Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.7 No.1 January 2022, P-ISSN : 2502-3470, E-ISSN : 2581-0367

A Survey on Deep Learning Algorithms in Facial Emotion Detection and Recognition

Prince Awuah Baffour[1], Henry Nunoo-Mensah[2], Eliel Keelson[3], Benjamin Kommey[4]

[1,2,3,4]*Department of Computer Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana*
[1]pawuahbaffour1@st.knust.edu.gh
[2]hnunoo-mensah@knust.edu.gh (*)
[3,4][ekeelson, bkommey.coe]@knust.edu.gh

Received: 2021-11-18; Accepted: 2022-01-15; Published: 2022-01-20

Abstract— Facial emotion recognition (FER) forms part of affective computing, where computers are trained to recognize human emotion from human expressions. Facial Emotion Recognition is very necessary for bridging the communication gap between humans and computers because facial expressions are a form of communication that transmits 55% of a person's emotional and mental state in a total face-to-face communication spectrum. Breakthroughs in this field also make computer systems (robotic systems) better serve or interact with humans. Research has far advanced for this cause, and Deep learning is at its heart. This paper systematically discusses state-of-the-art deep learning architectures and algorithms for facial emotion detection and recognition. The paper also reveals the dominance of CNN architectures over other known architectures like RNNs and SVMs, highlighting the contributions, model performance, and limitations of the reviewed state-of-the-art. It further identifies available opportunities and open issues worth considering by various FER research in the future. This paper will also discover how computation power and availability of large facial emotion datasets have also limited the pace of progress.

Keywords— Artificial Intelligence, Emotion Detection, Machine Learning, Facial Recognition, Datasets

I. INTRODUCTION

The use of emotions in human-to-human communication is essential as a lot can be derived from human emotions. Non-verbal communication accounts for about 66% of total communications [1]. Research has shown that 55% of emotions are visual, 38% are vocal, and 7% are verbal [2]. Facial Emotion (FE), as one of the non-verbal forms of communication, is used by humans to deduce cues that will otherwise be lost in a conversation. The human mind is trained to perceive facial emotions expressed openly even if the expression lasted for 1/25 fractions of a second to 4 seconds [3]. The field of facial expression analysis is a fascinating and challenging issue with implications in various fields, including human-computer interactions and medical applications. The pervasiveness is because computers are generally faster for computational analysis and are also a source of cheaper and more dynamic labor [4],[5]. These computers will have to understand human emotions to better relate to humans in most instances. Some used cases include the Emotion Detection System in mental health to detect mental and emotional disorders. Communication with future robots and intelligent assistance systems will be better since that system will detect the user's moods and assist appropriately. Future robotic personal assistants will be better at assisting humans because they will provide answers on demand and vary their reactions depending on the user's emotion. FER can also be employed in Market Research Survey to get customers' sentiments on a message, a product, or a brand. For instance, a gaming company can select a few game players and play a new game they are about to release. While they play the game, the gaming company can generate their facial emotions at every game stage and improve the final product. Several studies have used facial landmarks to derive specific characteristics from emotion detection. Deep Learning presents a wide range of algorithms capable of Facial Emotion Recognition. Considering how dynamic and irregular human emotions are, the task of FER has been deemed a huge one and has necessitated much research.

This work will review some of the research works on FER, the methods used, their performances and efficiencies, and the possible setbacks to develop some open issues and likely trends for future research in FER.

The rest of this paper is organized in this way; in section 2, we give the background of emotion detection, specifically facial emotion detection and the available datasets. In session 3, were viewed recent works in Facial Emotion Recognition. Open issues are discussed in section 4. Section 5 is the concluding part of this paper.

II. RESEARCH METHODOLOGY

A. *Psychology of Emotion Demystified*

According to researchers, Facial expressions form a major part of human communication, 55% of the entire human face-to-face communication spectrum [2],[6]. This is to imply that the inability to read facial expressions means you are missing out on more than half of the total meaning of what

DOI : http://dx.doi.org/10.25139/inform.v7i1.4282

someone wants to communicate. Aside from facial expressions, humans express emotions through other means like voice intonation, which accounts for about 38% of all conversations, and just 7% are the actual choice of words used to press emotions. [2],[6],[7]. There are a couple of models on emotions. These include, among others, Plutchik's Wheel of Emotions, Izard, Pankseep & Watt, Levenson, and Ekman [8]. Among these models, Ekman's list of basic emotions is recently used in FER research [9]. Robert Plutchik's wheel classifies emotions into eight primary categories, as displayed in Figure 1: joy, trust, fear, surprise, sadness, anticipation, anger, and disgust. Each of the primary emotions has a polar opposite. All other emotions are derived from the primary emotions' intensity variations [10].

