Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

A Survey on Deep Facial Expression Recognition

Manisha Balkrishna Sutar [1], Dr. Asha Ambhaikar [2]

[1,2] Kalinga University, Raipur, Chhattisgarh, India
[1] more.manisha3030@gmail.com, [2] dsw@kalingauniversity.ac.in

Article Info
Page Number: 470-485
Publication Issue:
Vol. 72 No. 1 (2023)

Article History
Article Received: 15 October 2022
Revised: 24 November 2022
Accepted: 18 December 2022

Abstract
This paper reviews state-of-the-art developments in facial emotion recognition using deep learning. We also have analyzed the performance and limitations of the reviewed state-of-the-art deep learning architectures and algorithms for facial emotion recognition. We discuss the contribution, model performance, and limitations of various architectures like CNNs and RNNs. We have also reviewed the performance of various state-of-the-art deep learning algorithms for facial emotion recognition. We discuss the contribution, model performance, and limitations of each algorithm. We have found that the most successful architectures are Hybrid CNN-RNNs, which combine convolutional layers with recurrent ones. This is due to their ability to learn hierarchical representations of data and exploit temporal dependencies in inputs. There are also architectures that use meta-parameters such as attention vectors or word embeddings; however, they do not provide significant improvements over RNNs alone. We have also discussed the application of deep learning to facial emotion recognition, as well as its limitations. We conclude by discussing some future work that can be done in this field. The paper reviews the literature on face recognition. It also explains the relationship between facial emotion and facial feature extraction, which is essential for emotion recognition.

Keywords: - Artificial Intelligence, Emotion Detection, Machine Learning, Facial Recognition, Datasets.

I. INTRODUCTION

When communicating with other people, it is crucial to make use of emotions since human feelings may be used to infer a great deal. A little over two thirds of all communications [1] are non-verbal in nature. According to research [2], 55% of emotions are expressed visually, 38% are expressed vocally, and 7% are expressed verbally. When communicating with other people, it is crucial to make use of emotions since human feelings may be used to infer a great deal. A person's facial expression, tone of voice, and body language can tell you a great deal about their feelings.

For example, when someone feels anxious or uncomfortable their movements are more erratic than usual. They may fidget with their hair or clothing, avoid eye contact with others, look around the room trying to find something of interest to focus on. Humans are able to discern signs that might otherwise be missed in a discussion by using a non-verbal type of communication known as facial emotion, often known as FE. The human mind is capable of perceiving freely stated facial emotions even if they only lasted for a fraction of a second to

four seconds [3]. This is often referred to as microexpressions, and they are involuntary reactions to our emotions. They can be used to determine if someone is lying or telling the truth, or simply communicating their feelings about a situation or topic of discussion. Microexpressions are brief and fleeting, lasting for only a fraction of a second before another expression replaces it [4]. This is because the mind may be taught to do so. Analysis of facial expressions is an interesting and difficult problem that has repercussions in a variety of domains, including human-computer interactions and medical applications, to name a few. The pervasiveness may be attributed to the fact that computers are a source of cheaper and more dynamic labour in addition to being typically quicker for computational analysis [4, 5].

The study of micro expressions is not a new phenomenon. The first article on the topic was published in 1979 by Paul Ekman and Wallace Vigod, who argued that facial expressions are universal across cultures and can be used to accurately detect emotions [2]. Since then, numerous studies have been conducted on the subject. These computers will need to have an understanding of human emotions in order to better connect to people in the majority of situations. One example of a utilized case is the Emotion Detection System, which is used in mental health to diagnose mental and emotional illnesses. Because it can gauge the user's state of mind and respond accordingly, future robots and intelligent support systems will have a far easier time communicating with their human counterparts.

The robotic personal assistants of the future will be better at supporting humans because they will be able to deliver answers on demand and modify their replies based on the emotion being shown by the user. In addition, FER may be used in Market Research Surveys to determine how people feel about a certain message, product, or brand. For instance, a gaming business could invite a small group of game players to test out a new game before it is officially released. The company could then use FER to determine how the players feel about the game and whether or not they would buy it. If they found that a large number of people were indifferent or negative towards the game, then they could tweak it to make it more appealing before releasing it to the public.

The gaming firm can produce the player's facial expressions at each step of the game while the player is playing the game, which helps them enhance the end result. Several research have taken emotion detection and utilised facial cues to extract particular attributes of the subject. Deep learning makes available a diverse selection of algorithms that are able to recognise facial expressions of emotion. Because human emotions are so fluid and unpredictable, the work of FER has been judged to be a very difficult one and has required a significant amount of study to be carried out. The human face is one of the most expressive parts of our body, and through it we can read a wealth of information about what someone is thinking or feeling. From social media to marketing, facial recognition has become a critical part of identifying people and their emotions.

The purpose of this work is to generate some open problems and potential trends for future study in FER by conducting a review of some of the previous research works that have been conducted on FER. The techniques that have been employed, as well as their performances and efficiency, will be discussed. The first section of this work will give an overview of the

literature that has been published in the past decade. The second section will focus on some of the most interesting works in FER and how they use machine learning techniques to achieve their goals. Finally, we conclude with some open problems and potential trends for future research in FER.

