IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Facial Sentiment Analysis using AI Techniques: State-of-the-Art, Taxonomies, and Challenges

KEYUR PATEL[1], DEV MEHTA[1], CHINMAY MISTRY[1], RAJESH GUPTA[1], SUDEEP TANWAR[1], NEERAJ KUMAR[2,3,4], MAMOUN ALAZAB[5]

[1]Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India (e-mails: 17bce080@nirmauni.ac.in, 17bit047@nirmauni.ac.in, 17bit052@nirmauni.ac.in, 18ftvphde31@nirmauni.ac.in, sudeep.tanwar@nirmauni.ac.in)
[2]Department of Computer Science Engineering, Thapar Institute of Engineering and Technology, Deemed to be University, Patiala, Punjab, India (e-mail: neeraj.kumar@thapar.edu)
[3]Department of Computer Science and Information Engineering, Asia University, Taiwan
[4]King Abdul Aziz University, Jeddah, Saudi Arabia
[5]College of Engineering, IT & Environment, Charles Darwin University, Casuarina, NT 0810, Australia (e-mail: mamoun.alazab@cdu.edu.au)

Corresponding author: Mamoun Alazab (mamoun.alazab@cdu.edu.au), Neeraj Kumar (neeraj.kumar@thapar.edu).

This work was supported by the Department of Corporate and Information Services, NTG of Australia.

ABSTRACT With the advancements in machine and deep learning algorithms, the envision of various critical real-life applications in computer vision becomes possible. One of the applications is facial sentiment analysis. Deep learning has made facial expression recognition the most trending research fields in computer vision area. Recently, deep learning-based FER models have suffered from various technological issues like under-fitting or over-fitting. It is due to either insufficient training and expression data. Motivated from the above facts, this paper presents a systematic and comprehensive survey on current state-of-art Artificial Intelligence techniques (datasets and algorithms) that provide a solution to the aforementioned issues. It also presents a taxonomy of existing facial sentiment analysis strategies in brief. Then, this paper reviews the existing novel machine and deep learning networks proposed by researchers that are specifically designed for facial expression recognition based on static images and present their merits and demerits and summarized their approach. Finally, this paper also presents the open issues and research challenges for the design of a robust facial expression recognition system.

INDEX TERMS Facial Sentiment Analysis, Machine Learning, Deep Learning, Convolutional Neural Network, Deep Belief Network, Artificial Intelligence.

I. INTRODUCTION

Emotions are efficacious and self-explanatory in normal day-to-day human interactions. The most noticeable human emotion is through their facial expressions. The Facial Expression Recognition (FER) is quite complex and tedious but helps in various applications areas such as healthcare [1]–[3], emotionally driven robots, and human-computer interaction. Although the advancements in FER increases its effectiveness, achieving a high accuracy is still a challenging task [4]. The six most generic emotions of a human are anger, happiness, sadness, disgust, fear, and surprise. Moreover, the emotion called *contempt* was added as one of the basic emotions [5].

FER is a baffling task and its accuracy is completely dependent on the parameters selected, such as illumination factors, occlusion, i.e., obstruction on the face like hand, age, and sunglasses. Researchers of the field are taking these parameters into consideration while making their FER models so that the considerable accuracy can be achieved. The description of some important factors for FER is as follows.

- *Illumination factor:* The light intensity falling on the object affects the classification of the model. The textural values increase the false acceptance rate due to either by minimizing the distance between classes or by increasing the contrast [6].
- *Expression Intensity:* The expression recognition is highly dependent on the intensity of the expression. The expression is recognized more accurately when the expression is less subtle. It highly affects the accuracy of the model.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

TABLE 1: Nomenclature

AI	Artificial Intelligence
ANN	Artificial Neural Network
BN	Batch Normalization
CK	Cohn-Kanade Dataset
CK+	Extended Cohn-Kanade Dataset
CNN	Convolutional Neural Network
CV	Computer Vision
DAGSVM	Directed Acyclic Graph Support Vector Machine
DBN	Deep Belief Network
DCT	Discrete Cosine Transform
DL	Deep Learning
FC	Fully Connected
FER	Facial Expression Recognition
FSA	Facial Sentiment Analysis
GRNN	General Regression Neural Network
HOG	Histogram of Oriented Gradients
IL-CNN	island loss Convolutional Neural Network
JAFFE	Japanese Female Facial Expression Database
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LLE	Locally Linear Embedding
MAC	Multiply-Accumulate
ML	Machine Learning
MSCNN	Multi-Signal Convolutional Neural Network
MUG	Multimedia Understanding Group
NLDA	Zero/Null Space Linear Discrim-inant Analysis
P-SIFT	Parallel Scale-Invariant Feature Transform
PCA	Principal Component Analysis
PReLU	Parametric rectified linear unit
RaFD	Radboud Faces Database
RBF	Radial Basis Kernel Function
RBM	Restricted Boltzmann Machine
ReLu	rectified linear unit
ResNet	Residual Networks
RNN	Recurrent Neural Network
SVM	Support Vector Machine
TFD	Toronto Face Database
ULTP	Uniform Local Ternary Pattern
VGG	Visual Geometry Group

- *Occlusion:* If occlusion is present on an image then it becomes difficult for the model to extract features from the occluded part due to inaccurate face alignment and imprecise feature location. It also introduces noise to outliers and the extracted features.

FER systems can be either static or dynamic based on image. Static FER considers only the face point location information from the feature representation of a single image, whereas, the Dynamic Image FER considers the temporal information with continuous frames [7], [8]. The static FER process over is exhibited in FIGURE 1 with description of steps as follows.

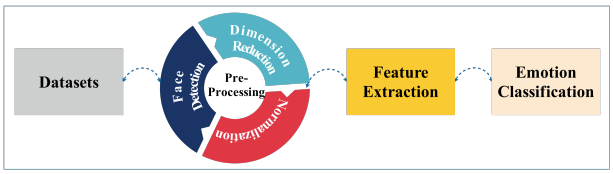

FIGURE 1: The general pipeline of FER systems

- *Dataset*: To avoid over-fitting, the following FER algorithms are discussed, which needs extensive training data. A dataset must have well-defined emotion tags of facial expression is essential for testing, training, and validating the algorithms for the development of FER. These datasets contain a sequence of images with distinct emotions, as mentioned above. We have reviewed many datasets to train different models for real-world profits. Table 4 provides an overview of different datasets available for FER.
- *Pre-Processing*: This step pre-processes the dataset by removing noise and data compression. Various steps involved in data pre-processing are: (i) *facial detection* is the power to detect the location of the face in any image or frame. It is often considered as a special case of object-class detection, which determines whether the face is present in an image or not, (ii) *dimension reduction* is used to reduce the variables by a set of principal variables. If the number of features is more, then it gets tougher to visualize the training set and to work on it. Here, PCA and LDA can be used to handle the aforementioned situation. (iii) *normalization:* It is also known as feature scaling. After the dimension reduction step, reduced features are normalized without distorting the differences in the range of values of features. There are various normalization methods, namely Z Normalization, Min-Max Normalization, Unit Vector Normalization, which improves the numerical stability and speeds up the training of the model.
- *Feature Extraction*: It is the process of extracting features that are important for FER. It results in smaller and richer sets of attributes that contain features like face edges, corners, diagonal, and other important information such as distance between lips and eyes, the distance between two eyes, which helps in speedy learning of trained data.
- *Emotion Classification*: It involves the algorithms to classify the emotions based on the extracted features. The classification has various methods, which classifies the images into various classes. The classification of a FER image is carried out after passing through pre-processing steps of face detection and feature extraction. Various classification techniques are discussed later in the proposed survey.

