IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 1, Ser. 3 (Jan. – Feb. 2024), PP 62-72
www.iosrjournals.org

A Survey On Facial Emotion Recognition Using Convolutional Neural Network

T Shilpa[1], M Siddappa[2]
[1]*(Department Of Information Science And Engineering, Bangalore Institute Of Technology, India)*
[2]*(Department Of Computer Science And Engineering, Sri Siddhartha Institute Of Technology, India)*

Abstract:
There is a great demand for an efficient facial emotion recognition (FER) technique which can recognize different emotions from facial images. The emotional state of a person can be identified through different facial expressions and automatic recognition of emotions from facial expressions can be advantageous in tasks such as human-computer interaction and computer vision applications. There are several FER models discussed in existing works. This paper provides a brief analysis of FER techniques and emphasizes the implementation of the convolutional neural network (CNN) model for emotion recognition. The study suggests that CNN can overcome the drawbacks of conventional machine learning classifiers in terms of achieving better recognition accuracy and reduction in computational complexity. The study discusses different CNN architectures such as VGGNet, AlexNet, Inception, and ResNet, databases such as JAFFE, CK+, BU-3DFE, DISFA etc, and feature extraction techniques. The performance of CNN architectures is evaluated with respect to dataset and accuracy. In addition, this review provides a critique of existing research works done on FER, and highlights recommendations that outline a few concepts that need further investigation.

Keywords: *Facial Emotion Recognition, Convolutional Neural Networks, CNN architecture, Recognition accuracy, Performance Evaluation.*

Date of Submission: 09-02-2024 Date of Acceptance: 19-02-2024

I. Introduction

Facial emotion recognition (FER) is the process of identifying human emotion based on facial expressions[1]. Emotion recognition deals with the identification and classification of different emotions based on facial expressions[2]. Facial expressions define human feelings since it corresponds to the emotions. In general, facial expressions are the nonverbal way of communicating the emotions and it can be used as a potential tool to determine the emotional state of an individual[3]. Emotion recognition has gained huge attention in recent times since the features obtained from facial emotions can be used in a wide range of real time applications such as human computer interaction (HCI), E-learning and in the analysis of human behavior[4]. Recently, the emergence of artificial intelligence (AI) has enabled automated emotion recognition wherein the computer systems interact with humans to understand human emotions[5]. This interaction can help in providing personal counseling in medical domains[6]. However, in most of the cases, the AI enabled computer systems detect only static emotions and they fail to recognize the user's feelings in real cases. Hence it is important to explore the models which can perform real-time FER and capture feelings dynamically. In addition, it is challenging to identify different types of emotions just by looking at someone. There are six fundamental expressions namely, anger, disgust, fear, happiness, sadness and surprise and FER models should be capable of recognizing these emotions with high accuracy[7].

Recently, the application of machine learning (ML) and deep learning (DL) models are being used extensively in the FER process[8,9,10]. DL is a subset of ML processes which uses artificial neural networks for performing automated tasks using cognitive intelligence. Among different DL algorithms, this research emphasizes Convolutional Neural Networks (CNNs) for recognizing emotions from facial expressions[11]. CNN belongs to the class of deep neural networks which uses convolution based mathematical operations for performing a task. With the extensive application of CNN for FER in recent years, it is important to investigate the state of the art of CNN for FER, evaluate its performance and identify the challenges associated with it. This review presents a comprehensive analysis of the application of CNN for emotion recognition.

II. Research Significance

Several research works have discussed the important of FER. However, the continuous evolution of FER techniques demands a comprehensive analysis which can introduce latest research trends on FER. Few research articles have focused on the application of traditional approaches based on visual information and multi-modal information [12]. Recent investigations suggest that DL-based FER provides better accuracy compared to traditional FER techniques. However, DL models are characterized by their computational complexity and require additional resources such as graphic processing units (GPU) for achieving better performance. There is a need for a detailed investigation on different DL algorithms for FER. In this context, this paper presents a comprehensive analysis of FER techniques and helps the researchers to gain insight about the recent advancements in the field of FER. The key differences between the traditional FER models and CNN based FER are determined in terms of accuracy and other performance measures. Furthermore, this paper outlines the observations identified from existing literary works and highlights future research direction for FER.

III. Overview of the Related Works on FER

This research focuses on the CNN architecture which is applied to recognize different emotions from facial expressions. The current review analyzes different types of research articles including journal articles from reputed journals. The review considered articles that incorporate different datasets with an emphasis on implementing CNN for FER applications. A novel CNN approach for FER is proposed in[13]. The FER process used a two part CNN wherein the first part was used to remove the backgrounds from the images and in the second part the facial features were extracted. In this model, an expression vector was used to capture 5 different types of facial expressions. The performance of CNN was evaluated using the data of 10,000 images and CNN was able to attain a phenomenal 96% accuracy for FER. A weighted mixture deep neural network (WDNN) for automatic feature extraction for FER is implemented in[14]. The model was tested CK+, JAFFE, and Oulu-CASIA datasets which achieved an accuracy of 97%, 92.2%, and 92.3% respectively. Results show that there is a need to improve the accuracy of recognition. The authors in [15] trained the CNN model on a Field Programmable Gate Array (FPGA) for FER. The model was tested on the FER 2013 dataset for detecting emotions. The CNN model exhibited better results with fine-tuning and increased depth of the neural network. A FERC model based on CNN for FER is proposed in [16]. The model uses an expressional vector (EV) for detecting five different types of facial expressions. The study discussed the performance of the CNN based FERC model including datasets and their ability to provide solutions for inherent issues. The preliminary aim of this study is to enhance the accuracy of the CNN model which can provide a better insight for future research. The performance of the existing CNN, Decision tree, and feed forward neural network model was compared and results show that the CNN model outperforms existing model in terms of recognition accuracy. CNN architectures such as ResNet and MobileNet can be implemented to improve the performance.

IV. Facial Emotion Recognition

In general, FER models are categorized into two types based on their inputs i.e., input images and dynamic sequences. The analysis of the FER process involves face acquisition, image preprocessing, feature extraction and representation, feature classification, and emotion recognition.