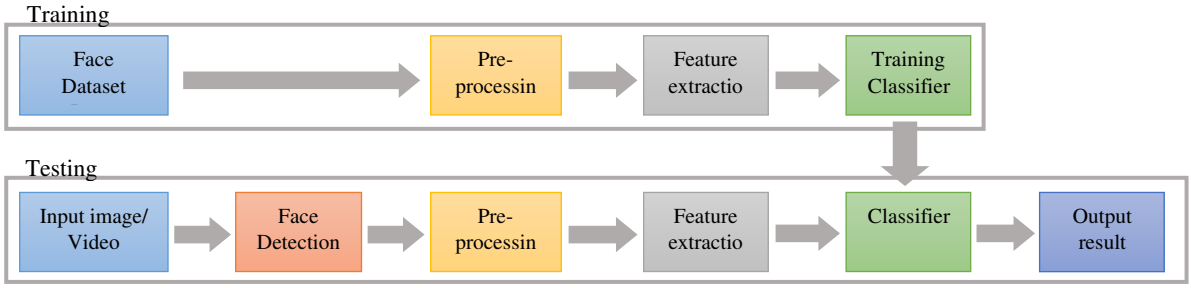

Figure 1. General Overview of FER system

B. *Facial Emotion Detection Systems*

FER combines three inherent tasks: detecting face area, extracting and representing data of interest, and recognizing the expression. Machine Learning models of FER have two phases; a training phase and a testing phase. Representation of a typical FER machine learning architecture is shown in Figure 2. Two approaches have been adopted for data representation, viz., the feature-based geometric and the appearance-based approaches. During the implementation of the feature-based geometric approach, image processing techniques to extract vital facial points (i.e., corners of the lip, middle of the eye, the ends of the eyebrows, and the tip of the nose). The resulting coordinates are used to construct a facial geometry made of the extracted characteristics vectors. The appearance-based approach examines video frame by frame and generates an attribute vector using an image filter. This can be used on the whole face or only a certain area [3].

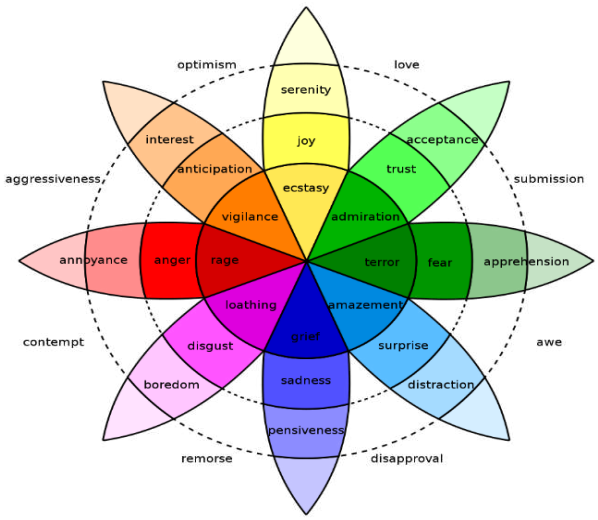

Figure 2. Pluchik's Wheel of Emotions

Recent studies have added on more facial emotion, contempt, which makes interest seven (7)[11]. Facial Emotion Recognition is a typical classification problem that can be solved using several classification methods such as k-Nearest Neighbours (KNN), Decision Tree (DT), Learning Vector Quantization (LVQ), and multilayer Feed-forward Neural Network (MFFNN) [12]. The minimum distance classifier (MDC), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Hidden Markov Model (HMM), and Artificial Neural Network (ANN) [1], [12].

In recent years ANNs such as Convolution Neural Network (CNN) and Deep CNN have been used for image classification with very high accuracies. ANNs are systems inspired by the brain's biological neural networks. They are made up of a collection of artificial neurons which are interconnected. Each neuron can send signals to another connected neuron which can also process the signal received and transmit it after that. Neurons are arranged in layers, and each layer may perform a different type of transformation on the receiving input signal. Some ANNs are capable of backpropagation which allows the data to flow backward in the network to adjust the network's effectiveness. Deep Learning (Deep Neural Network) is based on ANN. The term 'deep' is derived from the multiple layers in the network. Figure 3 shows a Deep CCN architecture for facial Emotion Recognition.