II RESEARCH METHODOLOGY

A. Psychology of Emotion Demystified

Researchers have shown that facial expressions account for a significant portion of human communication, namely 55% of the whole spectrum of face-to-face communication between humans. [2],[6]. If you are unable to interpret someone's facial expressions, it is safe to say that you are losing out on more than half of the whole meaning of what they are trying to convey with you. Facial expressions are a crucial part of human communication and are used to convey countless emotions that could range from happiness, sadness, anger, fear or surprise. These expressions allow us to communicate our thoughts and feelings in ways that words cannot always describeIn addition to facial expressions, individuals communicate their feelings by other ways, such as the intonation of their voices, which accounts for around 38% of all discussions. In comparison, the actual choice of words used to push emotions only accounts for 7% of all dialogues. [2],[6],[7]. There are a few different theories about feelings. Plutchik's Wheel of Emotions, Izard, Pankseep & Watt, Levenson, and Ekman are only few of the studies that fall within this category [8]. All of these theories are used to explain how we feel and what the different types of emotions are. Theories that use facial expressions, like Izard's and Ekman's, say that people can interpret emotions based on their facial expressions. Pankseep & Watt suggest that there is a specific part of the brain responsible for processing emotional information.

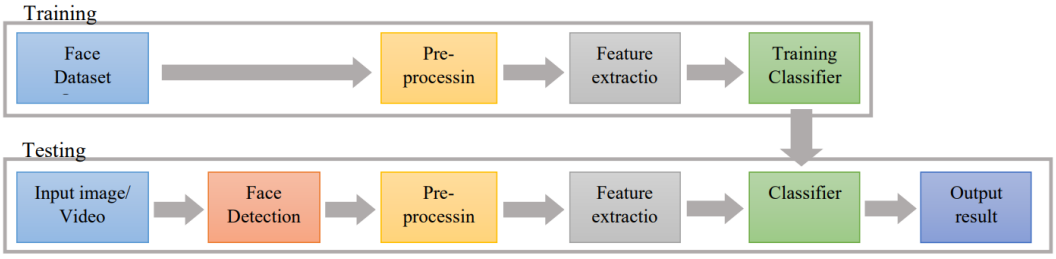

Figure 1. General Overview of FER system

B. Facial Emotion Detection Systems

The detection of the face region, the extraction and representation of data of interest, and the recognition of the expression are the three inherent tasks that are combined in FER. Both a training phase and a testing phase are included in the structure of FER models that are based on machine learning. Figure 2 is a representation of a typical machine learning architecture for FER and it can be seen here. The training phase can be seen as the learning process of a model that is used for the detection and recognition of facial expressions. In this phase, the input data are fed into the network which then learns from it to generate a model that can be used later on in the testing phase. The feature-based geometry method and the appearance-based approach

have both been used to describe the data so far. Both of these techniques have their advantages and disadvantages. Image processing methods were used throughout the development of the feature-based geometric approach in order to extract essential face points (i.e., corners of the lip, the middle of the eye, the ends of the eyebrows, and the tip of the nose).

The obtained coordinates are put to use in the process of building a face geometry out of the retrieved characteristic vectors. The appearance-based method analyses a video clip frame by frame and develops an attribute vector by applying an image filter to the processed images. This may be applied across the whole of the face, or only to a specific region [3].

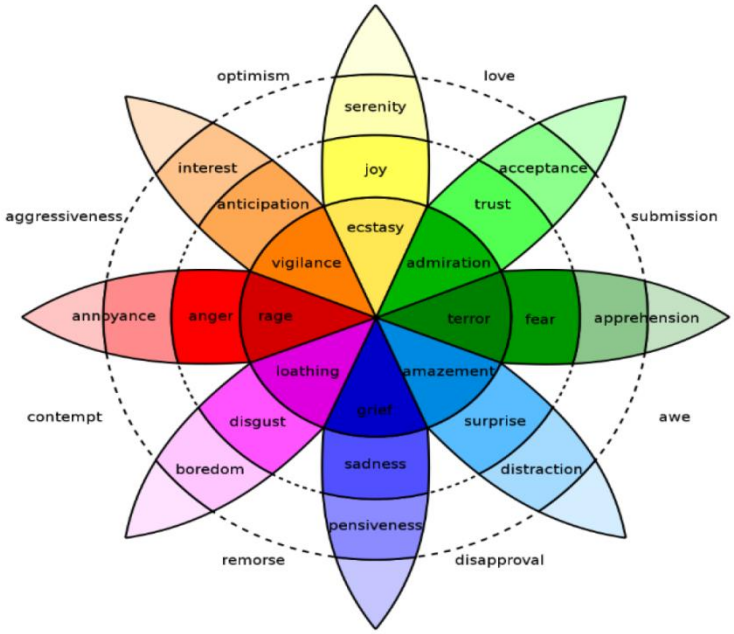

Figure 2. Pluchik's Wheel of Emotions

Recent research has added on one more facial emotion, disdain, which brings the total number of facial emotions studied to seven (7)[11]. Facial Emotion Recognition is a Typical Classification Problem that can be Solved Using Several Different Classification Methods [12]. Some of these classification methods include k-Nearest Neighbours (KNN), Decision Tree (DT), Learning Vector Quantization (LVQ), and multilayer Feed-forward Neural Network (MFFNN). The artificial neural network (ANN), the minimal distance classifier (MDC), the support vector machine (SVM), the linear discriminant analysis (LDA), and the hidden markov model (HMM) [1, 12].