Face Acquisition

Several techniques have been developed for detecting faces in real time scenarios[17,18]. It was observed that these techniques were only able to detect frontal view faces and underperform when used for detecting multi-view faces such as side views. With the advancements in FER technology, face acquisition is merged into a preprocessing phase wherein different preprocessing techniques are applied on the input images before the images are fed to the recognition models.

Image Preprocessing

The raw input images collected from the facial image dataset are subjected to preprocessing and during this stage, the input images are subjected to different operations such as denoising, resizing, and quality enhancement[19]. Preprocessing is carried out to remove the external noise, to prevent the influence of uncertainties during the emotion classification and recognition process. The images collected from the dataset are often distorted due to additive noise. In addition to the noise, few images may also incorporate complex backgrounds with poor light intensity, contrast, and occlusion, which affects the performance of the FER process. Hence, it is important to eliminate these interference factors before recognizing emotions[20].

The Steps involved in the preprocessing are as follows[21]:

Elimination of noise: The effect of noise in the input images can be eliminated using different processing filters such as Average Filter, Gaussian Filter, Median Filter, Adaptive Median Filter, and Bilateral Filter. These filters perform denoising which involves the removal of unnecessary noise from the input data for appropriate image processing and to obtain smooth image quality.

Face Detection: Face detection helps in detecting different face regions in different image frames. It processes the complex background information which might influence the recognition ability of FER techniques. Recently, face detection is performed as an independent process with an aim to localize and extract specific face regions.

Normalization: Normalization is performed to logically group the input images within the same range (usually the range is between 0 and 1). The size of the grayscale and color images are normalized in order to minimize the complexity of the FER process. Since the range of the input images vary randomly, it will have a negative impact on the recognition process. Hence, it is necessary to bring all the images under a common range via normalization and ensure the presence of important face features.

Image quality enhancement: The quality of the images are enhanced using histogram equalization techniques[22] such as Adaptive Histogram Equalization and Contrast Limited Adaptive Histogram Equalization (CLAHE) techniques. These techniques limit the amplification by preventing the contrast in the images.

Feature Extraction and Representation

The feature extraction module is used for extracting relevant features for the FER process. This is done to reduce the dimensionality of image data when most of the features are not contributing enough to the overall variance. Reducing unwanted and redundant features will reduce the computational time and improve the overall performance. In this stage, different types of features will be extracted using relevant feature extraction techniques. Most commonly, facial image features are categorized into two types namely geometric features and appearance features. Geometric features represent the location and shape of facial parts such as eyes, node, mouth and eyebrows for identifying facial expressions. These features consider muscle motion for determining facial expressions[23]. On the other hand, appearance features represent the textural changes and skin without considering the muscle motion. Techniques that use appearance features generate better results since they consider detailed image information such as color intensity, texture edges, pixel intensity, and wrinkles of the face. Some of the important techniques that use appearance features are local binary pattern (LBP), Haralick feature extraction, Gabor feature extraction, ROI based feature extraction, and Histogram of Oriented Gradients (HoG) descriptor[24,25,26,27].

Local binary pattern (LBP) feature extraction:

The LBP model identifies the level of intensity and neighborhood values in an image. The binary pattern in an image is computed by comparing the actual binary pattern with neighboring pixels. Based on the difference, variations in the image patterns are recorded[28]. Furthermore, the facial images are reconstructed based on the histograms obtained from all samples. In this process, the input image is divided into multiple smaller blocks and for each pixel in the block, a bit pattern is obtained using the steps discussed in below points:

- Binary bit patterns are calculated by obtaining eight adjacent pixels positioned either in clockwise or anticlockwise directions.
- The value of the center pixel is determined and if the value of the center pixel is less than the surrounding adjacent pixel, then the assigned value is ‘0’, else the assigned value will be ‘1’.
- The bit pattern in the form of binary string is obtained and the number of each bit pattern is calculated.
- Only uniform bit patterns are considered for further process and are normalized.
- Further, all the uniform bit patterns are arranged and the number of occurrences in the block for each column is calculated. Lastly, all columns of all blocks are added to form a bit pattern matrix, which is constituted as the image feature.

Haralick feature extraction:

In this process, the features are extracted using a gray level co-occurrence matrix (GLCM). The GLCM matrix evaluates the co-occurrence of the adjacent gray levels in the images. The haralick feature extraction helps in determining the texture of the image in terms of different parameters such as contrast, correlation, sum of squares, sum of average, homogeneity etc.

ROI based feature extraction:

In ROI based technique, the features from the contrast enhanced images are extracted using their region of interest (ROI). The ROI helps in selecting an appropriate area for recognizing and classifying facial expressions. In FER processes, the ROI of a facial image is extracted directly from the normalized feature point area.to extract appropriate features.

Gabor feature extraction:

Gabor feature extraction technique is based on the principles of fourier transform which integrates the wallet theory with the Gabor feature. FER based on Gabor feature extraction yields excellent results when implemented with other classification techniques[29]. For instance, Gabor wavelets with Discrete Wavelet

Transform (DWT) generate more detailed feature vectors compared to Gabor filters alone, to overcome the issue of data dimensionality. This technique is more robust to illumination intensity and is most suitable for extracting multi-scale and multi-directional texture features.

Histogram of Oriented Gradients (HoG) descriptor:

The HoG based feature extraction is used widely in computer vision applications for image classification and detection processes. This technique evaluates the occurrences of gradient orientation in the localized portion of an image. For each image sample, the HoG generates histograms based on the intensity and orientation of the HoG descriptor.

Studies that used different feature extraction techniques for FER and its strength and limitations are tabulated in Table 1 and Table 2 respectively.