The FER system has various applications such as computer-human interactions, healthcare system [9]–[14], and social marketing. In the proposed survey, we analyze the existing surveys pertaining to different approaches of FER proposed by the authors globally. We compare the surveys and develop a taxonomy on various pre-processing, feature extraction, and emotion classification steps. We also discuss the various open issues and future research challenges related to FER.

A. MOTIVATION

Paul Ekman first coined the term FER in the mid-1980s. Since then, various machine learning techniques like random forest classifiers, artificial neural networks, etc. were used

by the researchers to recognize the seven basic emotions. They also claimed good and effective results. Automated human emotion detection is all-important in security and surveillance applications these days. To further improve its performance, the researchers are trying hard to explore further in this field. Various challenges like occlusion in datasets, over-fitting of models, etc. have to be taken care of while implementing the FER. As per the literature explored and knowledge of the authors, no survey is available, which exhaustively compares the FER approaches from the perspective of AI. Motivated from the aforementioned fact, we present a comprehensive survey on FER using Artificial Intelligence (AI) techniques in which we have explored the state-of-the-art machine learning and DL (DL) approaches with their merits and demerits.

B. SCOPE OF THE SURVEY

Facial sentiment analysis is the most trending topics in Computer Vision area. A lot of literature has already been published by researchers across the globe in this field, but still, many researchers are trying to solve the challenges and issues in FER. Various surveys have been published in recent years [7], [15], [26], [27] on sentiment analysis. These surveys have mainly focused on traditional methods like support vector machine (SVM), decision tree classifiers, and artificial neural network (ANN). The DL methods [28], [29] have rarely been explored by the researchers working in the same field. So, in this paper, we analyzed the surveys on facial sentiment analysis and presented a comparative analysis. For example, Hemalatha *et al.* [15] surveyed various methods for facial detection, facial feature extraction and classification of FER, but not presented the proper comparison of methods considered and the dataset used. Later, the authors in [16] also presented the survey on FER, but they had not mentioned anything about datasets useful for emotions recognition. Another survey of Chengeta *et al.* [19] was on various traditional feature extraction techniques like principal component analysis (PCA), Linear Discriminant Analysis (LDA), and Locally Linear Embedding (LLE), and thereafter they proposed an ensemble classifier. They failed to compare the advanced DL approach, which is currently the most novel approach in FER. Again, Bhaskar *et al.* [18] also lacks in explaining various DL approaches.

Recently, DL-based FER approaches has been explored in [7], [27], which are the detailed surveys without the discussion on FER. Therefore, in the proposed survey, we make a systematic survey of various databases used for FER, various methods for face detection, facial feature extraction, and emotion classification, future challenges, and current issues in facial sentiment analysis. Our aim for this survey that it would be quite beneficial for those who want to explore in this field and they will get a complete overview of all the advanced systematic approaches in facial sentiment analysis. Table 2 presents relative differences between the existing surveys with the proposed survey.

C. RESEARCH CONTRIBUTIONS

In this paper, we surveyed various existing literature on Facial Sentiment Analysis focusing on the DL techniques, datasets, and the methodologies used to classify emotions. Following are the crisp contributions of the paper.

- We present an in-depth survey on FER methods and dataset used. Then, we highlight the advanced methods used for FER and their comparative analysis.
- We present a taxonomy on FER methods based on face detection, feature extraction, and emotion classification.
- Finally, we presented the open issues and research challenges in the Facial Sentiment Analysis.

D. ORGANIZATION

Structure of the survey is as shown in FIGURE 2. Section II focuses on the evolution of facial recognition techniques presented by the authors across the globe and the dataset used. It also describes the need for facial detection, dimension reduction, normalization, feature extraction, and emotion classification. In Section III, we highlighted the bibliometric analysis and methodology used for conducting the proposed survey. In Section IV, we discuss various facial expression databases available for analysis. Section V discusses the proposed taxonomy (facial sentiment analysis taxonomy). In Section VI, we discuss the open issues and research challenges of FER, and finally, Section VII concludes the survey. Table 1 lists all the acronyms used in the paper.

II. BACKGROUND

This section focuses on the background and importance of facial expressions for sentiment analysis. It is bifurcated into four subsections. Firstly, we discuss the evolution of the timeline of facial recognition methods. Secondly, we discuss the need for Facial Detection, Dimension Reduction, and Normalization for sentiment analysis. In the third subsection, we focus on the need for feature extraction from the face image. Finally, we highlight the need for emotion classification.

A. EVOLUTION TIMELINE

Figure 3 gives a brief overview on the evolutionary timeline of facial sentimental recognition methods given by the researchers across the globe along with the datasets. There exists various algorithms for FER such as traditional state-of-the-art algorithms and DL-based algorithms proposed by various researchers till 2020. The emotion recognition was first stated in the paper proposed by Bassili *et al.* [30] in 1978 where authors have classified the emotions into six basic gestures such as happiness, sadness, fear, surprise, anger, and disgust. Different algorithms (traditional and DL) were used for FER by the authors. For example, Padgett *et al.* [31], the first time (ANN) in 1996, SVM [32] in 2000, CNN [33] in 2003, Multi-SVM [34] in 2006, boosted DBN [34] in 2014, RNN [35] in 2015, and (PHRNN and MSCNN) [36] in 2017. Also, many datasets have been created for training ans testing these FER models. The time-line shows the list of datasets as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

TABLE 2: A relative comparison of the proposed survey with the existing FER surveys.