In recent years, ANNs such as Convolutional Neural Networks (CNN) and Deep Convolutional Neural Networks (Deep CNN) have been used for image categorization with extremely high degrees of accuracy. ANNs are computer programmes that mimic the biological neural networks seen in the human brain. They are composed of a network of artificial neurons that are all linked with one another. Each neuron is capable of sending signals to other neurons that are linked to it, and those other neurons are able to process the signals they receive and then pass them on. Neurons are stacked in layers, and each layer may conduct a different kind of modification on the input signal that it receives depending on the

kind of neuron that it contains. Some ANNs have the capability of backpropagation, which enables data to flow in the opposite direction of the network's normal flow in order to fine-tune the efficiency of the network. ANN is the foundation of Deep Learning, often known as Deep Neural Networks. The various levels that are present in the network are where the adjective "deep" comes from. A Deep Convolutional Neural Network design for recognising facial expressions of emotion is shown in Figure 3.

The convolutional layers of CNNs are responsible for extracting features from input pictures for the use of subsequent levels (i.e., pooling, dropout, and fully connected layers). The convolutional layer is made up of many tiny patches, each of which makes use of a different filter value to modify the whole picture and produce feature maps. This is accomplished by the use of equation (1). Where is the picture that is being read in, is the name of the filter being applied, and is the size of the matrix that is being formed as a consequence. Where is the picture that is being read in, is the name of the filter being applied, and is the size of the matrix that is being formed as a consequence.

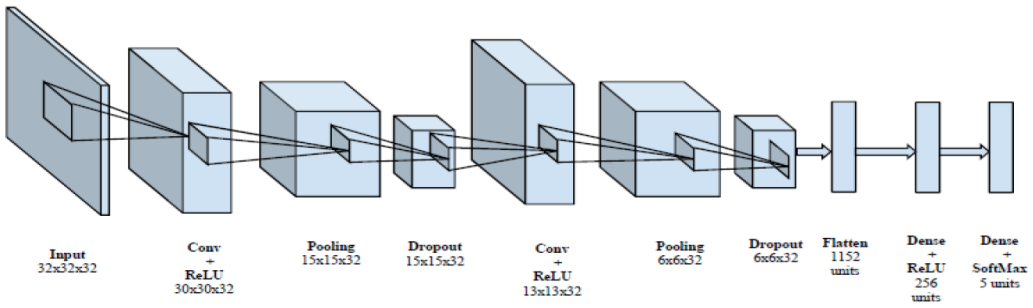

Figure 3. Deep CNN architecture for FER

The data that is produced from the convolution layer are sent to the pooling layer through a lossless transfer, which results in a reduction in the size of the image. Using the flatten layer, the final data is a two-dimensional array that has been transformed into a one-dimensional vector. This vector will then be input into the neural network so that it may be classified. The neural network backpropagates the errors of the network to change the weights, which then decreases the error (loss) function. This process is repeated until the error (loss) function is minimised. The equation (2) [5] is used to calculate the weight adjustment.

C. Dataset for Facial Emotion Detection Systems

There are datasets available for each of these feelings. Table I contains a representation of the information pertaining to the datasets. CASIA-Face-Africa is a dataset that takes into account the demographic imbalance that exists in the currently available dataset. This demographic imbalance has a tendency to impact the performance of face biometric systems when applied to African individuals. The collection consists of 38,546 pictures, and 1,183 of the subjects are people of African descent. Face biometrics, face image preprocessing, face feature analysis and matching, facial expression recognition, sex/age estimation, ethnic categorization, face image production, and other related topics may all be investigated with its help. For the purpose of facilitating the recognition of these face landmarks, the dataset has been manually

tagged with 68 landmark points. The collection of datasets includes a segment with 70 individuals who acted out seven different emotions (i.e., neutral, angry, sad, happy, surprised, fearful, and disgusted) [13].

III. RELATED WORK

Over the last several years, academics have experimented with a variety of architectural approaches for the purpose of face expression identification, with varying degrees of success. In this section of the study, we will examine current research that have been completed in this subject making use of deep learning techniques. This study also intends to map out trends and patterns in recent research and create a suggestion for future directions in the area of facial emotion detection, as indicated in Table II. Specifically, this work aims to map out trends and patterns in the field of facial emotion identification.

The authors Pranav et al. [5] created a Deep Convolution Neural Network (DCNN) model that could categorise all five human facial expressions. This model was constructed with the help of the Keras Deep Learning Library. The model that they suggested made use of two convolution layers, each of which was followed by a dropout layer, a pooling layer, and then another convolution layer. The Rectified Linear Unit, also known as ReLU, made use of the activation function, which changed all negative values to zero. After the formation of a two-dimensional array, the data is sent on to the flatten layer, where it is transformed into a one-dimensional vector. The transformed data is then sent on to a two-layered network, which is used to categorise the feelings. In order to provide a probabilistic output, the activation function for the output layer that is employed is called softmax. Using a learning rate of 0.01, the model was trained and validated over the course of 11 epochs. The accuracy of the model was calculated to be 78.04%. The dataset that was employed was compiled by hand with the use of a 48 megapixel camera. It consisted of a total of 2550 photos, each of which had a pixel resolution of 1920 by 2560. The dataset was partitioned into 2040 pictures for training, 255 photos for validation, and 255 images for testing. The following feelings were evaluated: anger, happiness, neutrality, sadness, and surprise.