Table 1. Feature extraction-based FER studies

Reference	Preprocessing technique	Feature extraction method	Accuracy of FER	Observations
[30]	Face detection	LDHRP, LDSP	96.25 %	The proposed feature extraction technique is tolerant against variations in the illumination intensity
[31]	Face detection	DCT, GF	97.10%	New features of face images are extracted using DCT and GF which achieved better recognition and classification accuracy compared to other methods
[32]	Cropping and Normalization	HOG, DWT	75%	Accuracy of the FER process can be improved by incorporating learned and engineered features
[33]	Noise removal using Gaussian filters	HOG, Haralick, GF, SBDP	94.11%	The extracted features helps in achieving a high recognition rate with a very minimum error rate
[34]	Cropping and Image quality enhancement	LBP	99.12%	The proposed approach achieves excellent texture recognition performance with less complexity and better efficiency
[35]	Face detection	HOG, LBP	97.66%	The HOG and LBP based approach is suitable for identifying neutral expressions and is capable of reducing dimensionality
[36]	Cropping and Normalization	GoF	89.41%	The proposed model extracts features of facial ROIs and requires less computational resources

LDHRP → Local directional rank histogram pattern, LDSP → Local directional strength pattern, DCT → Discrete cosine transform, GF → Gabor filter, DWT → Discrete wavelet transform, HOG → Histogram of Oriented Gradients, GoF → Gabor orientation filters, SBDP → Square-Based Diagonal Pattern

Table 2. Overview of different feature extraction techniques

Feature	Variants	Strength	Limitation
LBP	Uniform LBP, Rotated LBP, Complete LBP and rotation variant LBP	Less computational complexity, robust to grayscale variations	Extracts limited image formation and is highly affected by image rotation
GF	Log polar GF, and Gabor wavelets	High spatial frequency, and captures smaller variations in image attributes	Susceptibility to data dimensionality issues
HOG	Circular HOG, rectangular HOG	Provides large scale global information, tolerant to photometric transformation and illumination changes	Computation time for extracting complex features increases with the growing HOG descriptor vectors

Hybrid Features	Different types of geometric features	Features are highly correlated	More prone towards computational complexity
Learned Features	Combination of neural networks	Efficient descriptors	Requires more number of computational resources while processing large scale image data

As inferred from Table 1, feature extraction plays a significant role in maximizing the recognition performance with reduced complexity. The LBP and HOG features achieve an excellent recognition accuracy of 97.66 and LBP alone achieves 99.12% accuracy. In addition to LBP and HOG, the DCT and GF also exhibit better performance in terms of capturing new features from the facial images. When combined with classification techniques, the HOG, LBP, DCT and GF filters achieve phenomenal performance. Some of the prominent classifiers used in the FER are discussed in the next section.

Feature Classification for Emotion Recognition

After extracting a set of salient features, the features are classified with the aim of selecting only the most discriminative features to recognize different types of emotions. Several considerable techniques have been proposed over a period of time for classification of features for FER. Some of the prominent classifiers are: Support Vector Machines (SVM), Naive Bayes classifier, Random Forest (RF), Decision trees (DT), Artificial Neural Networks (ANN) and others.

SVM classifier: SVM is a supervised ML classifier used for performing classification tasks. SVM will identify the data patterns and use information from the previous process for classification[37]. In general, SVMs overcome the problem of overfitting and achieve precise classification in various cases. SVM possesses high generalization, slow convergence speed and is highly sensitive to local extrema.

NB classifier: The NB classifier belong to the family of fundamental probabilistic classifiers which employ a Bayes' theorem with robust naive assumptions between the features. The NB classifiers are one of the simplest network models whose assumptions are independent[38]. This property of the classifier allows easy training of the model without using any previous data. These classifiers are incorporated with kernel density estimation which improves the classification accuracy. The Naive Bayes classifiers possess high scalability and require fewer parameters for learning a problem. This will reduce the complexity of the network and minimize the computational burden on the network layers.

RF classifier: The RF classifier is a supervised classification algorithm which is suitable for both classification and regression tasks. To classify the input features, the RF algorithm creates the forest with a number of trees[39]. More the number of trees, more is the accuracy of the RF classifier.

Decision Tree (DT) classifier: Decision tree algorithm is an advanced supervised ML algorithms used for performing both classification and regression[40]. Decision trees are aimed to construct a mode which can estimate the target variable by creating certain decision based rules. These rules are derived from the extracted features. The mechanism involved in the decision tree algorithm is highly inductive and is used widely in image classification applications.

Artificial Neural Networks (ANN): An ANN model is one of the effective neural network algorithms whose architecture is similar to that of the biological nervous system as the brain processes the received data[41]. ANNs show superior performance in solving various complex nonlinear problems with superior accuracy and better precision. One of the prominent aspects of the ANN algorithm is the advanced and sophisticated architecture of the data analyzing system. ANNs aggregate their knowledge by identifying the data patterns and their correlation with other data elements and learn the behavior of the data through experience. ANNs are considered to be condescending when compared to similar classifiers mainly because of its versatility, robustness towards faults, high accuracy, well-structured architecture, and scalability.

Deep Belief Networks (DBN): DBN is designed using unsupervised Restricted Boltzmann Machines (RBM) and Belief networks wherein each layer constitutes a RBM which is arranged to form DBN. For classification, the DBN layers are trained to learn the features from the input facial images in order to maximize the recognition capability of the FER models[42].

Long Short Term Memory (LSTM): The LSTM model is structured to analyze the chronological sequences and their long-range dependencies with better accuracy than conventional RNNs. LSTMs are a special type of RNNs which can overcome the long-term dependency problem by remembering the information for a longer duration. Due to their superior memory, LSTMs are extensively used to perform various explicit functions such as classification and prediction[43].

Generative Adversarial Networks (GAN): GAN is an unsupervised ML algorithm which uses a supervised loss as part of the training process. For an input training dataset, GAN will automatically train itself to generate new data without altering the statistics as given in the training set. This will significantly minimize the computational complexities. DCGAN will be employed for training the system that can synthesize the raw image data.

The classification techniques incorporated in various existing studies are summarized in Table 3.