Author	Year	Objective	Merits	Demerits
Hemlatha *et al.* [15]	2014	Analyzed various facial detection, facial feature extraction, and classification methods of FER	Detailed review on few FER approaches	No proper comparison between the methods discussed
Deodhare *et al.* [16]	2015	Given a detailed survey on various approaches for FER based on computational paradigms	Reviewed different methods that are being used in day-to-day life	Did not explain how classification takes place using different algorithms like CNN, SVM, etc.
Asad *et al.* [17]	2017	Presented a survey on various latest developments in FER domain	Reviewed various related works and presented a comparison among them	Detailed information about various DL techniques, their pros and cons, related issues and challenges were not explained
Bhaskar *et al.* [18]	2018	To analyze different machine learning algorithms for FER	Surveyed 3 ML algorithms deeply and how their characteristics helped in FER	Not reviewed advanced DL techniques like CNN,RNN etc
Chengeta *et al.* [19]	2018	To Review the traditional feature extraction methods and to propose an ensemble classifier	Reviewed feature extraction methods like LBP and LDP, and proposed the ensemble classifier to improve results	No proper focus on novel DL techniques for FER
Rajeswari *et al.* [20]	2018	To study and survey various FER techniques	Different detection, extraction and classification techniques were discussed.	Lack of proper discussion regarding FER techniques, future challenges, current issues and datasets
Martinez *et al.* [21]	2019	To survey various detection, extraction and machine learning classification methods for FER	Provided a systematic survey on various techniques and focused exclusively on automatic AU analysis from RGB imagery.	DL techniques were not discussed.
Bhattacharya *et al.* [22]	2019	To review and analyze several challenges for pose, illumination and age invariance for FER	Explored various challenges for pose, illumination and age invariance in FER	Only CNN was discussed as DL approaches
Vyas *et al.* [23]	2019	To survey various FER techniques based on CNN	Detailed survey of various CNN architectures used for FER on various datasets	Issues and Challenges in CNN were not discussed
Li *et al.* [24]	2020	To provide a detailed survey of various AI techniques for FER	Reviewed various deep neural networks for FER and also provided information about datasets and future challenges and issues	Not provided taxonomy for the FER system
Fathima *et al.* [25]	2020	To present a literature summary of the various ML Techniques used in FER	Provided a comparative study of various ML pre-processing, feature extraction and classification techniques for FER	Not provided taxonomy, future challenges for FER system. Also, very few FER datasets were discussed
Proposed Survey	2020	To explore current state-of-art AI techniques for FER	Surveyed various AI approaches, their pros and cons, FER datasets, current issues and challenges in detail	-

per the creation year. Datasets are- JAFFE [37] in 1998, CK+ [38] in 2000, MMI [39] in 2002, Oulu-CASIA [40] in 2008, Multi-PIE [41] in 2009, (RaFD [42], MUG [43], and TFD [44]) in 2010, and FER-2013 [45] in 2013. As new datasets are being available supported with DL algorithms to solve the challenges in FER.

B. NEED FOR FACIAL DETECTION, DIMENSION REDUCTION AND NORMALIZATION

In the FER process, the first pre-requisite step is face detection, which involves the detection of a face in the image or frame and removes the insignificant pixels. The face detection algorithm gives the output in the form of coordinates of the bounding box, which is put over the face. Detecting a face is quite complex as the human faces can be in different sizes and shapes. So, the face detection algorithm plays a vital role in the aforementioned situation. Various algorithms for face detection are available such as Viola-Jones [46], PCA, LDA, and genetic algorithms. Viola-Jones algorithm is one of the most widely used algorithms for face detection. It differentiates faces from the non-faces. PCA is the other most widely used face detection method. It is used to reduce the image dimensions and has four main parts:- feature covariance, eigen decomposition, principal component transformation, and choosing components [47]. Reducing the dimensions from $m - dimensions$ to $n - dimensions : \forall\ m > n$ does not means we are losing the properties of the image, moreover it preserves [48]. After the dimension reduction, normalization can be used to scale-up the image.

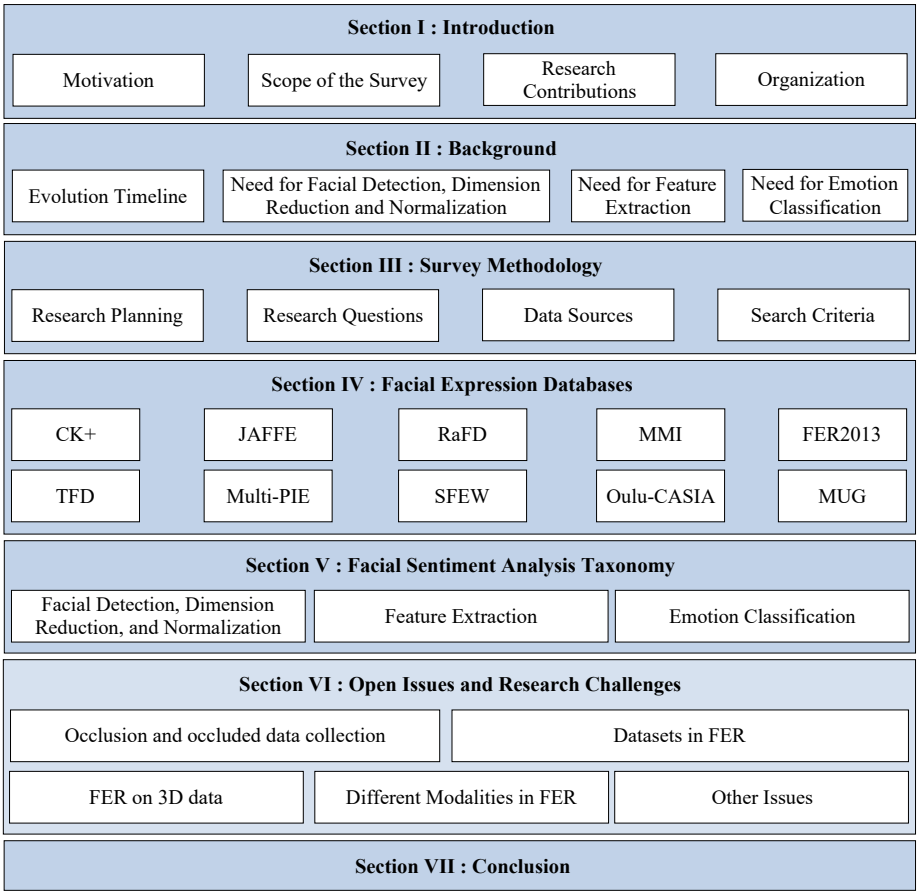

FIGURE 2: Roadmap of the survey

C. NEED FOR FEATURE EXTRACTION

Facial Expression analysis comprises of various methods such as facial landmark identification, feature extraction, and different feature extraction databases. Facial landmarks are drawn by the facial key points which are derived from the geometry of the face [16]. Feature Extraction is done after preprocessing phase [49]. There are two methods available for feature extraction are appearance-based extraction and geometric-based extraction. The geometric-based method extracts feature like edge features and corner features. Verma *et al.* [50] analyzed the performance of the feature extraction technique Gabor filter. They also tested the average gabor filter and compared both the filtering techniques to enhance the recognition rate.

- *Corners*: Corners of an image is a significant property, which can be inferred from the complex objects of the image. Cho *et al.* [51] developed the corner detection technique, which measures the distance and angle between two straight lines.
- *Edges*: They are one-dimensional features that represent the boundary of an image region.

The second method, which is an appearance-based method, takes care of the states of different points of the face, such as the position of the eye, shape of important points such as mouth and eyebrows using the salient point features. The majority of the traditional methods have used Local Binary Pattern (LBP) as the feature extraction technique, which is a generic-based framework for the extraction of features from the static image. It converts the most important features of the input image, as mentioned above, into a histogram [52].