An experiment FER system has been developed, and it is comprised of the following three processes: preprocessing with the Viola-Jones algorithm, feature extraction with local fisher discriminant analysis (LFDA) for dimensionality reduction and k-nearest neighbours (KNN), and classification with a feed-forward artificial neural network (ANN) [6]. JAFFE was the dataset that was utilised for this experiment; however, the researchers only focused on four different feelings. Out of the seven feelings represented in the sample, happy, neutral, surprised, and sad were the most prevalent. The efficiency of the two different classifiers was evaluated side by side. The 1NN algorithm fared better with neutral and unhappy emotions, but the ANN method performed well with joyful and surprised emotions. However, the ANN algorithm surpasses the 1NN approach when it comes to the average performance, with a performance of 66.66% as opposed to 54.16%.

In their paper [14], Jaiswal et al. developed a deep learning architecture for detecting facial emotions that consisted of two distinct CNN networks. The model that has been suggested

makes use of Keras, and it includes an input shape that is 48*48*1, as well as two models for feature extraction that have the same kernel size. Before being sent to a fully connected softmax layer for classification, the submodels are first "flattened" into vectors, and then those vectors are "concatenated" together to form one long vector. The performance of their architecture was assessed using two datasets, FER2013 and JAFFE, and the accuracies achieved for the two datasets were, respectively, 70.14% and 98.65%. The evaluation was based on the fact that the accuracies obtained for FER2013 were higher than those realised for JAFFE. The researchers selected the datasets with the intention of making the model more resilient in terms of its overall variety. In addition, Lasri et al. [15] presented a CNN architecture for the purpose of recognising the facial expressions of students. Their design is made up of four convolution layers, each of which has maximum pooling layers and two layers that are completely linked. The suggested method utilises softmax in order to create predictions about facial expressions. The FER2013 dataset was used for both the training and validation of the model. Additionally, the accuracy of their model was rated at 70%.

Hand-over-Face Gesture-based Facial Emotion Recognition Method (HFG FERM) is the name of a CNN model that was suggested by Naik and Mehta [16] for the purpose of recognising facial expressions of emotion while the hands are covering the face. As an example of occlusion, hand-over-face is normally regarded in other types of facial expression identification; hence, pictures with hand-over-face are not included in the trials designed to examine this phenomenon. The suggested study gives thorough coding schemas along with extra hand signals that assist detect unknown emotions in addition to fundamental feelings. These unexplored feelings include confidence, the ability to make a choice, as well as terrified, embarrassed, furious, and ok signs. Their approach was able to identify feelings in circumstances involving excessive hand occlusion as well as extreme head rotation. The scientists verified their model by using pictures from the Cam3d corpus, the FER2013 dataset, and public domains, which together totaled a total of 18 different emotional classifications. Comparisons were made between the performance of their model and that of two other models: the Multimodal Fusion Approach (MMFA) and the Emotion Recognition via Facial Gestures model (ERTFG). According to the results of their trials, their model HFG FERM performs better than both MMFA and ERTFG on a variety of levels.

The authors of [17] offered several designs of Convolution Neural Networks (CNNs) of two models in order to categorise seven facial expressions of emotion. The first CNN model consisted of five layers: one dropout layer, two fully connected layers, four convolutional layers, and four max pooling layers. The first model served as the basis for the second model, with the exception that the second model included more data. iCV MEFED was the dataset that was used for this experiment (Multi-Emotion Facial Expression Dataset). Their selection of the dataset was one that was not too old and had complex or contrasting feelings, such as being angry at being shocked or crying with happiness. The collection is comprised of 5,750 photos, each of which has a resolution of 5184 by 3456 pixels. It was shown that the model performed much better with photos that had not been distorted as opposed to ones that had been boosted. It was also discovered that some feelings, such as disappointment and disdain, are less accurately anticipated than the other emotions.

A model was suggested by Bouzakraoui et al. [18] that has the capability of automatically detecting the facial expressions that are produced by customers in response to a product or service. After first extracting geometric information from the customers' faces, the researchers employed an altered version of the support vector machine to make predictions about the customers' levels of pleasure. During the geometric feature extraction process, the picture is turned into geometric primitives such as points and curves by measuring the relative distances between various characteristics such as the eyes, eyebrows, nose, mouth, and chin. These distances are measured relative to one another. When these distances are considered, a vector consisting of 19 values is produced to reflect the customer's face expression. The JAFFE dataset was used for the purposes of this particular experiment. The researchers reorganise the feelings that were reflected in the dataset into three distinct categories: pleased, not satisfied, and neutral.

A model that is based on a single Deep Convolutional Neural Network was suggested by Jain and colleagues [19]. (DNNs). The model that is being presented has a total of six layers of convolution, two layers of deep residual blocks, and two layers that are completely linked. Each of these layers has a ReLU serving as the activation function and dropout. The model is able to learn more nuanced traits that are associated with certain emotions with the aid of these aspects of the model. Two residual blocks each have four convolution layers of varied sizes, one skip connection, two short connections, and two short connections between them. Softmax is the method that is used for the categorization of feelings. It was discovered that the combination of fully connected networks and residual block increased the overall performance of their proposed model, which was trained and evaluated on two datasets: CK+ and JAFFE. These datasets were used for training and testing the suggested model.