Table 3. Existing classification techniques for FER

Reference	Classification model	Observations	Recognition Accuracy	Limitations
[44]	SVM	The FER model extracts geometric features for identifying facial muscle movements. The SVM model is optimized using genetic algorithm to achieve optimal FER accuracy	96.29%	The SVM based FER model is not tested for classifying frontal and side views of human faces from 3D facial images
[45]	RF, SVM, K-Nearest Neighbor (KNN), and Multi-Layer Perceptron (MLP)	The performance of different machine learning based classifier is analyzed for FER, by incorporating a feature ranking based approach	KNN - 94.93% RF - 93.95% MLP - 89.89% SVM - 89.43%	The classifiers are tested for a limited number of data features and the classifiers were able to extract only a fewer number of features
[46]	LSTM	An optimized LSTM model is applied for classifying different emotions	77.68%	The LSTM model suffered from the problem of premature convergence which significantly affected the classification performance
[47]	LSTM-CNN	The LSTM is implemented to address the occurrence of gradient disappearance and explosion, which occurs during the training of deep learning model	88.3%	The hybrid LSTM-CNN model achieves satisfactory performance in terms of recognition accuracy and can be improved
[48]	LSTM-CNN	The advantages of LSTM and CNN are leveraged to identify different facial expressions and emotions from speech text	92.00%	The performance of the LSTM-CNN model is not robust enough to background noise, which affected the recognition accuracy
[49]	RF	The emotion from speech text is recognized using a Fourier transform method combined with RF classifier	89.60%	The lack of natural emotional speech dataset limits the performance of RF classifier to classify integrated speech and facial emotions

[50]	DT, SVM, RF, CNN	The performance of the FER is tested using different classifiers to categorize each emotion to a particular class	DT = 52%
			SVM = 73%
			RF = 65%
			CNN = 78.20%
				The loss of the CNN model is higher due to larger number of training parameters

[51]	DBN	The DBN classifier is trained to extract high level features and multimodal expressions from facial image dataset	90.89%	The performance of the fusion model needs to be improved in terms of analyzing relative attributes
[52]	GAN	A feature improving GAN is implemented for face frontalization in order to enhance the recognition performance for identifying large face poses	98.3%	The GAN model underperforms to recognize emotions from face images with large pitch angle

The performance of different classifiers in terms of recognition accuracy are summarized in Table 3. It can be inferred that GAN, SVM and LSTM exhibit superior classification performance compared to other classifiers. Despite the phenomenal performance in terms of achieving excellent recognition performance, the performance of the existing classifiers is affected in terms of intensity variation, changes in illumination, poor resolution, occlusion, difficulty in identifying instant facial expressions, etc. In addition, the efficiency of the classifiers is also affected due to high computational complexity, increased detection time and space complexity. These drawbacks motivate the researchers to focus on improving the classification of FER processes that lead to the implementation of CNN models.

V. CNN for Facial Emotion Recognition

Recently, the application of convolutional neural networks (CNN) has been drawing huge attention among various researchers in the field of image classification and to enhance the efficiency and accuracy of the FER process[53]. The main advantages of using CNN for classification is the best use of the shared weight of the conventional layers and their efficiency in image recognition and in the classification process. CNNs can extract relevant features from facial images and make effective use of the image filters which is a common phenomenon in most of the image processing and classification applications. CNNs basically employs very little preprocessing of images when compared to other image classification techniques. This shows that the techniques used in basic networks other than CNN which are responsible for interpreting the filters were not able to train themselves and possessed reduced performance ability[54]. CNNs are the best solution to overcome the limitation of these traditional algorithms hence are more often implemented in image recognition applications. The basic architecture of CNN is illustrated in Figure 1.

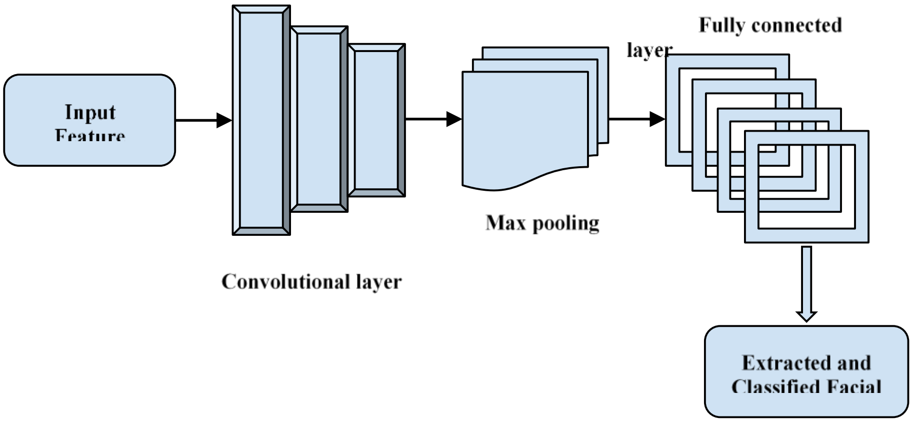

Figure 1. CNN Architecture

For classification, an input image from the raw image dataset is given to the CNN model and create a feature map using different filters. Each CNN layer can generate representative features including low-level features for achieving better classification performance [55]. The architecture of CNN consists of three main layers namely; convolution layer, max pooling layer, and a fully connected (FC) layer as shown in figure 2. Among them, the convolution and max pooling layers are used for feature extraction and the FC layer combined with a softmax classifier is used for classification. The description of the CNN layers are as follows:

Convolutional layer: It is the fundamental layer which is considered as the building block of the architecture. This layer captures all important salient features from facial images such as edges, color, gradient orientation, etc. Small sized filters that are embedded in this layer extracts information about the spatial location of the image and generates an output matrix for further processing.

Max pooling layer: This is the second layer in the CNN architecture which reduces the dimension of the feature maps while maintaining the same image depth. This layer is responsible for reducing overfitting in the model by minimizing the network parameters and computational time.

Fully connected (FC) layer: The FC layer connects all the previous CNN layers using activation functions. It generates high-level features and selects features that are highly correlated. In addition, a softmax layer in the CNN to predict the output based on the features obtained from the FC layer. This is done by determining the probability of each image sample and the sample that has the highest probability is generated as the classification output.