The convolutional layers of CNNs extract feature from the input images for the other layers (i.e., pooling, dropout, and fully connected layers). The convolutional layer comprises small patches that use filter values to transform whole images and create feature maps. This is achieved using equation (1). Where, is the input image, is the filter, (is the size of the resulting matrix generated. Where, f is the input image, h is the filter, (m,n) is the size of the resulting matrix generated.

$$G[m,n] = (f * h)[m,n] = \sum_j \sum_k h[j,k] f[m-j, n-k] \quad (1)$$

DOI : http://dx.doi.org/10.25139/inform.v7i1.4282

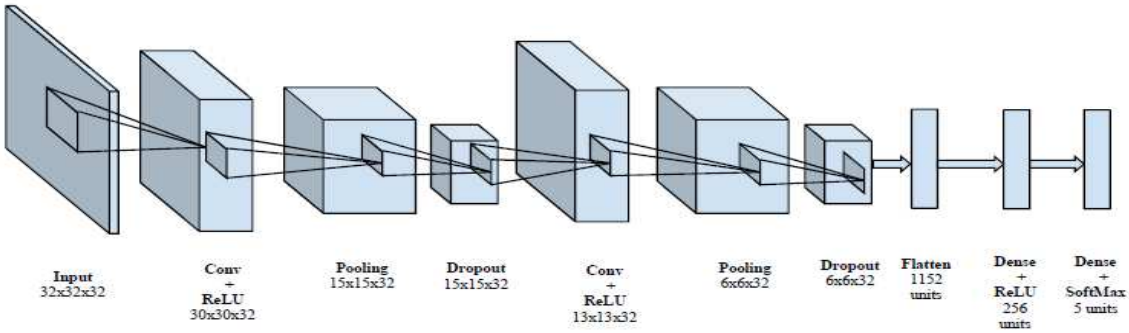

Figure 3. Deep CNN architecture for FER

The output data from the convolution layer, through a lossless transfer, is passed to the pooling layer, thus reducing the size. The resulting data is a 2-dimensional array converted to a single-dimensional vector using the flatten layer to be passed to the neural network for classification. The neural network backpropagates the errors of the network to adjust the weights, which in turn reduces the error (loss) function. The weight adjustment is made using equation (2) [5].

$$W_i = W_i + \Delta W_i \quad (2)$$

Where W_i is the weight and ΔW_i is given by the delta rule equation (3).

$$\Delta W_i = n \frac{dE}{dW_i} x_i \quad (3)$$

Where *n* variable is the learning rate,the E variable is the error function, and the x_i variable is the input.

C. *Dataset for Facial Emotion Detection Systems*

Datasets are available for these emotions. The details of the datasets are represented in Table I. CASIA-Face-Africa is a dataset that considers the demographic imbalance of the existing dataset, which tends to affect the performance of face biometric systems for African subjects. The dataset contains 38,546 images, out of which 1,183 subjects are Africans. It is used to study face biometrics, face image preprocessing, face feature analysis and matching, facial expression recognition, sex/age estimation, ethnic classification, face image generation, etc. The dataset is manually labeled with 68 landmark points to facilitate these facial landmark detection. The dataset set has a section with 70 subjects who acted out seven emotions (i.e., Neutral, Angry, Sad, Happy, Surprise, Fear, and Disgust) [13].

TABLE I
DATASETS FOR FACIAL EMOTION DETECTION [2]

