of recent research on facial emotion recognition (FER) through deep learning. In this, highlight the contributions that have been addressed as well as the architecture and databases that have been employed and illustrate the progress made by evaluating the recommended approaches with the results acquired. Deep learning models, particularly convolutional neural networks (CNNs), have shown tremendous potential among all FER approaches because of their powerful automated feature extraction and computational efficiency.**

Keywords- Facial emotion recognition, Deep learning, Facial Databases

I. INTRODUCTION

Facial expression recognition (FER) is a technique that uses human facial images to predict fundamental facial expressions. Because of its potential use in human abnormal detection, computer interfaces, autonomous driving, health management, and other comparable activities, FER has gained a lot of attention. In recent years, as pattern recognition and artificial intelligence have grown in popularity, more and more research has been conducted on the subject of human-computer interaction. In social communication, human facial expressions are significant. In most cases, both verbal and nonverbal communication are used. Eye contact, movements, facial expressions, body language, and languages are examples of nonverbal communication between humans. Despite the fact that other emotion description models, and the continuous model using the affect dimensions, are thought to represent a broader range of emotions, the categorical model, which describes emotions in terms of discrete basic emotions, remains the popular perspective for FER.

Deep learning employs artificial neural networks., which is a type of machine learning, which are analogues of the human brain, learn from massive amounts of data. The deep learning system would repeat a task, similar to how we learn from experience. It will get better each time until it reaches the desired result. Computer image data is used to train a model for purposes of identification, verification, or control. Deep neural networks are used to analyse the patterns and characteristics contained in the images. According to the feature representations, FER systems may be split into two categories: static image FER and dynamic sequence FER. In static-based approaches, the feature representation is encoded using simple spatial information from the current images, while in dynamic-based methods, the provided facial expression sequences' temporal connection between consecutive frames is taken into consideration.

In most recent studies, facial expression recognition is still facing some challenges because of several factors such as, head deflection, partial occlusion of face areas, and lighting changes. Face detection performance may be significantly impaired by these interferences, and FER accuracy may be reduced. As a result, deep learning may have been a suitable solution to these problems.

Convolutional neural networks (CNNs) have already made significant progress in pattern recognition, particularly in the identification of faces and the recognition of handwritten mathematical expressions. With a deep network, CNN can automatically interpret and learn the target's abstract signatures. Due to deep layers and complex architecture, i.e., CNN, or any other deep network, can properly execute FER under extreme situations.

The common drawbacks of emotion recognition systems are: (a) misclassification problems; (b) small alignment problems affect the performance; (c) a fully connected neural network can't learn the complicated local pixel relationship in image data well; and (d) in convolutional neural network-based architecture, the local spatial features like eyes, noses, etc. are learned well, but global spatial features like the animal as a whole or face as a whole are not much learned; and (e) dataset problems that affect the performance; and (f) contains mislabeled data.

The rest of this survey is constructed as follows. Section II provides a thorough examination of related work on FER approaches. Section III provides the details of the proposed model's performance and its comparison. Section IV describes the conclusion.

II. LITERATURE SURVEY

K. Zhang et al. [1] proposed evolutional spatial-temporal networks for facial emotion identification to extract many types of features. To extract dynamic geometry information, they propose a PHRNN model. Landmarks are divided into separate portions based on facial morphological variances, making it easier to describe dynamical expression evolution. After that, they developed an MSCNN model that uses both recognition and verification signals as supervision in order to supplement the still appearance information. The two signals correspond to two different loss functions that help to increase the variation of different expressions while decreasing the difference between identical expressions. Here, the models used in the database are CK+, Oulu-CASIA, and MMI databases. This approach minimises the error rates of the previousbest methods by 45.5%, 25.8%, and 24.4%, respectively, across three facial expression databases. They then found that it is difficult to capture spatial-temporal information about expressions with little motion in the confusion matrices across thethree databases. As a result, more sophisticated structures to describe the movements of these important regions are needed, asare particular approaches to assess these emotions, such as metriclearning.

In 2018, B. Yang et al. [2] proposed a weighted mixture deep neural network (WMDNN) to automatically extract the parameters that are essential for FER tasks. For input facial data, several preprocessing techniques are necessary, namely face detection, rotation rectification, and data augmentation. WMDNN has two channels of facial images, including facial grayscale images and their equivalent local binary pattern (LBP) images. For facial grayscale images, a partial VGG16 network with starting parameters derived from the VGG16 model pretrained on ImageNet is constructed to automatically extract expression-related attributes. A shallow convolutional neural network (CNN) based on DeepID features is extracted from LBP facial images as shown in Fig 1. The datasets used are CK+, JAFFE, and Oulu-CASIA. The proposed approach can successfully classify facial expressions with accuracy. Even though the accuracy obtained, some recognition becomes failed. Then later, work will concentrate on simplifying the network used to speed and then focus on other channels of facial images that can be used to improve the fusion network. It works well on some specific features only.