Network architectures of CNN for FER

The architecture of each CNN differs in terms of depth, composition of layers, and number of parameters. Most of the CNN models are shallow since they lack proper resources for selecting suitable architectures. The main reason for this is not known [56]. Several CNN architectures have been introduced to overcome this bottleneck. This section discusses some of the prominent CNN architectures for FER.

VGGNet: The VGGNet architecture can process large input images of pixel size 224 x 224 and has 4096 convolutional features. The VGGNet - CNN model is computationally efficient and is used in several computer vision applications such as image processing, object detection, and classification tasks. However, training VGGNet with large filters is difficult since it requires a large amount of data. In addition, it is computationally intensive and expensive to implement VGGNet for image classification tasks where the size of the input images range from 100 x 100 pixel to 350 x 350 pixels.

Inception: The architecture of inception-CNN model is similar to GoogLeNet but is characterized with robust neural network structure without pooling of strided convolutions. The inception model incorporates various convolutional kernels inside a single block. This model is capable of extracting different features from different environments and then it concatenates all the features.

AlexNet: AlexNet is trained from the first layer of the CNN by using arbitrary adjustments of weights which requires more time. In this study, AlexNet is trained by learning the features of CNN which is quite faster and requires a relatively smaller number of training samples. In AlexNet there are 5 convolutional layers with 3 FC layers and each layer is provided with convolution, Rectified Linear Unit (ReLU) pooling feature identification layers and each neuron in the AlexNet is connected with each other by means of FC layers which allows the generation of classification output using softmax function.

ResNet: ResNet is a form of deep convolutional network (DCNN) developed by Microsoft. ResNet consists of a framework which trains numerous layers (up to thousands of layers) without affecting the performance of the network. The robust representation capacity of the ResNet enhances the performance of object detection and face recognition. The ResNet architecture consists of 152 layers with more than 1 million parameters. Hence, it is considered a deep network for CNNs.

The performance of different CNN network architectures for the FER process is summarized in Table 4.

Table 4. CNN architectures for FER

Reference	CNN Architecture	Architecture Description	Dataset used	FER Accuracy
[57]	Shallow CNN (SHCNN)	A shallow CNN architecture is designed with three layers for classifying static and micro-expressions simultaneously without requiring large datasets for training CNN.	FER2013, FERPlus, CASME, CASME II, and SAMM datasets	FER 2013 - 69.10% FERPlus - 86.54% CASME - 63.33% CASME II - 69.81% SAMM - 86.47%
[58]	Deep CNN	The DCNN model is developed using 5 convolutional layer, 3 max pooling, 1 average pooling and an up sampling layer	FER2013, JAFFE, CK+, KDEF, and RAF	FER2013- 78%, JAFFE - 98%, CK+ - 98%, KDEF - 96%, RAF - 83%
[59]	GoogLeNet	The proposed multi-task cascaded CNN model consists of 4 residual depth-wise separable convolutions activated using a ReLU activation function	FER-2013	Training accuracy = 71% and Testing accuracy = 67%
[60]	Racial Identity Network (RI-Net)	The RI-Net model is constructed using 5 convolutional layer and a Softmax classifier to classify facial expressions	CK+ dataset	Accuracy = 100% and 92% while recognizing happy and angry emotions respectively
[61]	ResNet	A Light-CNN model is developed using six depth wise separable residual convolution modules to overcome the shortcoming of shallow CNN model in terms of overfitting	CK+ dataset, multi-view BU-3DEF, and FER2013 dataset	CK+ dataset - 92.86% , BU-3DEF - 86.20% FER2013 - 68%
[62]	VGG-Net	A VGG-Net based DCNN model is designed using a deeper architecture composed of a 3x3 small convolution kernel and a 2 x 2 small pool kernel	FER2013 dataset	73.06%
[63]	ResNet	Two ResNet architectures namely ResNet-12 and ResNet-18 are implemented with reduced number of parameters	CK+ dataset	97.56%
[64]	AlexNet	A pre-trained AlexNet architecture is designed and the model is fine-tuned and trained using a Imagenet dataset	CK+ dataset, FER dataset	CK+ dataset - 99.44% FER dataset -

				70.52%
[65]	Inception and VGG	The inception and VGG model are trained using a ImageNet database with a feature length of 4096 and image size of 224 x 224	CK+, JAFFE and FACES	For VGG16 model: CK+ - 88.27%, JAFFE - 81.33% FACES - 95.25% For VGG-Face CK+ - 91.37%, JAFFE - 86.67% FACES - 95.06%

As inferred from Table 4, different types of CNN architectures are implemented for achieving better FER performance. Compared to different models, ResNet and AlexNet exhibit superior FER accuracy and use a lesser number of parameters for training the CNN model. The second best model is the VGGNet model which is suitable for designing a deeper network architecture.

Database for FER

This section discusses prominent facial image datasets for FER:

Japanese Female Facial Expressions (JAFFE) dataset: The JAFFE dataset incorporates 7 different facial emotions with six basic and one neutral emotion depicted using 213 images[66].

Extended Cohn–Kanade (CK+) dataset: This dataset is an extended version of CK dataset consisting of 7 facial emotions collected from 593 video sequences. The average age group of the participants is 18 to 30 and includes multiple races.

Compound Emotion (CE) dataset: The dataset consists of 5060 images with facial expressions collected from 230 participants. The facial occlusion is reduced since the dataset contains images of individuals without glasses. The dataset consists of coloured images with a resolution of 3000 x 4000 pixels.

Denver Intensity of Spontaneous Facial Action Dataset (DISFA): The DISFA dataset incorporates facial emotion data aggregated from 27 individuals and 130,000 video sequences. The resolution of the input images is 1024 x 768 pixels and the intensity of the action unit is varied from 0-5 scale.

MMI Facial Expression Dataset: The MMI dataset consists of high resolution images collected from 2900 video samples and 75 participants. Each AU in the videos are annotated for yielding better resolution.