D. NEED FOR EMOTION CLASSIFICATION

The third step in the FER is the Emotion Classification. There are various methods that are used for the classification of emotions after applying face detection and feature extraction algorithms. The various classification algorithms are convolutional neural network (CNN) [53], SVM, and restricted boltzmann machine (RBM). The most widely used method for classification is CNN. It is the most efficient algorithm as it can be applied directly to the input image without applying any feature extraction and face detection algorithms and still gets better accuracy over the input data [54]. The number of images in the training data set also has a huge impact on classification results. CNN faces a huge challenge in the training of limited-image dataset. So, the models which are built on a limited dataset can use the SVM algorithm for feature extraction and face detection. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

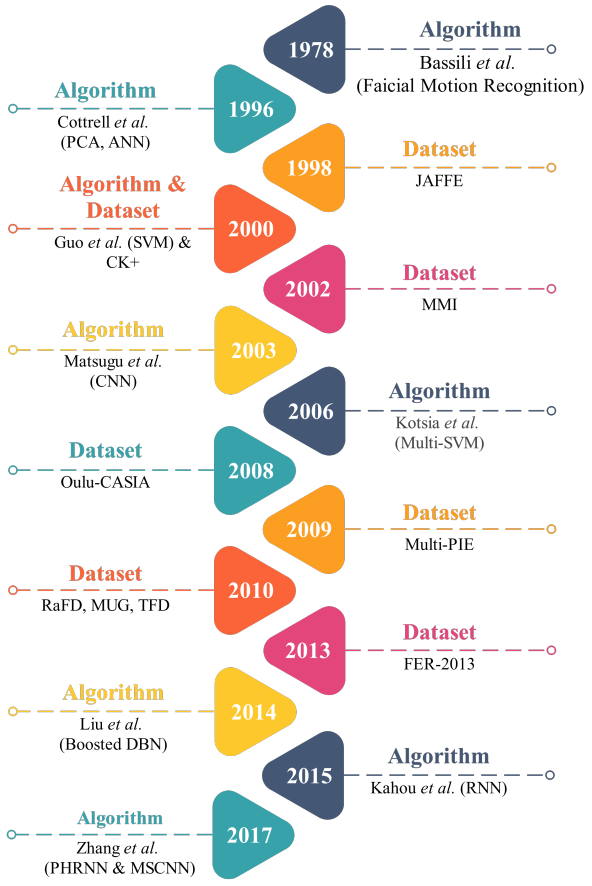

FIGURE 3: Evolution of facial recognition methods and datasets.

TABLE 3: Research questions and their objectives

Question No	Identified Research Questions	Objective
RQ 1	What are the recent techniques that are incorporated for FER and what are their contributions?	It is necessary to explore current state-of-art techniques used for FER and to compare each technique with others and to find their contributions.
RQ 2	What type of issues exists in FER?	It is necessary to explore each issue existing in FER and also discuss how to address these issues.
RQ 3	Discuss various taxonomies and comparisons of various surveys of the methods of FER.	Various taxonomies based on the existing surveys of Facial Sentiment Analysis methods to compare the methods and identify the most suitable method.
RQ 4	What are the research challenges faced by FER?	It aims to identify the challenges faced by the Facial Sentiment Analysis system in the near future and explore the solutions which can be applied to these challenges.

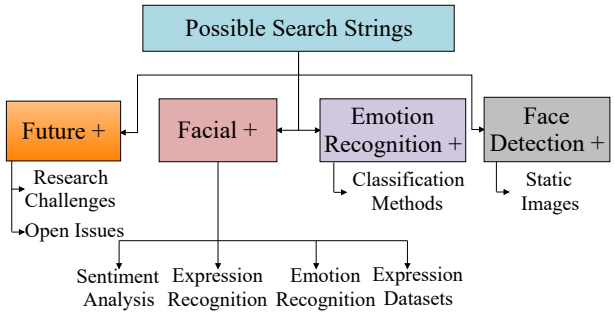

FIGURE 4: Possible strings used to search the literature.

emotions of a human are not static, it varies time-to-time. So, the classification of situation-based emotions is challenging.

III. SURVEY METHODOLOGY

In this section, we present the methodology followed to conduct the proposed survey such as search strings used, research questions, and the authentic data sources.

A. RESEARCH PLANNING

The proposed survey initiated with the discussion and identification of various quality research questions, data sources, as well as search criteria. We identified the relevant surveys proposed by various researchers and if data is found relevant, then we extracted data from it [55].

B. RESEARCH QUESTIONS

The proposed survey found out the existing literature on Facial Sentiment Analysis. The identified research questions are specified in Table 3.

C. DATA SOURCES

Various literature has been studied for a thorough and complete survey. We followed the genuine digital databases such as Springer, IEEEXplore (early access, magazine, and transaction articles), Science Direct (elsevier), ACM digital library, Google Scholar for accessing the existing literature surveys on Facial Sentiment Analysis [56].

D. SEARCH CRITERIA

The search is performed using some standard keywords like *"Facial Sentiment Analysis", "Facial Emotion Recognition"* and other matching keywords as mentioned in FIGURE 4. There exist many articles in different digital libraries, where the search string is not present either in the title or abstract [57], then a manual search process was done for such research papers.

IV. FACIAL EXPRESSION DATABASES

In this section, we discuss various datasets that are currently used for training and testing in FER. We also presented a comparative analysis of these datasets based on the articles

published till-date. The comparative analysis of the datasets is presented in Table 4.

A. EXTENDED COHN-KANADE (CK+)

CK+ is the most widely used dataset for FER systems [58], [59], which contains almost 593 different sequences captured from 123 different subjects. Sequences may vary between 10 to 50 frames, whereas frames shows the shift from a neutral face to a specific expression [59], [59].

B. THE JAPANESE FEMALE FACIAL EXPRESSION

JAFFE data-set includes 219 different images with 7 facial expressions captured from 10 Japanese female models [75]. Images of each and every model were captured while looking through a camers with semi-reflective plastic sheet. To remove less illumination problem on the face, tungsten lights were used.

C. RADBOUD FACES DATABASE (RAFD)

RaFD database is the database of portrait images of 49 subjects of 39 Dutch adults and 10 Dutch children [76]. All models have 8 facial emotions like neutral, sadness, happiness, disgust, anger, contempt, surprise, and fear with three gaze directions. Every emotion in the image was shown with eyes coordinated straight ahead, deflected to the left side, and turned away to the right side. Pictures were captured with white background from five different camera angles at the same time from left to right with 45°angle. There are total of 120 images for each model.

D. DELIBERATE EXPRESSION DATASET MMI

MMI database is a laboratory-controlled database which contains 326 sequences from 32 subjects [77], [78]. There are total of 213 sequences labeled with 6 expressions and the main advantage of this database is that there are 205 sequences captured in frontal view. The difference between CK+ and this database is that this database contains sequences which starts with a neutral expression and reaches to the specific expression at the middle of the sequence and then returns to the neutral expression. It is a complex database because there exist large relational variations due to the images have the same expressions and non-uniformity (many of the images have glasses, long hair, and mustache). Researchers widely use the first and the last three frames to the final expression to perform emotion recognition task [59].