The Valence-Arousal Dimensional Emotion Model was used by the researchers in [20] to present a face expression recognition system that is based on the model. The model that is being suggested makes use of a valence dimension prediction that consists of nine different levels. The suggested model requires the use of CNN in order to forecast a result that is equivalent to the weighted fusion of valence value and the probability that is associated with it. The architecture of the CNN comprises of four layers of convolution, each of which uses ReLU as its activation function; three maximum pooling layers, which are put after convolution layers 2, 3, and 4 sequentially; and two layers that are completely linked. The categorization in this model is accomplished with the help of softmax. The CK+ and FER2013 datasets were used throughout the training process for the suggested model. The researchers took the photos from both the training set and the test set and averaged them using the SAM method. Each picture was assigned a valence dimension between 1 and 9, and the researchers utilised 10 annotations to do this. The effectiveness of the system was evaluated by observing the reactions of participants while they watched a film and having the system identify their facial expressions.

According to the findings of researchers in [21], an attention-based ACNN model for facial emotion recognition works effectively for a partial block on the face. Occlusion is the condition that happens when anything, such as a hand, hat, hair, or any other object, covers a portion of the face. Because of this, it is more difficult for a FER system to interpret the

emotions. The problem of occlusion is addressed by the model by concentrating on and coordinating the emotions derived from the symmetrical parts of the face that are referred to as the informative areas. The aforementioned model makes use of two distinct variants of ACNN in order to identify the emotional zone as well as the block region. It is necessary to utilise the face of the Path-Gated Unit (PGunit) in order to identify the blocked area, while the Global-Gated Unit is required for the identification of the whole face region (GG-unit). The attention unit will acquire the scalar weights for the patch area via adaptive learning. Nevertheless, the ACNNs identify feelings based on facial landmarks. The aforementioned model was educated and validated with the help of the Indian Spontaneous Expression Database (ISED), which is broken down into four (4) categories (disgust, happy, sad, and surprise)

Navaz et al. [22] put out the hypothesis that preprocessing has the potential to improve the performance of deep learning architecture. The approaches that they suggested for use in the pretreatment step include picture data management, in which the photographs are examined to see whether or not they have been incorrectly labelled by use of a Perl file that also sorts the images into the appropriate folders. They made use of an additional script in order to transfer photographs of poor quality into a certain folder. Following the use of the Very-Deep Super Resolution method comes the following step in the preprocessing sequence, which is known as quality enhancement. This step involves enhancing the resolution quality of the picture via the use of bicubic interpolation. Following the completion of the preprocessing step, the collected data was fed into a CNN architecture in order to identify facial expressions. The Indian Movie Face Database served as the source material for this investigation's dataset (IMFDB). It should come as no surprise that the preprocessing improves the performance of the facial emotion recognition system.

In addition, researchers that worked on the project [23] developed a model of CNN architecture with enhanced preprocessing and feature extraction for FER. In the first step of the procedure, called preprocessing, the researchers cut out the undesired parts of the picture, leaving just the face visible. After that, the picture is shrunk down to the dimensions required by the CNN's input. After that, the brightness and contrast of the picture are fixed by using MinMax normalisation, which is an intensity normalising technique. For the purpose of feature extraction, a technique that combines CNN and Histogram Oriented Gradients (HOG) is used. The JAFFE and FER2013 datasets were used for the training and validation of the model.

An algorithm for recognising emotions in a person's face that use Gabor filters and CNN in order to recognise facial expressions. [24]. The model extracted features with the help of two different Gabor filters, the first one's output serving as the input for the second one. The greatest responsiveness is achieved using a Gabor filter at the edges as well as the locations where the texture changes. After applying the filter to the picture, the characteristics of the face that are used for analysing facial expressions, such as the form of the brows, the eyes, the nose, and the lips, are brought into focus. After the classification filters have been applied, the picture is sent on to be processed by the CNN architecture. The CNN has three convolutional layers, each with a MaxPooling activation function, as well as a flatten layer, two dense layers, a dropout layer, and a softmax layer. Additionally, the CNN features a dropout layer. The

accuracy of the suggested model was measured using the JAFFE dataset, and it was found to be 97% after 25 iterations. At the conclusion of the trial, it was discovered that the model that they had presented was far quicker than the other CNN models.

The researchers that published [25] suggested employing Multi-Task CNN (MTCNN) for face identification and the usage of ShuffleNet V2 for emotion recognition. The pointwise group convolution and the channel shuffle are two processes that are used in the ShaffleNet architecture that are not utilised in the standard CNN. The convolution is split up and distributed over many CPUs in order to do concurrent separable convolution operations using group convolution. This may be highly expensive in terms of the complexity of the calculation required. Consequently, the channel sparse connect is a component of the design. The accuracy of the model was measured at 71.19% when it was applied to the FER2013 dataset.