Datasets	Description	Emotions Captured
CK+	640x490 or 640x480 grayscale or full-scale image sequences of frontal and 30-degree views. The dataset has 593 sequences from 123 subjects, with each sequence containing 10 to 60 frames.	Surprise, Happiness, Sadness, Disgust, Contempt, Fear. and Anger
JAFFE	The JAFEE dataset had 213 grayscale images made up of only Japanese females.	Surprise, Happiness, Sadness, Disgust, Fear. and Anger and Neutral
FER2013	The FER2013 comprises 35887 48x48 images collected from Google image search.	Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise
MultiPie	The dataset is comprised of almost 750,000 images. These images were captured by 15 views and 19 illumination conditions	Squint, Anger, Surprise Neutral, Happy, Scream, Disgust
MMI	It contained 2900 videos indicating the neutral, onset, apex, and offset	Surprise, Happiness, Sadness, Disgust, Fear. and Anger and Neutral
GEMEP FERA	The dataset comprised 289 image sequences	Happiness, Fear, Relief, Anger, and Sadness
SFEW	There were seven hundred images with different ages, occlusion, illumination, and head pose.	Surprise, Happiness, Sadness, Disgust, Fear. and Anger and Neutral
BU-3DFE	The dataset comprises 2500 3D facial images captured on two views, i.e., -45°, +45°	Surprise, Happiness, Sadness, Disgust, Fear. and Anger and Neutral
CASME II	This dataset is made up of 247 micro-expression sequences.	Regression, Disgust, Happy, Surprise and others
Oulu-CASIA	The Oulu-CASIA comprises 2880 videos captured in three different illumination conditions.	Surprise, Happiness, Sadness, Disgust, Fear. and Anger
AffectNet	AffectNet comprises more than 440,000 images collected from the internet.	Surprise, Happiness, Sadness, Disgust, Fear. and Anger and Neutral
RAFD-DB	There are 30000 images acquired from the real world, making up the RAFD-DB dataset.	Surprise, Happiness, Sadness, Disgust, Fear. and Anger and Neutral
RaFD	There are 8040 images with different face pose, age, gender, sexes making up the dataset.	Surprise, Happiness, Sadness, Disgust, Fear. and Anger, Contempt, and Neutral

DOI : http://dx.doi.org/10.25139/inform.v7i1.4282

III. RELATED WORK

In recent years researchers have used various architectures for facial emotion recognition with various levels of success. This session of this paper reviews the recent works done in this field using deep learning algorithms. This paper also seeks to map out trends and patterns in recent research and develop a proposal for future directions in the field of facial emotion recognition, as shown in Table II.

Pranav et al. [5] proposed a Deep Convolution Neural Network (DCNN) model which was built using the Keras Deep Learning Library to classify five human facial emotions. Their proposed model used two convolution layers with a pooling layer followed by a dropout layer after each convolution layer. ReLU (Rectified Linear Unit) used the activation function, which converted all negative values to zeros. A 2-dimensional array is created, which is then passed to the flatten layer to be converted to a single-dimensional vector and then passed to a two-layered network used to classify the emotions. For a probabilistic output, softmax is used as the activation function for the output layer. The model was trained and tested for 11 epochs with a learning rate of 0.01. The model had an accuracy of 78.04%. The dataset used was manually created using a 48 MP camera and contained a total of 2550 images, each having a pixel size of 1920x2560. The dataset was split into 2040 training images, 255 validation images, and 255 testing images. The emotions tested for were angry, happy, neutral, sad, and surprised.

An experiment FER system was developed, which is divided into three processes: preprocessing with the Viola-Jones algorithm, feature extraction with local fisher discriminant analysis (LFDA) for dimensionality reduction and k-nearest neighbors (KNN), and classification with a feed-forward artificial neural network (ANN) [6]. The dataset used for this experiment is JAFFE, but the researchers focused on four (4) emotions. Happy, Neutral, Surprised, and Sad out of the seven emotions present in the dataset. The two classifiers were compared in terms of performance. The 1NN algorithm performed better with sad and neutral emotions, while the ANN algorithm performed well with happy and surprise emotions, but for the average performance, the ANN algorithm outperforms the 1NN algorithm with an average performance of 66.66% as against 54.16%.

Jaiswal et al. [14] proposed a deep learning architecture of two different CNN networks for facial emotion detection. The proposed model uses Keras and has an input shape of 48*48*1 and two models with the same kernel size for feature extraction. The submodels are flattened into vectors and then concatenated into one long vector before transmitting to a fully connected softmax layer for classification. The performance of their architecture was evaluated using two datasets, FER2013 and JAFFE, and the accuracies realized were 70.14% and 98.65%, respectively, for the two datasets. The choice of datasets by the researchers was to make the model more robust in terms of diversity. Lasri et al. [15] also proposed a CNN architecture to recognize students' facial emotions. Their architecture consists of 4 convolution layers, each with maximum pooling layers and two fully connected layers. The proposed make use of softmax to predict facial emotions. The model was trained and tested using the FER2013 dataset. Their model also scored 70% on accuracy.