Binghamton University 3D Facial Expression (BU-3DFE): This dataset is created mainly for FER from 3D images in order to understand human behavior. An overall 100 participants with different ethnicities. Six emotions are included in the dataset with a set of manually annotated facial expressions. The resolution of the images is 1040 x 1329 pixels.

Large MPI Facial Expression (MPI) dataset: The MPI dataset is an experimentally tested database consisting of both emotional and conversational facial expressions. The dataset consists of 55 expressions collected from 19 German individuals with 9 males and 10 females. The facial expressions are obtained from different shooting angles

VI. Conclusion

FER has gained significant attention in recent times. The research related to FER models has attracted various researchers and as a result several FER models have been introduced in the past decades. This review emphasizes the implementation of CNN model for FER processes. The study discusses a comprehensive analysis of the state of the art of different models adopted for recognizing different facial emotions and provides a brief overview of the existing studies on FER. Different stages involved in the FER processes such as image acquisition, feature extraction, classification and emotion recognition are discussed with an emphasis on the role of different feature extraction techniques and classifiers in improving the accuracy of the FER process. Different CNN architectures and datasets used in the FER tasks are reviewed and the observations in terms of accuracy are tabulated. In addition, different performance evaluation metrics used for evaluating the efficacy of the CNN model are outlined. Lastly, the study outlines the observations and suggests some of the future directions to improve the FER process. This review article aims to assist the researchers carrying out their research in this domain and intends to contribute to the existing works conducted in this field. Some of the prominent observations are as follows:

- Although the CNN model requires a larger dataset for training, it is highly efficient in enhancing the FER performance because of its superior classification ability.
- It is essential to perform preprocessing before classification and emotion recognition in order to achieve better FER accuracy and minimize the complexity of the CNN model.

- It was observed from existing works that CNN requires a large volume of data in case the dataset lacks sufficient training samples. In such cases, it is suggested to perform data augmentation which can increase the size of the dataset by performing certain predefined actions such as rescaling, rotating, zooming and horizontal and vertical flipping.

Compared to shallow CNN models, Deep CNN is highly efficient in solving problems related to FER. However, the increase in the depth of the conventional CNN architecture does not ensure the increase in the recognition accuracy.

For future research, there is a need to validate and benchmark the implementation of private datasets for FER and deeper research is required to integrate CNN models with other deep learning models such as LSTM, and GAN and test more ensemble models for FER.

References

[1]. Khaireddin, Y., & Chen, Z. (2021). Facial Emotion Recognition: State Of The Art Performance On Fer2013. Arxiv Preprint Arxiv:2105.03588.

[2]. Hu, M., Wang, H., Wang, X., Yang, J., & Wang, R. (2019). Video Facial Emotion Recognition Based On Local Enhanced Motion History Image And Cnn-Ctslstm Networks. Journal Of Visual Communication And Image Representation, 59, 176-185.

[3]. Lasri, I., Solh, A. R., & El Belkacemi, M. (2019, October). Facial Emotion Recognition Of Students Using Convolutional Neural Network. In 2019 Third International Conference On Intelligent Computing In Data Sciences (Icds) (Pp. 1-6). Ieee.

[4]. Jain, D. K., Shamsolmoali, P., & Sehdev, P. (2019). Extended Deep Neural Network For Facial Emotion Recognition. Pattern Recognition Letters, 120, 69-74.

[5]. Gervasi, O., Franzoni, V., Riganelli, M., & Tasso, S. (2019, January). Automating Facial Emotion Recognition. In Web Intelligence (Vol. 17, No. 1, Pp. 17-27). Ios Press.

[6]. Flynn, M., Effraimidis, D., Angelopoulou, A., Kapetanios, E., Williams, D., Hemanth, J., & Towell, T. (2020). Assessing The Effectiveness Of Automated Emotion Recognition In Adults And Children For Clinical Investigation. Frontiers In Human Neuroscience, 14, 70.

[7]. Zadeh, M. M. T., Imani, M., & Majidi, B. (2019, February). Fast Facial Emotion Recognition Using Convolutional Neural Networks And Gabor Filters. In 2019 5th Conference On Knowledge Based Engineering And Innovation (Kbei) (Pp. 577-581). Ieee.

[8]. Georgescu, M. I., Ionescu, R. T., & Popescu, M. (2019). Local Learning With Deep And Handcrafted Features For Facial Expression Recognition. Ieee Access, 7, 64827-64836.

[9]. Hassouneh, A., Mutawa, A. M., & Murugappan, M. (2020). Development Of A Real-Time Emotion Recognition System Using Facial Expressions And Eeg Based On Machine Learning And Deep Neural Network Methods. Informatics In Medicine Unlocked, 20, 100372.

[10]. Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2021). Deep Learning-Based Facial Emotion Recognition For Human–Computer Interaction Applications. Neural Computing And Applications, 1-18.

[11]. Mohammadpour, M., Khaliliardali, H., Hashemi, S. M. R., & Alyannezhadi, M. M. (2017, December). Facial Emotion Recognition Using Deep Convolutional Networks. In 2017 Ieee 4th International Conference On Knowledge-Based Engineering And Innovation (Kbei) (Pp. 0017-0021). Ieee.

[12]. Ko, B. C. (2018). A Brief Review Of Facial Emotion Recognition Based On Visual Information. Sensors, 18(2), 401.

[13]. Mehendale, N. (2020). Facial Emotion Recognition Using Convolutional Neural Networks (Ferc). Sn Applied Sciences, 2(3), 1-8.

[14]. Yang, B., Cao, J., Ni, R., & Zhang, Y. (2017). Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based On Double-Channel Facial Images. Ieee Access, 6, 4630-4640.

[15]. Phan-Xuan, H., Le-Tien, T., & Nguyen-Tan, S. (2019). Fpga Platform Applied For Facial Expression Recognition System Using Convolutional Neural Networks. Procedia Computer Science, 151, 651-658.