Researchers in [26] came to the conclusion that the difficulties inherent in unimodal emotion detection systems lead to a reduction in accuracy. As a result, they presented a multimodal emotion identification system that uses information that is complimentary to emotional cues conveyed by facial expressions and voice. The model that has been suggested utilises a 1D CNN and a bi-directional LSTM to extract acoustic data from speech and a 2D CNN to extract high-level information from facial expressions. Both of these components are convolutional neural networks. The algorithm SoftMax is used to complete the joint classification. The suggested model was educated and evaluated with the use of IEMOCAP, a dataset that contains 12 hours' worth of audio-visual data. After combining "happy" and "excited," the researchers came up with a dataset that consisted of four categories: "excited," "angry," "sad," and "neutral." The suggested model demonstrated an improvement in accuracy of 10.05% and 11.27%, respectively, in comparison to speech emotion recognition using a single modality and facial expression recognition using a single modality.

A multimodal emotion recognition system was also presented by researchers in paper 27. [27] The model that was presented here had a total of four neural networks, each of which was responsible for extracting different aspects from the audio, face, and gesture datasets. On the validation set, the grid-search method was investigated by the model for the purpose of performance optimization. The model is made up of two sub-models for face expression, which are referred to as the feature embedding model and the frame attention model. Deep convolutional neural networks (CNNs) are used in the feature embedding model to produce a feature vector from the facial picture. The frame attention model is responsible for learning the weights and dynamically aggregating the feature vectors in order to provide a single video representation that is discriminative. For the body movement and gesture model, the researchers made use of a tool called the Temporal Shift Module, or TSM. TSM provides both great levels of efficiency and performance. The researchers carried out the classification job after first extracting characteristics from the audio model with the help of the openEar programme. Following that step, the researchers combined the scores from all of the models by using a weighted total. 76.43% was the greatest overall accuracy that could be achieved throughout the testing.

Transfer Learning was evaluated for its performance in comparison to training from scratch for deep facial emotion recognition by the authors of [28]. In this comparison, Alexnet and VGG16 were the two networks that were used. Training Alexnet from scratch, training Alexnet with transfer learning, training VGG16 from scratch, and training VGG16 with transfer learning were the four steps of training that were carried out. The Alexnet CNN architecture is well-known for its ability to categorise photos into one thousand different object categories and to learn from the ImageNet dataset. It has seven activation layers, three completely connected layers, five convolutional layers, three pooling layers, two dropout layers, and three fully connected layers. VGG16 is likewise a CNN architecture that was trained using ImageNet to categorise 1000 different objects, similar to Alexnet. On the other hand, the topology of its network is distinct, consisting of 13 convolutional layers, five pooling layers, two dropout layers, 15 activation layers, and three fully-connected layers. The researchers started from scratch when they redesigned the two training architectures and obtained the pre-trained models so that they could do transfer learning. Every one of the models was put through its paces using the RaFD data. According to the results of the experiment, Alexnet and VGG16 obtained 95% and 95.33%, respectively, for transfer learning, but they only achieved 84% and 16.67% for training from scratch. The researchers, on the other hand, concluded that the inadequate training data was to blame for the poor performance of the VGG16 training from scratch since the network is too broad and has an imbalance in the input data.

TABLE II EXISTING DEEP LEARNING TECHNIQUES FOR FER

[5]	Self-trained two convolution layers with a pooling layer followed by a dropout layer after each convolution layer and ReLU as activation function (AF). softmax as AF for the output layer	Validation accuracy of 78.04%	
[6]	LFDA for dimensionality reduction, 1NN and Feed-forward ANN for classification	The ANN algorithm outperforms the 1NN algorithm with 66.66% as against 54.16%	1NN performed better with sad and neutral, and whiles ANN performed well with happy and surprise
[15]	CNN model with 4 convolutional layers, 4 pooling layers, and 2 fully connected layers. The model uses softmax at AF for the output.	70% validation accuracy	The model, in some cases, wrongly predicts fear as the sad face.
[17]	Two CNN models; the first model had four conv layers, four	The first model performs better than	The model, in some instances,

	max pooling, one dropout, and two fully connected layers. The second model is the same as the first model with data augmentation	the second model.	confused fear with surprise and contempt with sadness.
[20]	Valence-Arousal dimensional model which is a CNN architecture consisting of 4 conv layers each with ReLU as AF, 3 max pooling layers after the last 3 conv layers and two fully connected layers. The model uses softmax for classification	The model achieved an RMSE index of 0.0857 0.0064	There is some overfitting in the training and testing phase, possibly due to the insufficient number of facial images with valence dimensions.
[21]	Two variations of attention-based ACNN, one to detect blocked region of the face and the other to detect the region with the emotion.	Not deterministic	Model is limited to video datasets but not video streams due to latency for the later.
[23]	CNN architecture with improved preprocessing and feature extraction	accuracy of 91.2% and 74.4% were obtained on JAFFE and FER2013 database respectively	Applying both HoG and facial landmarks causes over-fitting, thereby decreasing performance

IV. RESULT AND DISCUSSION

The challenge of face expression detection is an area of interest in human-computer interaction (HCI) and affective computing. As a result, some amount of work has been done on the subject, and as a result, there is still a significant amount of work being done on the problem. Nevertheless, further input is necessary for some of the difficulties.