Naik and Mehta [16] proposed a CNN model that is used for FER in a hand-over-face situation called Hand-over-Face Gesture-based Facial Emotion Recognition Method (HFG_FERM). Hand-over-face is typically considered in other facial emotion recognition as an instance of occlusion, and as such, images with hand-over-face are exempted from the experiments. The proposed research provides extensive coding schemas with additional hand signs that help identify unexplored emotions such as confidence, making a decision, and scared, ashamed, angry, and ok signs along with basic emotions. Their model identified emotions in extreme hand occlusion and extreme head rotation cases. The authors validated their model with images from the Cam3d corpus, FER2013 dataset, and public domains summing up to a total of 18 emotion categories. The performance of their model was compared with two other models, Multimodal Fusion Approach (MMFA) and Emotion Recognition through Facial Gestures (ERTFG). From their experiments, their model HFG_FERM outperforms MMFA and ERTFG on different levels.

The researchers in [17] proposed different architectures of Convolution Neural Networks (CNNs) of two models to classify seven facial emotions. The first CNN model had four convolutional layers, four max pooling, one dropout, and two fully connected layers. The second model used the same model as the first but with data augmentation. The dataset used in this experiment is iCV MEFED (Multi-Emotion Facial Expression Dataset). Their choice of the dataset was relatively new and had compound/mixed emotion, e.g., angrily surprised, tears of joy. The dataset contains 5,750 images, each of size 5184x3456 pixels. It was shown that the model performed better with images that are not distorted than with augmented images. It was also revealed that emotions such as sadness and contempt are under-predicted compared to the other emotions.

Bouzakraoui et al. [18] proposed a model that automatically detects facial expressions displayed by clients when they react to a product or service. First, the researchers extracted geometric features from the customers' faces and then used adapted SVM to predict the customers' satisfaction. The image is converted to geometric primitives such as points and curves during the geometric feature extraction by measuring relative distances between distinct features like the eyes, eyebrows, nose, mouth, and chin. A vector of 19 values that represents the customer's facial expression from these distances is generated. The dataset used for this experiment is the JAFFE dataset. The researchers reclassify the emotion in the dataset into three classes; satisfied, not-satisfied and neutral.

Jain et al. [19] proposed a model based on a single Deep Convolutional Neural Networks (DNNs). The proposed model consists of six convolution layers, two deep residual blocks,

DOI : http://dx.doi.org/10.25139/inform.v7i1.4282

and two fully connected layers, each with a ReLU as activation function and dropout. These features of the model help the model to learn subtle features which are related to certain emotions. Two residual blocks contain four convolution layers with varying sizes, two short connections, and one skip connection. For the classification of emotion, softmax is used. Their proposed model was trained and tested on two datasets, CK+ and JAFFE, and it was found that the combination of fully connected networks and residual block improved the overall performance of the proposed model.

Researchers in [20] proposed a facial expression recognition based on the Valence-Arousal dimensional emotion model. The proposed model uses a valence dimension prediction which has nine levels. The proposed model implores the use of CNN to predict a result equal to the weighted fusion of valence value and its corresponding probability. The CNN architecture consists of 4 convolution layers, each with ReLU as activation function, three maximum pooling layers placed after convolution layers 2, 3, and 4 respectively, and Two (2) fully connected layers. The model uses softmax for the classification. The proposed model was trained using CK+ and FER2013 datasets. The researchers used ten annotations to average the images of the training set and the test set based on the SAM system; each picture is annotated with a valence dimension of 1-9. The performance was verified by letting the system recognize the facial expressions of volunteers when watching a video.

Researchers in [21] a Facial Emotion Recognition model with attention-based ACNN performs well for a partial block on the face. Occlusion occurs when there is a hand, hat, hair, or anything covering part of the face. This makes it harder for a FER system to read the emotions. The model proposed tackles occlusion by focusing and interoperating the emotions from the symmetric part of the face known as the informative regions. The proposed model implores two variations of ACNN to detect the block region and the region with emotion. To recognize the blocked section of the Path-Gated Unit (PG-unit) face is used, whereas recognition of full-face region makes use of Global-Gated Unit (GG-unit). The attention unit learns the scalar weights adaptively for the patch region. However, the ACNNs detect emotion by facial landmarks. The proposed model was trained and tested using Indian Spontaneous Expression Database (ISED), which has four (4) categories (disgust, happy, sad, and surprise)