E. FER2013

This database was first established during the ICML 2013 challenges of Kaggle [45], [59]. It includes a large number of images and important characteristics. It also includes unconstrained images collected automatically by Google image search API. It contains 28,709 images for the training set, 3,589 images for validation set, and 3,589 images for test set, which sums up to total 35,887 images with 7 expression labels [59]. All these images have been reduced to the size of (48×48) pixels. This is one of the most challenging datasets in FER.

F. TORONTO FACE DATABASE (TFD)

It is the combination of different FER datasets [44]. It includes 1,12,234 images, 4,178 of which are labeled with one of the seven expression labels, such as sadness, surprise, fear, happiness, anger, disgust, and neutral [59]. The main advantage of this dataset is that it contains images that have faces been already detected and reduced to the size of (48×48). There are 5 folds in this database where each fold has 70% of images for the training set, 10% of images for the validation set, and remaining for test set [59].

G. MULTI-PIE

This database contains 755,370 images with 337 subjects [41]. The main advantage of this database is that it includes images with 15 different viewpoints and 19 diverse illumination conditions and each image is named as one of the 6 expression. This is mainly used for multi-view 3D FER [59].

H. STATIC FACIAL EXPRESSIONS IN THE WILD (SFEW)

IT was designed by selecting frames (static) from the acted facial expressions in wild (AFEW) [69]. The advanced SFEW 2.0 was the benchmarking data used for EmotiW 2015 challenge. It includes 958 images for train, 436 images for validation, and 372 for test set labeled with 7 expression labels [59].

I. OULU-CASIA

Oulu-CASIA NIR (near-infrared) and VIS (visible light) facial expression database with six diverse expressions (surprise, happiness, sadness, anger, fear, and disgust) having 80 subjects between the ages of 23 and 58 years [79]. 73.8% of the subjects are males and the rest are females. The image resolution is 320×240 pixels.

J. MULTIMEDIA UNDERSTANDING GROUP

It is also known as MUG [43] dataset, which is a laboratory-controlled dataset with 6 basic emotions-anger, disgust, fear, happy, sad, surprise, and neutral. This database is divided into two segments. In the initial segment, the subjects were approached to play out the six basic emotions. The subsequent part contains a research facility that prompted feelings. There is a total of 86 subjects and 1462 image sequences. A camera has the option to click pictures at a pace of 19 frames/second. Each picture is in jpg format with (896×896) pixels and 240 to 340 KB size. This dataset was made to defeat the restrictions of other comparable datasets in FER, such as high goals, uniform lighting, numerous subjects, and numerous clicks per subject.

K. AFFECTNET

It is a dataset for the identification of wild human expressions. It has above 1 million different images to be collected

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

TABLE 4: Different facial expression databases

Dataset	Year	No. of Subjects	No. of Images	Classes	Resolution /Color or Gray	Author	URL	Type
Japanese Female Facial Expressions (JAFFE)	1998	10	213 static images	N, S, Sr, H, F, A, and D	256x256 /Gray	[37]	[60]	Posed
Extended Cohn-Kanade (CK+)	2000	123	593 sequences	N, S, Sr,H, F, A, C, and D	640x490/ Gray	[38]	[61]	Posed and Sponta-neous
MMI	2002	75	740 images	N, S, Sr, H, F, A, and D	720x576/ Color	[39]	[62]	Posed and Sponta-neous
Oulu-CASIA	2008	80		N, S, Sr, H, F, A, and D	320×240/ Color	[40]	[63]	Posed
Multi-PIE	2009	337	750,000	Sm, Sr, sq, D, sc, and N		[41]	[64]	Posed
Multimedia Understanding Group (MUG)	2010	86	1462 sequences	N, S, Sr, H, F, A, and D	896×896	[43]	[65]	Posed
Toronto Faces Dataset (TFD)	2010		1,12,234	N, S, Sr, H, F, A, and D	48x48/ Gray	[44]	[66]	posed
Radbound Faces Database (RaFD)	2010	67	8040	N, S, Sr, H, F, A, C, and D	681x1024/ Color	[42]	[67]	Posed and Sponta-neous
FER-2013	2013	-	35,887	N, S, Sr, H, F, A, and D	48x48/ Gray	[45]	[68]	Posed and Sponta-neous
SFEW (EmotiW)	2015	-	1766	N, S, Sr, H, F, A, and D	48x48/ Gray	[69]	[70]	posed
AffectNet	2017	450000	1,000,000	Affection, A, Annoyance, Anticipation, Aversion, Confidence, Disapproval, Disconnection	variable/ color	[71]	[72]	Wild
EMOTIC	2019	34320	23571	26 Classes	variable/ color	[73]	[74]	Wild

N:Neutral, S:Sadness, Sr:Surprise, H:Happiness, F:Fear, A:Anger, D:Disgust, C:Contempt, Sm:Smile, Sq:Squint, Sc:Scream

from the internet source by using over 1200 keywords related to human emotions.

L. EMOTIC

It is a facial emotion dataset with EMOTions In Context. It collected the images of people in the real environment with apparent emotions. It has widest 26 emotions categories.

V. FACIAL SENTIMENT ANALYSIS: THE PROPOSED TAXONOMY

This section presents the taxonomy for Facial Sentiment Analysis, which is splitted into three subsections such as pre-processing (face detection, dimension reduction, and normalization), feature extraction, and emotion classification. The detailed taxonomy for FER is shown in FIGURE 5.

A. FACIAL DETECTION, DIMENSION REDUCTION, AND NORMALIZATION

In this section, we discuss the various face detection, dimension reduction, and normalization techniques that are widely used in FER models. We also highlight and compare the various face detection techniques proposed by different researches. Table 5 shows the comparison of various state-of-art detection techniques available in existing literature.