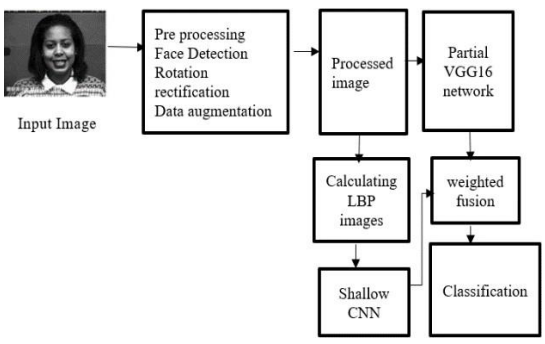

Fig 1. Block diagram of WMDNN model, B. Yang et al. [2]

In 2019, H. Zhang et al. [3] proposed a facial expression detection technique based on a CNN model and image edge detection that efficiently extracts the facial attributes. The edgeof each layer of the data is retrieved in the convolution process after the facial expression image is normalized. To maintain the texture image's edge structure information, the retrieved edge information is placed on each feature image. The maximum pooling approach is then used to reduce the dimensionality of the retrieved implicit features. The suggested technique is compared to the traditional neural network FRR-CNN model and the R-CNN algorithm in order to validate the robustness of this method for facial expression recognition under complicated backgrounds, as shown in Fig 2. It was created by combining the Fer-2013 facial expression database withthe LFW data set in a scientific way. The suggested method achieves an average recognition rate of 88.56% with fewer iterations, and the training speed on the training set is roughly 1.5times quicker than the contrast approach. The issue of datasets is very challenging, and noisy variation (such as face posture, occlusion, and blurring) of these datasets affects the performance and needs more robust models that satisfy real conditions.

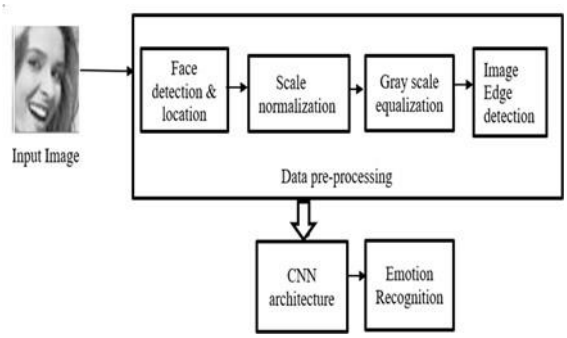

Fig 2. Block diagram of CNN approach, H. Zhang et al. [3]

In 2020, Garima Verma et al. [4] proposed a convolution neural network (CNN)-based deep learning model for analysing facial expressions and predicting emotions, which is commonly used to handle computer vision issues such as object identification, object tracking, image classification, and image segmentation. Two component models are included in the suggested approach. The first is a CNN model that recognises a secondary emotion, such as love or affection, while the second is a CNN model that classifies the major emotion shown in an image, such as pleasant or sad. The FER2013 and JAFFE datasets are used in this model. For developing the model, preprocessing steps are included. Three convolution layers and three fully linked layers with 1024 neurons each comprise the primary CNN. The last fully connected layer is a two-neuron layer that uses the SoftMax function to apply the classification. After that, the secondary CNN has five 2D convolution layers, each of which is connected to the max-pooling layer. Two dropout layers with a dropout rate of 0.2 are added into the network to reduce overfitting and decrease training time. The model was trained using the FER2013 and JAFFE datasets and achieved accuracy rate. The mislabeled data is still challenging. Therefore, we need advanced CNN-based models.

In 2021, N. Zhou et al. [5] propose and construct a lightweight convolutional neural network (CNN) for recognising facial emotions in real time and in bulk. To accomplish face identification and transmit the collected face coordinates to the facial emotion classification model they constructed initially, this system uses multi-task cascaded convolutional networks (MTCNN). Multi-task cascaded convolutional networks contain a cascade detection function, one of which may be used independently, minimising the memory usage. To replace the fully connected layer in the usual deep convolution neural network model, the classification model uses global average pooling. The fully connected layer's black box features are somewhat eliminated by associating each channel of the feature map with the relevant category. Simultaneously, the model combines residual modules with depth-wise separable convolutions and adds the normalisation term, resulting in a reduction of large numbers of parameters and increased portability. Here, the FER2013 dataset is used. The model achieved accuracy. Even though this model attains accuracy as compared to other recent works, there may be a lot of noise in real-life facial expressions, such as images with too bright or too dark illumination, blurred images, the majority of the face being blocked, and other variables that make recognition difficult.