Navaz et al. [22] proposed that preprocessing can increase the performance of deep learning architecture. The techniques they proposed in preprocessing stage are; image data handling, where images are checked for wrongful labeling using a Perl file that segregates the images into the corresponding folders. They used another script to move images with a low resolution into a specific folder. The next preprocessing technique is quality improvement, where the images resolution quality is improved by bicubic interpolation, which is done after using the Very-Deep Super-Resolution algorithm. After preprocessing, the resulting data was passed through a CNN architecture for facial emotion recognition. The dataset used in this experiment was the Indian Movie Face Database (IMFDB). It was evident that the preprocessing increases the Facial Emotion Recognition performance. Researchers in [23] also proposed a model of CNN architecture with improved preprocessing and feature extraction for FER. First, in preprocessing, researchers cropped the unwanted part of the image so that only the face is shown. The image is then resized to fit the input size of the CNN. After that, intensity normalization is done using MinMax normalization to correct image brightness and contrast. A hybrid method of CNN and Histogram Oriented Gradients (HOG) is used for feature extraction. The model was trained and tested using the JAFFE and FER2013 datasets.

A face emotion recognition algorithm that makes use of Gabor filters and CNN to detect facial expressions. [24]. The model used two Gabor filters for feature extraction, with the first output as the input to the second. A Gabor filter gives the highest response at the edges and at points where texture changes. After applying the filter to the image, the features for facial emotion detection like the eyebrows shape, eyes, nose, and mouth are highlighted. The image is passed to the CNN architecture after applying the filters for classification. The CNN has three convolutional layers with ReLU activation function and MaxPooling, and a flatten layer, two dense layers, a dropout layer, and a softmax. The proposed model was tested using the JAFFE dataset and attained 97% accuracy at 25 epochs. At the end of the experiment, it was realized that their proposed model was faster than other CNN models.

Researchers in [25] proposed using Multi-Task CNN (MTCNN) for face detection and the use of ShuffleNet V2 for emotion recognition. ShaffleNet architecture involves two operations different from the usual CNN: the pointwise group convolution and channel shuffle. The group convolution distributes the convolution over different CPUs for parallel separable convolution operations. This can be very costly in terms of computational complexity. Therefore, the architecture features the channel sparse connect. The model achieved an accuracy of 71.19% on the FER2013 dataset.

Researchers in [26] realized that challenges in unimodal emotion detection systems make them less accurate. Therefore, they proposed a multimodal emotion recognition system using complementary emotional information from facial expression and speech. The proposed model consists of a 1D CNN and bi-directional LSTM to extract acoustic features from speech, and a 2D CNN to extract high-level features from facial expressions. The joint classification is done using SoftMax. The proposed model was trained and tested using IEMOCAP, a dataset containing 12 hours of audio-visual data. The team combined happy and excited, and the resorting dataset had four categories: excited, angry, sad, and neutral. Compared with single modal speech emotion recognition and facial expression recognition, the proposed model proved a 10.05% and 11.27% increase in accuracy, respectively.

DOI : http://dx.doi.org/10.25139/inform.v7i1.4282

Researchers in [27] also proposed a multimodal emotion recognition system. The model proposed here consisted of four neural networks to extract features from the audio, facial, and gesture dataset. The model explored the grid-search strategy for performance optimization on the validation set. The model consists of two sub-models for facial expression: feature embedding and frame attention models. The feature embedding model is a deep CNN that generates a feature vector from the face image. The frame attention model learns the weights and flexibly aggregates the feature vectors to form a single discriminative video representation. The researchers used the Temporal Shift Module (TSM) for the body movement and gesture model. TSM gives high efficiency and high performance. The researchers used openEar software for the audio model to extract features and then performed the classification task. The researchers then used a weighted sum to fuse the scores from each model. The best overall accuracy of the experiments was 76.43%.

The authors of [28] did a performance analysis of Transfer Learning as against training from scratch for deep facial expression recognition. Two networks, Alexnet and VGG16, were used in this comparison. Four training stages were performed; training Alexnet from scratch, training Alexnet using transfer learning, training VGG16 from scratch, and training VGG16 using transfer learning. Alexnet is a famous CNN architecture that can classify images into 1000 object categories and train with the ImageNet dataset. It contains five convolutional layers, three pooling layers, two dropout layers, seven activation layers, and three fully-connected layers. Like Alexnet, VGG16 is also a CNN architecture trained with ImageNet to classify 1000 objects. However, its network structure is different, with 13 convolutional layers, five pooling layers, two dropout layers, 15 activation layers, and three fully-connected layers. The researchers remodeled the two architectures for the training from scratch and got the pre-trained models for transfer learning. All the models were tested with the RaFD dataset. The experiment showed that Alexnet and VGG16 achieved 95% and 95.33%, respectively, for transfer learning against 84% and 16.67% for the training from scratch. However, the researchers attributed the low performance of the VGG16 training from scratch to insufficient training data since the network is too extensive and imbalance in the input data.