[16]. Sarvakar, K., Senkamalavalli, R., Raghavendra, S., Kumar, J. S., Manjunath, R., & Jaiswal, S. (2021). Facial Emotion Recognition Using Convolutional Neural Networks. Materials Today: Proceedings.

[17]. Meena, D., & Sharan, R. (2016, December). An Approach To Face Detection And Recognition. In 2016 International Conference On Recent Advances And Innovations In Engineering (Icraie) (Pp. 1-6). Ieee.

[18]. Khan, M., Chakraborty, S., Astya, R., & Khepra, S. (2019, October). Face Detection And Recognition Using Opencv. In 2019 International Conference On Computing, Communication, And Intelligent Systems (Icccis) (Pp. 116-119). Ieee.

[19]. Mellouk, W., & Handouzi, W. (2020). Facial Emotion Recognition Using Deep Learning: Review And Insights. Procedia Computer Science, 175, 689-694.

[20]. Li, S., & Deng, W. (2020). Deep Facial Expression Recognition: A Survey. Ieee Transactions On Affective Computing.

[21]. Huang, Y., Chen, F., Lv, S., & Wang, X. (2019). Facial Expression Recognition: A Survey. Symmetry, 11(10), 1189.

[22]. Mungra, D., Agrawal, A., Sharma, P., Tanwar, S., & Obaidat, M. S. (2020). Pratit: A Cnn-Based Emotion Recognition System Using Histogram Equalization And Data Augmentation. Multimedia Tools And Applications, 79(3), 2285-2307.

[23]. Meriem, S. A. R. İ., Moussaoui, A., & Hadid, A. (2020). Automated Facial Expression Recognition Using Deep Learning Techniques: An Overview. International Journal Of Informatics And Applied Mathematics, 3(1), 39-53.

[24]. Kim, J. H., Kim, B. G., Roy, P. P., & Jeong, D. M. (2019). Efficient Facial Expression Recognition Algorithm Based On Hierarchical Deep Neural Network Structure. Ieee Access, 7, 41273-41285.

[25]. Sun, X., Zheng, S., & Fu, H. (2020). Roi-Attention Vectorized Cnn Model For Static Facial Expression Recognition. Ieee Access, 8, 7183-7194.

[26]. Hu, M., Yang, C., Zheng, Y., Wang, X., He, L., & Ren, F. (2019). Facial Expression Recognition Based On Fusion Features Of Center-Symmetric Local Signal Magnitude Pattern. Ieee Access, 7, 118435-118445.

[27]. Nazir, M., Jan, Z., & Sajjad, M. (2018). Facial Expression Recognition Using Histogram Of Oriented Gradients Based Transformed Features. Cluster Computing, 21(1), 539-548.

[28]. Kola, D. G. R., & Samayamantula, S. K. (2021). A Novel Approach For Facial Expression Recognition Using Local Binary Pattern With Adaptive Window. Multimedia Tools And Applications, 80(2), 2243-2262.

[29]. Mattela, G., & Gupta, S. K. (2018, February). Facial Expression Recognition Using Gabor-Mean-Dwt Feature Extraction Technique. In 2018 5th International Conference On Signal Processing And Integrated Networks (Spin) (Pp. 575-580). Ieee.

[30]. Uddin, M. Z., Khaksar, W., & Torresen, J. (2017). Facial Expression Recognition Using Salient Features And Convolutional Neural Network. Ieee Access, 5, 26146-26161.

[31]. Tsai, H. H., & Chang, Y. C. (2018). Facial Expression Recognition Using A Combination Of Multiple Facial Features And Support Vector Machine. Soft Computing, 22(13), 4389-4405.

[32]. Nigam, S., Singh, R., & Misra, A. K. (2018). Efficient Facial Expression Recognition Using Histogram Of Oriented Gradients In Wavelet Domain. Multimedia Tools And Applications, 77(21), 28725-28747.

[33]. Jain, N., Kumar, S., & Kumar, A. (2019). Effective Approach For Facial Expression Recognition Using Hybrid Square-Based Diagonal Pattern Geometric Model. Multimedia Tools And Applications, 78(20), 29555-29571.

[34]. Yasmin, S., Pathan, R. K., Biswas, M., Khandaker, M. U., & Faruque, M. R. I. (2020). Development Of A Robust Multi-Scale Featured Local Binary Pattern For Improved Facial Expression Recognition. Sensors, 20(18), 5391.

[35]. Lakshmi, D., & Ponnusamy, R. (2021). Facial Emotion Recognition Using Modified Hog And Lbp Features With Deep Stacked Autoencoders. Microprocessors And Microsystems, 82, 103834.

[36]. Jiang, P., Wan, B., Wang, Q., & Wu, J. (2020). Fast And Efficient Facial Expression Recognition Using A Gabor Convolutional Network. Ieee Signal Processing Letters, 27, 1954-1958.

[37]. Zhang, Y. D., Yang, Z. J., Lu, H. M., Zhou, X. X., Phillips, P., Liu, Q. M., & Wang, S. H. (2016). Facial Emotion Recognition Based On Biorthogonal Wavelet Entropy, Fuzzy Support Vector Machine, And Stratified Cross Validation. Ieee Access, 4, 8375-8385.

[38]. Kumar, P., Roy, P. P., & Dogra, D. P. (2018). Independent Bayesian Classifier Combination Based Sign Language Recognition Using Facial Expression. Information Sciences, 428, 30-48.

[39]. Yang, B., Cao, J. M., Jiang, D. P., & Lv, J. D. (2018). Facial Expression Recognition Based On Dual-Feature Fusion And Improved Random Forest Classifier. Multimedia Tools And Applications, 77(16), 20477-20499.

[40]. Saravanan, A., Perichetla, G., & Gayathri, D. K. (2019). Facial Emotion Recognition Using Convolutional Neural Networks. Arxiv Preprint Arxiv:1910.05602.

[41]. Liyakat, K. K. S., Mane, V. A., Paradeshi, K. P., Kadam, D. B., & Pandyaji, K. K. (2022). Development Of Pose Invariant Face Recognition Method Based On Pca And Artificial Neural Network. Journal Of Algebraic Statistics, 13(3), 3676-3684.