A. EMOTION RECOGNITION TECHNOLOGY

Neural Networks: These are computer systems made up of connected nodes that function similarly to neurons in the brain. Using algorithms, itcan find hidden patterns and correlations in raw data, cluster and categorise it, and learn and improve over time. Most of the researchuses a model called a convolutional neural network. CNN is a feed-forward neural network that processes input in a grid-like structure to evaluate visual images. CNN is employed to recognise and categorise items in an image. Multiple hidden layers enable the extraction of informationfrom an image in a convolution neural network. There are four layers to CNN. The technique of collecting important characteristics from an image using a convolutional layer consistsof many filters in a convolution layer. Next, ReLU conducts an element-by-element procedure, setting all negative pixels to zero. Pooling is a down-sampling technique that reduces the dimensionality of a feature map. To create a pooled feature map, the corrected feature map is now sent via a pooling layer. Recurrent neural networks are artificial neural networks that are extensively used in speech recognition and natural language processing. RNNs act on the idea of preserving a layer's output and feeding it back into the input in terms of addressing the layer's output. Attention is a strategy in neural networks that mimics cognitive attention. The effect amplifies specific sections of the input data while detracting from others, with the idea that the network should pay greater attention to that tiny but significant portion of the data.

B. FACIAL DATABASES USED IN EMOTION RECOGNITION

The training of the neuron network with examples is one of the key performance indicators of deep learning, and several facial emotion recognition databases are now available to researchers to help them do so. Each one varies from the others interms of the number and size of images and videos, variations in illumination, population, and facial pose. Some of them are discussed below:

CK+: These are the most extensively used laboratory-controlled database for FER system assessment is the Extended (CK+) database. The CK+ dataset contains 593 video sequences from 123 different people, ranging in age from 18 to 50, gender, and ethnicity. Each image depicts a face transition from neural to a selected peak emotion, captured at 30 frames per second in 640x640 pixel resolution.

JAFFE: The Japanese Female Facial Expression database is a lab-controlled image library with 213 examples of posed emotions from ten Japanese women. Each individual possesses three to four images representing each of the six fundamental facial expressions, as well as one neural image. Because there are few samples per subject or expression, the database is difficult to utilize.

MMI: The MMI database contains 326 sequences from 32 people and is lab-controlled. A total of 213 sequences are labelled with six fundamental expressions, and 205 sequencesare caught in frontal view. Furthermore, MI has more difficult conditions, such as substantial inter-personal variances due to respondents' non-uniform expressions and the fact that many of them wear accessories.

KDEF: The laboratory-controlled Karolinska directed emotional faces (KDEF) database consists of images of 70 actors with five different angles labelled with six basic facial expressions plus neutral. In addition to these commonly used datasets, others that are suitable for training deep neural networks have emerged in the last two years.

Oulu-CASIA: There are 2,880 records in the Oulu-CASIA database. Only the final three peak frames and the first frame are usually used. From the 480 videos gathered by the VIS system. The 10-fold cross validation studies were then carried out using regular indoor lights.

FER2013. The Google image search API automatically collects FER 2013, a large-scale and unrestricted database, as shown in Fig 3. After rejecting incorrectly tagged frames and modifying the cropped region, all images were registered and resized to 48*48 pixels. FER2013 comprises 28,709 training images, 3,589 validation images, and 3,589 test images.

Fig 3. Sample images of different datasets.

III. DISCUSSION AND COMPARISON

Many of the existing research on facial emotion recognition has been deeply analysed in terms of architecture, dataset used and the recognition rate obtained. The comparison table illustrates the detailed review of previous research in Table I.

In one of the research, Evolutional Spatial-Temporal Networks is proposed, by using CK+, Oulu-CASIA, MMI databases. The obtained accuracy rate is 98.50%,86.25%,81.18%. In the research, using FER approach based on WMDNN and CK+, Oulu-CASIA, JAFFE databases are used. The average recognition accuracies obtained are 0.970, 0.922, and 0.923, respectively. In the next research, the method based on CNN and edge detection is proposed. Then scientifically mixed FER 2013 database with LRF dataset. Hence, it produces recognition rate of 88.56%. The training set is 1.5 times faster than the contrast algorithm in terms of training speed. In the research, CNN model is proposed. Then it is trained on FER2013 and JAFFE datasets are used. The obtained accuracies are 97.07% and 94.12% were obtained. In the next research, CNN model is used and then it is trained on FER2013 dataset and obtained 67% accuracy. As illustrated by the previous research, deep learning models are stronger and generate effective results. Finally, based on all of the datasets, it is estimated that FER 2013 achieves lower accuracy than others.