TABLE II
EXISTING DEEP LEARNING TECHNIQUES FOR FER

Authors	Datasets	Algorithms	Contribution	Performance	Limitations
Pranav et al. [5]	2550 manually-taken images with 5 classes	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	Classification of 5 human facial emotions using DCNN, which was built using the Keras DL Library	Validation accuracy of 78.04%	
Ranjan and Sahana [6]	4 emotions out of the JAFFE datasets	LFDA for dimensionality reduction, 1NN and Feed-forward ANN for classification	Classification of 4 human facial emotions	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
Jaiswal et al. [14]	FER2013, JAFFE	CNN architecture using the Keras Library with two similar parallel sub-models. Each sub-model contains conv layer, local contrast normalization, max pooling, another conv layer, man pooling and flatten layer. The two models are concatenated at the output, which uses softmax as AF.	Classification of 7 facial emotions	70.14% on FER2013, 98.65% on JAFFE	
Lasri et al. [15]	FER2013	CNN model with 4 convolutional layers, 4 pooling layers, and 2 fully connected layers. The model uses softmax at AF for the output.	Automatic system that analyses student's facial expression in relation to a teacher's presentation.	70% validation accuracy	The model, in some cases, wrongly predicts fear as the sad face.
Naik and Mehta [16]	Images from Cam3d corpus, FER2013 and public	Hand-over-Face Gesture-based Facial Emotion Recognition Method (HFG_FERM). The CNN model contains two conv layers, each followed by a max-pooling layer. The output uses softmax as the AF	Facial emotion detection in extreme hand occlusion and head rotation. Hand-over-face gestures help detect other emotions like confident, thinking/ deciding, ashamed, and basic emotions (in total 18 emotion categories).	The proposed model outperforms MMFA and ERTFG.	

DOI : http://dx.doi.org/10.25139/inform.v7i1.4282

Authors	Datasets	Algorithms	Contribution	Performance	Limitations
Begaj et al. [17]	iCV MEFED	Two CNN models; the first model had four conv layers, four max pooling, one dropout, and two fully connected layers. The second model is the same as the first model with data augmentation	Classification of mixed/compound emotions such as tears of joy and angrily surprised.	The first model performs better than the second model.	The model, in some instances, confused fear with surprise and contempt with sadness.
Bouzakraoui et al. [18]	JAFFE dataset	Supervised SVM with input as geometric facial features which is computed by relative Euclidean distances between landmark points.	Customer satisfaction recognition using facial expressions. Classification into three classes; satisfied, non-satisfied, and neutral	SVM model attained a global Receiver Operating Characteristic (ROC) curve with area = 0.92	
Jain et al. [19]	CK+, JAFFE	DNNs which consists of 6 conv layers, 2 residual blocks and two fully connected layers each with ReLU as AF. Each Residual block contains 4 conv layers with varying sizes, two short connections and one skip connection	Classification of images into 6 facial emotion classes.	95.23% accuracy on JAFFE dataset and 93.24% accuracy on CK+	
Liu et al. [20]	CK+, JAFFE	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	Distinguishing 3 emotional categories on the Valence-Arousal dimension	The model achieved an RMSE index of 0.0857 ± 0.0064	There is some over-fitting in the training and testing phase, possibly due to the insufficient number of facial images with valence dimensions.
Engoor et al. [21]	ISED	Two variations of attention-based ACNN, one to detect blocked region of the face and the other to detect the region with the emotion.	Classification of 4 classes of emotions with instances of partial occlusion on the face from a video dataset.	Not deterministic	Model is limited to video datasets but not video streams due to latency for the later.
Navaz et al. [22]	IMFDB	Transfer learning using Pre-trained networks (Alexnet, GoogLeNet, and another CNN), and using AutoML	Very extensive image preprocessing; checking wrong labeling with a Perl script and Using VDSR to improve the resolution of the images to achieve a better performance,	Improved image quality generally improves performance. AutoML has $\approx 20\%$ prediction accuracy better than pre-trained networks	
John et al. [23]	JAFFE, FER2013	CNN architecture with improved preprocessing and feature extraction.	Preprocessing methods such as face cropping, intensity normalization, and feature extraction is used to increase the accuracy of the image.	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
Taghi Zadeh et al. [24]	JAFFE	CNN of 3 conv layers with ReLU as activation function and max pooling, a flatten layer, two dense layers, a dropout layer, and softmax as the activation function for the output layer	The use of two Gabor filters for feature extraction for higher accuracy	97% accuracy at 25 epochs	
Ghofrani et al. [25]	FER2013	Multi-Task CNN (MTCNN) for face detection and Shuffle Net V2 for emotion recognition.	Using a three-staged MTCNN to accurately narrow down on the face area, thereby reducing the work done during facial emotion detection	71.19% accuracy	
Cai et al. [26]	IEMOCAP	1D CNN and bi-directional LSTM for extracting acoustic features and 2D CNN for Facial expression. Joint classification is done using softmax.	Multimodal emotion recognition system using complementary emotional information from facial expression and speech for classification of 4 emotions	10% increase in accuracy for speech emotion recognition and 11.27% increase for facial emotion recognition	
Wei et al. [27]	Not stated	CNN feature embedding and frame attention for facial emotion detection, Temporal Shift Module (TSM) for body movement and gestures, and openEAR for extraction and classification of audio emotions	Recognition of face, audio, and gesture expressions via multimodal emotion recognition.	76.43% overall accuracy	