To begin, there are no big datasets available that are capable of training very massive networks. Because there are not enough datasets that are a sufficient size to train extremely deep networks, architects of deep learning systems are forced to restrict their designs to function with just a modest scale network. It is impossible to ignore the fact that a dataset of that magnitude is extremely difficult to acquire. This is due to the fact that it will require a significant amount of money to acquire a dataset of posed images, as well as a significant amount of time and effort to generate a dataset of that magnitude in the wide. The second problem is that in order to train such a network, a system with hardware specs at least as high as those required to run the network itself would be required.

Achieving a high level of human face expression recognition via the use of computer technologies is one thing. Nevertheless, emotion detection amongst people is a significant

challenge. This is due to the fact that humans are able to interpret emotional signals in a variety of ways, including via facial expression, voice intonation, word choice, body language, and written text. There is also an increasing worry over individuals leaving a conventional double life, particularly on the internet, which is referred to as "posing for the cameras." Getting to the bottom of problems like that requires a significant amount of cross-disciplinary education. Although there is some work being done in that sector, a significant amount of further effort is necessary to bridge that gap.

Transfer learning has a lot to offer when it comes to attaining large-scale holistic emotion detection and identification for HCI, as was revealed throughout the process of putting together this study. [Citation needed] A great deal of investigation has been carried out in a variety of subfields of affective computing, and the results of this work have resulted in the production of very impressive architectural designs in terms of both the computational complexity and the level of performance they offer in their respective emotion recognition subfields. The use of transfer learning to train already-established networks may be the primary focus of research in emotional computing in the years to come.

V. CONCLUSION

In this research, we took a look at the subject of emotion recognition, more especially emotion identification in facial expressions, which has become more important in the area of human-computer interaction (HCI). In addition to this, we discussed the usage of deep learning models and the datasets that are now accessible for the purpose of addressing the issue. Research efforts that have been done in the topic of facial emotion have been evaluated, with particular focus placed on the different architectures, tools, and algorithms that have been used, as well as the datasets that have been collected. We ultimately reviewed the outstanding problems and the future developments as far as face expression identification and emotion recognition are concerned. Your investigation revealed, in general, that ANNs, and more especially CNNs, are now in the lead in the architectural or FER system since they offer superior outcomes. This was found out to be the case. It has also been noticed that in order for certain researchers to improve their outcomes, they work to enhance their preprocessing procedures and feature extraction methods. Last but not least, we found that multimodal systems are becoming more popular. These systems provide a more accurate emotion detection system by using other emotional clues, such as ECG and verbal cues. Even though there have been significant advancements in the field of facial expression identification, the algorithms are still restricted to static photos, which is fine for the perfect situation. Nevertheless, human emotions may be rather complicated, and situations that take place in the real world call for the development of methods that can record continuous, spontaneous, and delicate facial expressions. It has also been shown that some feelings are simpler to train than others due to the number of training samples available for them.

References

[1] B. Balasubramanian, P. Diwan, R. Nadar, and A. Bhatia, 'Analysis of Facial Emotion Recognition, in 2019 3rd International Conference on Trends in Electronics and

Informatics (ICOEI), Tirunelveli, India, Apr. 2019, pp. 945–949. doi: 10.1109/ICOEI.2019.8862731.

[2] W. Mellouk and W. Handouzi, 'Facial emotion recognition using deep learning: review and insights', Procedia Comput. Sci., vol. 175, pp. 689–694, 2020, doi: 10.1016/j.procs.2020.07.101.

[3] L. Stanciu and A. Albu, 'Analysis on Emotion Detection and Recognition Methods using Facial Microexpressions. A Review', in 2019 E-Health and Bioengineering Conference (EHB), Iasi, Romania, Nov. 2019, pp. 1–4. doi: 10.1109/EHB47216.2019.8969925.

[4] A. Hassouneh, A. M. Mutawa, and M. Murugappan, 'Development of a Real-Time Emotion Recognition System Using Facial Expressions and EEG based on machine learning and deep neural network methods', Inform. Med. Unlocked, vol. 20, p. 100372, Jan. 2020, doi: 10.1016/j.imu.2020.100372.

[5] E. Pranav, S. Kamal, C. Satheesh Chandran, and M. H. Supriya, 'Facial Emotion Recognition Using Deep Convolutional Neural Network', in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, Mar. 2020, pp. 317–320. doi: 10.1109/ICACCS48705.2020.9074302.

[6] R. Ranjan and B. C. Sahana, 'An Efficient Facial Feature Extraction Method Based Supervised Classification Model for Human Facial Emotion Identification', in 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Ajman, United Arab Emirates, Dec. 2019, pp. 1–6. doi: 10.1109/ISSPIT47144.2019.9001839.

[7] T. Wu, S. Fu, and G. Yang, 'Survey of the Facial Expression Recognition Research', in Advances in Brain Inspired Cognitive Systems, Berlin, Heidelberg, 2012, pp. 392–402. doi: 10.1007/978-3- 642-31561-9_44.

[8] J. L. Tracy and D. Randles, 'Four Models of Basic Emotions: A Review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt', Emot. Rev., vol. 3, no. 4, pp. 397–405, Oct. 2011, doi: 10.1177/1754073911410747.

[9] P. Ekman, Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life. Henry Holt and Company, 2004. [Online]. Available: https://books.google.com.gh/books?id=AoIUp5fJkIcC

[10] R. Plutchik, 'The Nature of Emotions: Clinical Implications', in Emotions and Psychopathology, M. Clynes and J. Panksepp, Eds. Boston, MA: Springer US, 1988, pp. 1–20. doi: 10.1007/978-1-4757- 1987-1_1.