Table I. A comparison table for emotion recognition based on previous researches.

Author	Model	Dataset	Recognition Rate
K. Zhang et al. [1]	PHRNN & MSCNN	CK+, Oulu-CASIA, MMI	97.0%,92.02%, 92.3%
B. Yang et al. [2]	WMDNN	CK+, Oulu-CASIA, JAFFE	97.0%,92.2%, 92.3%
H. Zhang et al. [3]	CNN	FER2013	88.56%
G. Verma et al. [4]	CNN	JAFFE	94.12%
N. Zhou et al. [5]	CNN	FER2013	67%
D.H. Kim et al.[6]	CNN-LSTM	MMI, CASMEII	78.61%,60.98 %
Y.Li et al. [7]	ACNN	RAF-DB, AffectNet	80.54%,54.84 %
A.Agrawal et al. [8]	CNN	FER2013	65%
D.Liang et al.[9]	DCBiLSTM	CK+, Oulu-CASIA, MMI	80.71%
Z. Yu et al. [10]	STCNLSM	CK+, Oulu-CASIA, MMI	99.8%,93.5 %,84.53%
M.Mollahos seini et al. [11]	CNN	MMI, DISFA, FER2013	77.9%,55%, 61.1%
E. Pranav et al. [12]	DCNN	FER2013	78.04%
L.Liu et al. [13]	CNN	FER2013	49.8%

IV. CONCLUSION

Recognizing facial expressions is a difficult task. To this accomplish goal of recognition, variety of methodologies and procedures have been developed. This paper presented recent FER research, allowing us to keep up-to-date on the most recent discoveries in this field. We have discussed different architectures recently introduced by various researchers as well as various databases comprising spontaneous image collected in the real world and rest of them are created in laboratories, in orderto have and accomplish accurate emotion recognition. The convolutional neural network is employed by the majority of themin all of the above studies since it provides better accuracy. Even though they achieve better accuracy, some of them fail to extract multiple features. Therefore, we need a hybrid model for using attention-based vision transformers with transfer learning.

REFERENCES

[1] K. Zhang, Y. Huang, Y. Du and L. Wang, "Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks," *in IEEE Transactions on Image Processing, vol. 26, pp. 4193-4203, Sept. 2017.*

[2] B. Yang, J. Cao, R. Ni and Y. Zhang, "Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on Double-Channel Facial Images," *in IEEE Access, vol. 6, pp. 4630-4640, 2018.*

[3] H. Zhang, A. Jolfaei and M. Alazab, "A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing," *in IEEE Access, vol. 7, pp. 159081-159089, 2019.*

[4] Garima Verma, Hemraj Verma, "Hybrid Deep Learning Model for Emotion Recognition Using Facial Expressions", *Rev of Socionetwork Strat, vol.14,* pp. 171-180, 2020.

[5] N. Zhou, R. Liang and W. Shi, "A Lightweight Convolutional Neural Network for Real-Time Facial Expression Detection," *in IEEE Access, vol. 9, pp. 5573-5584, 2021.*

[6] D. H. Kim, W. J. Baddar, J. Jang, et Y. M. Ro, "Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition ", *IEEE Trans. Affect. Comput., vol. 10, pp. 223-236, avr. 2019.*

[7] Y. Li, J. Zeng, S. Shan, et X. Chen," Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism *", IEEE Trans. Image Process., vol. 28, no 5, pp. 2439-2450, 2019.*

[8] A. Agrawal et N. Mittal, "Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy ", *Vis. Comput., 2019.*

[9] D. Liang, H. Liang, Z. Yu, et Y. Zhang,"Deep convolutional BiLSTM fusion network for facial expression recognition ", *Vis. Comput., vol.36, pp. 499-508,2020.*

[10] Z. Yu, G. Liu, Q. Liu, et J. Deng, "Spatio-temporal convolutional features with nested LSTM for facial expression recognition ", *Neurocomputing, vol. 317, pp. 50-57, Nov. 2018.*

[11] M. Mohammadpour, H. Khaliliardali, S. M. R. Hashemi, et M. M. AlyanNezhadi, “Facial emotion recognition using deep convolutional networks “, *in 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), pp. 0017-0021,2017.*

[12] E. Pranav, S. Kamal, C. Satheesh Chandran and M. H. Supriya, "Facial Emotion Recognition Using Deep Convolutional Neural Network,"*6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 317-320, 2020.*

[13] Lu Lingling liu, "Human Face Expression Recognition Based on Deep Learning-Deep Convolutional Neural Network",*International Conference on Smart Grid and Electrical Automation (ICSGEA),2019.*