1) Viola-Jones Face Detection Algorithm

Viola-Jones is extensively used to perceive the face from an image. The training time of this algorithm quite long, but face identification is fast. It needs the full front view of the face as an input image. It has four stages, such as haar-like features, integral graphs, AdaBoost training, and cascading classifier.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

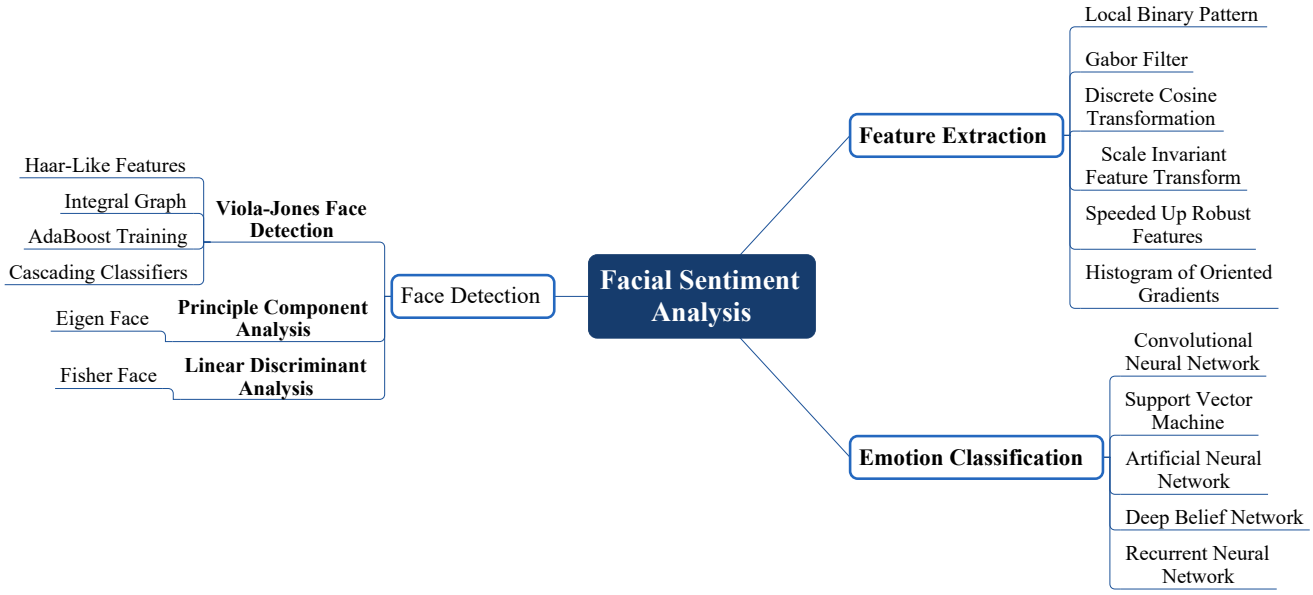

FIGURE 5: The Proposed Taxonomy for facial sentiment analysis

TABLE 5: A relative comparison of various state-of-the-art facial detection techniques

Author	Year	Algorithm Used	Approach
Jayaleks- hmi *et al.* [80]	2014	Viola-Jones	Used Viola Jones technique for face localization under the feature based classification
Li *et al.* [81]	2015	Viola-Jones	Used naive Viola-Jones to detect the face from the image
Ding *et al.* [82]	2017	Viola-Jones and IntraFace	Applied the Viola Jones algorithm and IntraFace for face detection and landmark detection
Lopes *et al.* [83]	2018	Viola-Jones	It detects the face points
Chaudhari *et al.* [84]	2018	Viola-Jones	They have used Viola-Jones to detect facial areas.
Kar *et al.* [85]	2019	PCA, Viola-Jones	Used Viola-Jones to detect face from image and then applied PCA and LDA combined to reduce the number of features which are redundant among various classes
Shah *et al.* [86]	2019	Haar Cascades, DNN	1) Haar-Like features are divided into different cascade for face detection. 2) They also proposed another approach to detect face, that is, a DNN based on ingle shot detector framework and ResNet.

- *Haar-Like Features*: Viola-Jones algorithm uses Haar-like features, 1.e., a scalar product between the image and some Haar-like templates [87]. As shown in FIGURE 6, edge features, linear features, center features, and diagonal features are the four Haar features used in Viola-Jones face detection algorithm [88]. There are two regions, as shown in the figure, black shaded and white shaded regions. The eigenvalue is calculated using the difference between those two regions for linear features [88].

$$eigenvalue(v) = \sum_{white} - \sum_{black} \quad (1)$$

$$eigenvalue(v) = \sum_{white} -2 \times \sum_{black}, \quad (2)$$

- *Integral Graph*: As the dimension of the generated Haar feature is large, a technique called integral map can be used to isolate the picture cells, such as 2D coordinates of the gray-scale picture and the estimations of every pixel point [88]. The procedure to make an integral graph is that each pixel is made equivalent to the total of all pixels above and to one side of the concerned pixel. Henceforth, the total of all pixels in the rectangle shape is determined.
- *AdaBoost Training*: This training algorithm is a weak classifier and is made to learn multiple times to become good. The two things that we need to consider for the classification of an image. First, the locale of the eyes is darker than the district of the nose and the cheeks,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

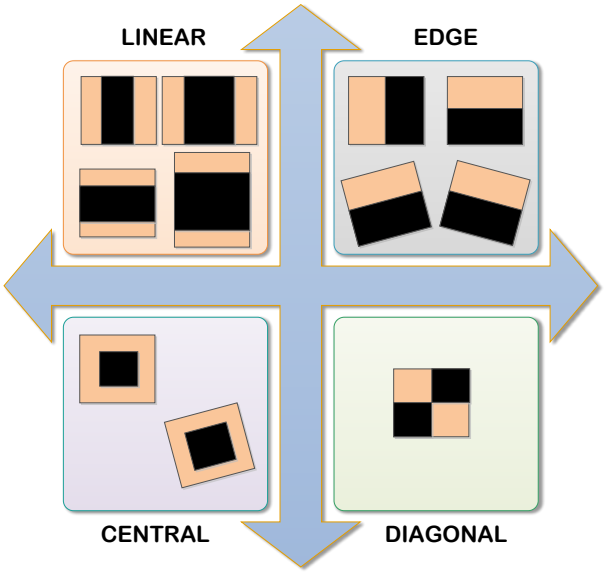

FIGURE 6: Four different types of Haar-like feature representations [88]

whereas the other thing is, eyes should be darker than nasal scaffold [80].

- *Cascading Classifiers*: We can classify the face or not a face in a single image using a single classifier, but the result of that might not reach our expectations. So, in Viola-Jones, we use cascading classifiers to make this classification accurate. Stages of the cascading classifier is shown in FIGURE 7. If one of the stages fails, then the image is not a face, but if it successfully reaches the last classifier, then the image is classified as a face and store it into the corresponding database.

Viola-Jones is not capable of handling occlusion and rigid objects. So the algorithm might generate false detection of face. To overcome this issue, the authors in [88] proposed the modified Viola-Jones algorithm using composite features. The procedure to detect the face using the modified Viola-Jones algorithm is as follows.

- A rectangular frame of the face is determined using Viola-Jones.
- The face in the rectangular frame is then calibrated and handled into four sorts of sub-images.
- The features are extracted from the acquired face and four sub-images using Zero/Null Space Linear Discriminant Analysis (NLDA).
- Then, the extracted features are evaluated by using discriminant distance.
- Now new composite feature vectors are generated from the discriminant values and then they are fed to a classifier for face recognition.

2) Principal Component Analysis

The fundamental thought behind the PCA is that multi-attribute data is projected onto a linear lower-dimensional space. This subspace is known as the *principal subspace*. The human face can be recognized using eigen face. Eigen space (the basis of faces) is a set of eigen vectors (the covariance network of the face space) is to classify the faces according to their basis representation. Steps to create the eigen faces are described as follows.