DOI : http://dx.doi.org/10.25139/inform.v7i1.4282

Authors	Datasets	Algorithms	Contribution	Performance	Limitations
Oztel et al [28]	RaFD	Alexnet and VGG16 were trained from scratch and trained using transfer learning.	Performance analysis of Transfer Learning and training from scratch in the area of deep facial expression recognition.	95% and 95.33% accuracy for Alexnet and VGG16 respectively for transfer learning and 84% and 16.67% for training from scratch	

IV. RESULT AND DISCUSSION

The problem of facial emotion recognition is an area of interest in HCI and Affective Computing, and as such, some amount of work has been done, and there is still a lot of work being done. However, some challenges need more input.

Firstly, there are no large datasets that can train huge networks. Designers of deep learning architectures have to stick to a small-scale network because there are no datasets large enough to train very deep networks. It cannot go unnoticed that a dataset of that magnitude is almost impossible to get since it will take a lot of funds to get posed images dataset and a lot of time and effort to generate a dataset of that magnitude in the wide. The other issue is that it will take a system of equally higher hardware specifications to train such a network.

Success at human facial emotion recognition for computer systems is one thing. Still, human-to-human emotion detection is a huge affair because humans can recognize emotional cues in many ways, from facial expression, voice intonation, choice of words, body language, and written text. There is also a growing concern of people leaving a standard double life especially on the internet, which is termed as 'posing for the cameras. Solving issues like that demand a lot of multimodal training. While there is some work being done in that field, a lot of other works are required to bridge that gap.

As discovered whiles putting this paper together, transfer learning has a lot to offer when it comes to achieving a large-scale holistic emotion detection and recognition for HCI. A lot of research has been undertaken in various fields in affective computing, and these works have produced very fine architectures in terms of computational complexity and performance in their respective emotion recognition fields. Future research in affective computing can focus on using transfer learning to train existing networks.

V. CONCLUSION

In this paper, we had a look at the problem of emotion detection, specifically facial emotion detection, which has prevalently become relevant in the field of HCI. We also exposed the available datasets for this cause and the use of deep learning models in solving the problem. A Series of research works done in the field of facial emotion was analyzed, looking at the various architectures, tools, and algorithms and the datasets used. We finally discussed the open issues and the future trends as far as facial emotion recognition and emotion recognition are concerned. In general, it was discovered in your research that ANNs, specifically CNNs, are currently leading in the architectures or FER system because they produce better results. It is also observed that some researchers turn to improve on preprocessing techniques and feature extraction to achieve better results. Lastly, we also recorded a growing trend in multimodal systems when including other emotional cues like ECG and verbal cues to produce a more accurate Emotion detection system. Although facial emotion detection has come a long way, the systems are still limited with still images, which is good for the ideal case. Still, human emotions can be very complex, and real-world scenarios would require techniques to capture continuous, spontaneous, and subtle facial expressions. It is also noticed that some emotions are easier to train because of the abundance of their training samples.

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',

DOI : http://dx.doi.org/10.25139/inform.v7i1.4282

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, 'CASIA-Face-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 Hand-over-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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

DOI : http://dx.doi.org/10.25139/inform.v7i1.4282