[11] [A. R. Dores, F. Barbosa, C. Queirós, I. P. Carvalho, and M. D. Griffiths, 'Recognizing Emotions through Facial Expressions: A Largescale Experimental Study', Int. J. Environ. Res. Public. Health, vol. 17, no. 20, p. 7420, Oct. 2020, doi: 10.3390/ijerph17207420.

[12] Dasharath. K. Bhadangkar, J. D. Pujari, and R. Yakkundimath, 'Comparison of Tuplet of Techniques for Facial Emotion Detection', in 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, Oct. 2020, pp. 725–730. doi: 10.1109/I-SMAC49090.2020.9243439.

[13] J. Muhammad, Y. Wang, C. Wang, K. Zhang, and Z. Sun, 'CASIAFace-Africa: A Large-scale African Face Image Database', ArXiv210503632 Cs, May 2021, Accessed: Jul. 07, 2021. [Online]. Available: http://arxiv.org/abs/2105.03632

[14] A. Jaiswal, A. Krishnama Raju, and S. Deb, 'Facial Emotion Detection Using Deep Learning', in 2020 International Conference for Emerging Technology (INCET), Belgaum, India, Jun. 2020, pp. 1–5. doi: 10.1109/INCET49848.2020.9154121.

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

[16] N. Naik and M. A. Mehta, 'An Improved Method to Recognize Handover-Face Gesture based Facial Emotion using Convolutional Neural Network', in 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, Jul. 2020, pp. 1–6. doi: 10.1109/CONECCT50063.2020.9198376.

[17] S. Begaj, A. O. Topal, and M. Ali, 'Emotion Recognition Based on Facial Expressions Using Convolutional Neural Network (CNN)', in 2020 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA), Tirana, Albania, Dec. 2020, pp. 58–63. doi: 10.1109/CoNTESA50436.2020.9302866.

[18] M. S. Bouzakraoui, A. Sadiq, and A. Y. Alaoui, 'Appreciation of Customer Satisfaction Through Analysis Facial Expressions and Emotions Recognition', in 2019 4th World Conference on Complex Systems (WCCS), Ouarzazate, Morocco, Apr. 2019, pp. 1–5. doi: 10.1109/ICoCS.2019.8930761.

[19] D. K. Jain, P. Shamsolmoali, and P. Sehdev, 'Extended deep neural network for facial emotion recognition', Pattern Recognit. Lett., vol. 120, pp. 69–74, Apr. 2019, doi: 10.1016/j.patrec.2019.01.008. [20] S. Liu, D. Li, Q. Gao, and Y. Song, 'Facial Emotion Recognition Based on CNN', in 2020 Chinese Automation Congress (CAC), Shanghai, China, Nov. 2020, pp. 398–403. doi: 10.1109/CAC51589.2020.9327432.

[20] S. Engoor, S. SendhilKumar, C. Hepsibah Sharon, and G. S. Mahalakshmi, 'Occlusion-aware Dynamic Human Emotion Recognition Using Landmark Detection', in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, Mar. 2020, pp. 795–799. doi: 10.1109/ICACCS48705.2020.9074318.

[21] A. N. Navaz, S. M. Adel, and S. S. Mathew, 'Facial Image Preprocessing and Emotion Classification: A Deep Learning Approach', in 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, Nov. 2019, pp. 1–8. doi: 10.1109/AICCSA47632.2019.9035268.

[22] A. John, A. Mc, A. S. Ajayan, S. Sanoop, and V. R. Kumar, 'Real-Time Facial Emotion Recognition System With Improved Preprocessing and Feature Extraction', in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, Aug. 2020, pp. 1328–1333. doi: 10.1109/ICSSIT48917.2020.9214207.

[23] M. M. Taghi Zadeh, M. Imani, and B. Majidi, 'Fast Facial emotion recognition Using Convolutional Neural Networks and Gabor Filters', in 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI), Tehran, Iran, Feb. 2019, pp. 577–581. doi: 10.1109/KBEI.2019.8734943.

[24] A. Ghofrani, R. M. Toroghi, and S. Ghanbari, 'Realtime Face-Detection and Emotion Recognition Using MTCNN and miniShuffleNet V2', in 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI), Tehran, Iran, Feb. 2019, pp. 817–821. doi: 10.1109/KBEI.2019.8734924.

[25] L. Cai, J. Dong, and M. Wei, 'Multimodal Emotion Recognition From Speech and Facial Expression Based on Deep Learning', in 2020 Chinese Automation Congress (CAC), Shanghai, China, Nov. 2020, pp. 5726–5729. doi: 10.1109/CAC51589.2020.9327178.

[26] G. Wei, L. Jian, and S. Mo, 'Multimodal(Audio, Facial and Gesture) based Emotion Recognition challenge', in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, Nov. 2020, pp. 908–911. doi: 10.1109/FG47880.2020.00142.

[27] I. Oztel, G. Yolcu, and C. Oz, 'Performance Comparison of Transfer Learning and Training from Scratch Approaches for Deep Facial Expression Recognition', in 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, Sep. 2019, pp. 1–6. doi: 10.1109/UBMK.2019.8907203.