- Provide a training set of faces with pixel resolution of $w \times h$.
- Then the mean is calculated and subtracted from each image in the matrix. Weight of the k^{th} eigen face is $w_k = V_k^T(U-M)$, // where Input image vector $U \in R^n$ (training set) and mean M.
  Then $W = [w_1, w_2, \ldots, w_k, \ldots, w_n]$
- Calculate Euclidean distance (D) of W_x and W.
- After calculating Euclidean distance, the image is classified as face or not a face.

Islam *et al.* [89] used PCA to reduce the redundant features. They have used downsampling to eliminate the number of redundant features. Consider the size of an image as $(m \times n)$. The size gets $(40 \times m \times n)$ after filtering its size, but after downsampling its size reduced to $(10 \times m \times n)$, i.e., the dimension is reduced by a factor of 4. Later, Luo *et al.* [90] used PCA to extract global features from an image that are important, but they can be environment-sensitive for facial expression. To overcome such issue, some local features are also selected using LBP.

3) Linear Discriminant Analysis

LDA is the same as PCA, which is used to reduce the dimensions of a given data. Like the eigenface in PCA, LDA uses fisher face (enhancement of eigenface) for reducing the dimensions of the features and the identification of face in an image. Fisher's face is usually used when the images have a contrast in illumination. Various steps to create the fisher face is same as PCA and performs better than PCA [91].

B. FEATURE EXTRACTION

This section discussed the various feature extraction techniques and explained how they could be used in FER models. We also compare the various Feature Extraction techniques and also analyzed various works done using the various techniques, as shown in Table 6.

1) Local Binary Pattern

It is the recent texture descriptor that converts the value of the original pixel of the image with a decimal value and converts it into codes known as LBP codes [102]–[104]. Labels are formed by thresholding the 3×3 neighborhood with a central value and considers the result as a binary number. But the basic LBP had a limitation with large-scale structures and highly sensitive to noise [105], [106]. It is also invariant to the rotations and size of the features, which are increasing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

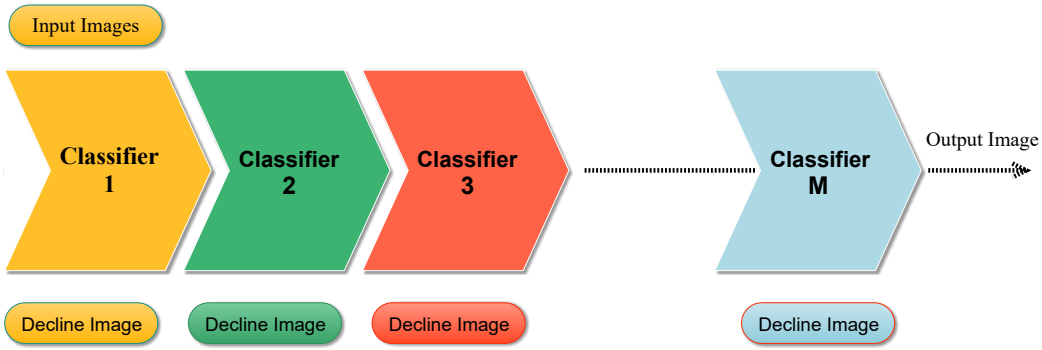

FIGURE 7: Cascading Classifier

TABLE 6: Comparison of various state-of-art feature extraction techniques

Author	Year	Method used	Approach
Lui *et al.* [92]	2010	Gabor Filter	proposed a local filter bank LG $(m \times n)$ to reduce the redundancy in eigen face values
Huang *et al.* [93]	2012	Speeded and Robust Features	Implemented the SURF technique that uses Haar Wavelet features and constructs a square region around it.
Luo *et al.* [90]	2013	Local Binary Pattern	Used normal LBP and micro-mode information such as edge and spots for feature extraction
Biswas *et al.* [94]	2015	LBP	Divided the images into blocks and modified the LBP algorithm by generating the compressed binary pattern of those divided images
Kravets *et al.* [95]	2015	Scale Invariant Feature Transform	Proposed a parallel SIFT algorithm that divides the entire feature extraction task into sub-tasks.
Mollahosseini *et al.* [96]	2016	IntraFace	Adapted a new method of SIFT which uses bidirectional mapping and extracts 49 points from the input image.
Mehta *et al.* [97]	2016	Gabor Filter	Extracted the features of face image by filtering it using a Log Gabor Filter Bank of (5×8). This resulted in 40 Log Gabor Filter filtered images.
Jayalekshmi *et al.* [80]	2017	Zernike moment, LBP, DCT	Used zernike moments which are translation and rotation invariant by providing extra normalization that are used by LBP to enclose textual data within image and works on fixed discrete sequences.
Sajjad *et al.* [98]	2018	HOG, ULTP	Described the scattering of edge directions or intensity gradients also describing Local appearance and shape features. Provided more robustness against noise. Obtained feature vector by fusing both ULTP and HoG into a single vector.
Shrivastava *et al.* [99]	2018	CNN	used a Convolutional Neural Network (CNN) to extract 15 essential feature points distributed in regions such as eyes, eyebrows and mouth on a given face.
Lopes *et al.* [83]	2018	Gabor Filter	The cropped image with an elder face is sent to a Gabor Filter function as it reduces power required.
Islam *et al.* [89]	2018	Gabor Filter	Used Gabor Filter bank with five scales and eight orientations after orientation at different points.
Ravi *et al.* [100]	2020	LBP	Used LBP by taking 8 neighborhood pixels and thresholding each pixel to 8 binary digits.
Ramos *et al.* [101]	2020	Gabor Filter	Used eigenfaces for feature extraction and identified the distinctive points of the face

exponentially with the increase in neighbors.

A Uniform LBP was proposed that considers the U pattern, which has at most 2 bitwise transitions from 0 to 1. So, various extensions of the LBP were proposed for the neighborhoods of any size. A circular neighborhood was proposed with any number of pixels and radius. It is represented by the notation (P, R) where P means the number of sampling points on a circle of radius R. There are various applications

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

where LBP is used, such as-texture analysis, face analysis, and classification. It codifies the local primitives, including the edges, corners, different spots, and flat areas. Nowadays, LBP converts the important pixels of an image into a histogram which is known as the Histogram of Oriented Graph approach (HoG), that stores the information of local-micro patterns of the faces.

A modified algorithm for LBP was proposed by the authors of [94]. The steps for generating the threshold are as follows.

- The input preprocessed image is divided into 3×3 blocks.
- For each block, calculate the minimum and maximum of block representing the pixel intensity value of the block.
- Now, calculate the threshold value of block B by taking an average of both the minimum and maximum values.
- If any element of the block is greater than threshold, then write '1' to it, else write '0'.
- The eight-bit pattern is converted to a decimal number, representing transformed block B.

2) Gabor Filter

It extracts both time and frequency domains [107] of the image. Means, it analyzes whether there is any particular frequency content in the image in a particular direction around the point of analysis. The use of 2D Gabor Filter is made in the spatial domain. Gabor Filters is quite successful in FER models. Multi-resolution structures are applied to images which consist of multi-frequencies and multi orientations. These structures relate Gabor Filters to wavelets [108].