[42]. Hassan, M. M., Alam, M. G. R., Uddin, M. Z., Huda, S., Almogren, A., & Fortino, G. (2019). Human Emotion Recognition Using Deep Belief Network Architecture. Information Fusion, 51, 10-18.

[43]. Yang, J., Huang, X., Wu, H., & Yang, X. (2020). Eeg-Based Emotion Classification Based On Bidirectional Long Short-Term Memory Network. Procedia Computer Science, 174, 491-504.

[44]. Liu, X., Cheng, X., & Lee, K. (2020). Ga-Svm-Based Facial Emotion Recognition Using Facial Geometric Features. Ieee Sensors Journal, 21(10), 11532-11542.

[45]. Abdulrazaq, M. B., Mahmood, M. R., Zeebaree, S. R., Abdulwahab, M. H., Zebari, R. R., & Sallow, A. B. (2021, February). An Analytical Appraisal For Supervised Classifiers' Performance On Facial Expression Recognition Based On Relief-F Feature Selection. In Journal Of Physics: Conference Series (Vol. 1804, No. 1, P. 012055). Iop Publishing.

[46]. Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., & Chandran, V. (2018). Long Short Term Memory Hyperparameter Optimization For A Neural Network Based Emotion Recognition Framework. Ieee Access, 6, 49325-49338.

[47]. An, F., & Liu, Z. (2020). Facial Expression Recognition Algorithm Based On Parameter Adaptive Initialization Of Cnn And Lstm. The Visual Computer, 36(3), 483-498.

[48]. Wang, X., Chen, X., & Cao, C. (2020). Human Emotion Recognition By Optimally Fusing Facial Expression And Speech Feature. Signal Processing: Image Communication, 84, 115831.

[49]. Hamsa, S., Shahin, I., Iraqi, Y., & Werghi, N. (2020). Emotion Recognition From Speech Using Wavelet Packet Transform Cochlear Filter Bank And Random Forest Classifier. Ieee Access, 8, 96994-97006.

[50]. Christy, A., Vaithyasubramanian, S., Jesudoss, A., & Praveena, M. D. (2020). Multimodal Speech Emotion Recognition And Classification Using Convolutional Neural Network Techniques. International Journal Of Speech Technology, 23(2), 381-388.

[51]. Liu, D., Chen, L., Wang, Z., & Diao, G. (2021). Speech Expression Multimodal Emotion Recognition Based On Deep Belief Network. Journal Of Grid Computing, 19(2), 1-13.

[52]. Rong, C., Zhang, X., & Lin, Y. (2020). Feature-Improving Generative Adversarial Network For Face Frontalization. Ieee Access, 8, 68842-68851.

[53]. Xie, S., & Hu, H. (2018). Facial Expression Recognition Using Hierarchical Features With Deep Comprehensive Multipatches Aggregation Convolutional Neural Networks. Ieee Transactions On Multimedia, 21(1), 211-220.

[54]. Lopes, A. T., De Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial Expression Recognition With Convolutional Neural Networks: Coping With Few Data And The Training Sample Order. Pattern Recognition, 61, 610-628.

[55]. Said, Y., & Barr, M. (2021). Human Emotion Recognition Based On Facial Expressions Via Deep Learning On High-Resolution Images. Multimedia Tools And Applications, 80(16), 25241-25253.

[56]. Pramerdorfer, C., & Kampel, M. (2016). Facial Expression Recognition Using Convolutional Neural Networks: State Of The Art. Arxiv Preprint Arxiv:1612.02903.

[57]. Miao, S., Xu, H., Han, Z., & Zhu, Y. (2019). Recognizing Facial Expressions Using A Shallow Convolutional Neural Network. Ieee Access, 7, 78000-78011.

[58]. Mohan, K., Seal, A., Krejcar, O., & Yazidi, A. (2020). Facial Expression Recognition Using Local Gravitational Force Descriptor-Based Deep Convolution Neural Networks. Ieee Transactions On Instrumentation And Measurement, 70, 1-12.

[59]. Zhou, N., Liang, R., & Shi, W. (2020). A Lightweight Convolutional Neural Network For Real-Time Facial Expression Detection. Ieee Access, 9, 5573-5584.

[60]. Sohail, M., Ali, G., Rashid, J., Ahmad, I., Almotiri, S. H., Alghamdi, M. A., ... & Masood, K. (2021). Racial Identity-Aware Facial Expression Recognition Using Deep Convolutional Neural Networks. Applied Sciences, 12(1), 88.

[61]. Shao, J., & Qian, Y. (2019). Three Convolutional Neural Network Models For Facial Expression Recognition In The Wild. Neurocomputing, 355, 82-92.

[62]. Jun, H., Shuai, L., Jinming, S., Yue, L., Jingwei, W., & Peng, J. (2018, November). Facial Expression Recognition Based On Vggnet Convolutional Neural Network. In 2018 Chinese Automation Congress (Cac) (Pp. 4146-4151). Ieee.

[63]. Li, M., Xu, H., Huang, X., Song, Z., Liu, X., & Li, X. (2018). Facial Expression Recognition With Identity And Emotion Joint Learning. Ieee Transactions On Affective Computing, 12(2), 544-550.

[64]. Sekaran, S. A. R., Lee, C. P., & Lim, K. M. (2021, August). Facial Emotion Recognition Using Transfer Learning Of Alexnet. In 2021 9th International Conference On Information And Communication Technology (Icoict) (Pp. 170-174). Ieee.

[65]. Sajjanhar, A., Wu, Z., & Wen, Q. (2018, December). Deep Learning Models For Facial Expression Recognition. In 2018 Digital Image Computing: Techniques And Applications (Dicta) (Pp. 1-6). Ieee.

[66]. Chen, A., Xing, H., & Wang, F. (2020). A Facial Expression Recognition Method Using Deep Convolutional Neural Networks Based On Edge Computing. Ieee Access, 8, 49741-49751.