The filters having real and imaginary components represents the orthogonal directions. The equations are shown below [109]:

$$\Psi_{\omega,\theta}(a,b) = \frac{1}{2\pi\sigma_a\sigma_b}\exp\left[-\frac{1}{2}\left(\frac{a^{'2}}{\sigma_a^2}+\frac{b^{'2}}{\sigma_b^2}\right)\right]\exp\left[j\omega a^{'}\right] \tag{3}$$

$$a^{'} = a\cos\theta + b\sin\theta \tag{4}$$

$$b^{'} = -a\sin\theta + b\cos\theta \tag{5}$$

where (a,b) is the pixel position in the spatial domain, θ is the orientation of Gabor filter, ω is the radial central frequency and σ is the standard deviation of the Gaussian Filter which means it controls the size of the Gabor Envelope.

Liu *et al.* [92] proposed in his paper the local Gabor filter bank LG $(m \times n)$ which spreads all over. It also contains multi-scale information of features those having a global filter or the image as well as it also reduces the redundancy in eigen values. This reduces the time for extracting the features. [89] used a Gabor Filter bank with eight orientations and five scales. The formed bank is used to filter the generated divided images 40 times each. This created a computational burden from them, so they reduced the number of features by dimension reduction techniques. Dimension reduction using PCA is explained under section PCA.

3) Discrete Cosine Transform (DCT)

A finite sequence of data or feature points as the sum of cosine functions are oscillating at different frequencies is represented by DCT. It is a way to compressing the data/2D-image without losing its original meaning [110]. In such type of applications (data compression), an input of 8×8 size is used for DCT [111] for feature extraction. It has two stages:

- The first stage is to apply DCT on the image.
- The second step is the selection of co-efficients [112].

By applying DCT on an $U \times V$ image then a 2D $U \times V$ co-efficient matrix is formed.

$$G_x(0) = \frac{\sqrt{2}}{M}\sum_{m=0}^{M-1} X(m) \tag{6}$$

$$G_x(k) = \frac{2}{M}\sum_{m=0}^{M-1} X(m)\cos\frac{(2m+1)k\Pi}{2M}, \tag{7}$$

$$where \quad k = 1, 2, \ldots, (M-1) \tag{8}$$

where $G_x(k)$ is the kth DCT co-efficient [113]. Jayalekshmi *et al.* [80] in their work have used DCT over fixed discrete sequences to convert the data into elementary frequency components.

4) Scale Invariant Feature Transform (SIFT)

It transforms the input image data into scale-invariant coordinates relative to the local features and stored into a database [114]. SIFT features are highly distinctive i.e., it can match a single feature with a large probability from the database. SIFT also features scale and rotation invariant, which means that even if we scale or rotate the image, the features remain preserved. This is useful in FER when a rotated image comes as an input. If we rotate the image, then also the features are maintained, and we can get those features efficiently [115]. The stages of the SIFT procedure are as follows.

- *Scale-Space Extrema Detection*: The first stage searches all image locations and scales using the Gaussian method.
- *Keypoint localization*: It determines the location and scale at each point of the image.
- *Orientation Assignment*: Rotation Invariance is performed by assigning one or more dominant orientations to each key point.
- *Keypoint Descriptor*: A descriptor is made to represent each keypoint, which supports the assigned orientation in the preceding stage. It supports the histogram of the gradient within the image. The changes in illumination are scaled back by the descriptor to the key point. When the keypoint descriptors are received, they are often used as a feature or keypoint for data to solve various problems. More detailed information on SIFT computation are often found in [116], [117].

Kravets *et al.* [95] proposed P-SIFT (Parallel SIFT) algorithm, which reduces the computation time and increases

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

the processing speed. In P-SIFT, the problem is divided into sub-tasks, and multiple processors are used for feature extraction. The program reads the input image and generates the key points. After that, it matches the key points with respect to each image in the database. Matching is done using Euclidean distance. The images that have a ratio of first least distance to the second least distance is less than 0.8 are taken into consideration. Then the image is given to the classifier that classifies the emotion [118].

A newly adapted method of SIFT to extract features from an image called IntraFace was proposed in [96]. It uses multi-directional warping of active visualization model and a supervised descent model [119], which uses the SIFT feature extraction technique for feature mapping and trains a method to extract 49 points from the image. These points are used for registering an average face, which is then termed as the face region.

5) Speeded Up Robust Features (SURF)

Speeded Robust Features is a type of local feature detector as well as descriptor. It is inspired from SIFT and the authors claim that it is faster than SIFT. Around the interesting point, a certain reproducible orientation is fixed based on information from a circular region. From this selected orientation, we make a squared region, and the SURF descriptor is extracted [120]. The SURF has two parts (i) a detector and (ii) a descriptor. The location of the key points of the image is provided by the detector and the descriptor expresses the features of those key points. SURF uses Hessian matrix for the fast assessment of box filters. The integral images are expressed as

$$J(p,q) = \sum_{l=0}^{p}\sum_{j=0}^{q} I(l,m) \tag{9}$$

The Hessian matrix is represented as:

$$H(X,\sigma) = \begin{bmatrix} O_{aa}(M,\sigma) & O_{ab}(M,\sigma) \\ O_{ab}(M,\sigma) & O_{ab}(M,\sigma) \end{bmatrix} \tag{10}$$

The introduction of pyramid scale space is done in SURF because of box filters. The descriptor makes use of the sum of Haar wavelet features that increases the robustness and decreases the computation time. For extraction, it constructs a square region around the keypoint and oriented along the orientation decided by a method. Each region is split into 4×4 sub-regions, which keeps the important spatial features. The Haar wavelet computes at 5×5 sampled points [93].

6) Histogram of Oriented Gradients (HoG)

The features extracted by HoG are tough against photometric and geometric deviations. Many applications use HoG as the feature extraction technique, such as human detection [116]. The steps to calculate these features are:

- In the first step, it divides the entire image into small cells.
- Then, its direction and the magnitude is calculated.
- Calculate the Bin for each direction as well as magnitude using HoG.
- The blocks from adjacent cells are calculated, and block normalization are created to calculate feature vector from it.

During implementation, 9 bins of histogram were calculated using the cells of 8×8 pixels and blocks of 2×2 pixels with unsigned orientation. [121] used a fusion of HoG and LBP features. Firstly, the HoG and LBP features were extracted from the segmented parts. The final feature vector consists of both the features of HoG and LBP, which had 1892 features out of which 1656 came from HoG while the rest came from LBP.

C. EMOTION CLASSIFICATION

In this section, we discuss the various Classification techniques that are used to classify human face emotions. We also presented a comparative analysis of various emotion classification techniques, as shown in Table 7.

1) Convolutional Neural Network (CNN)

CNN is most widely used architectures in computer vision techniques as well as in machine learning [136]. A massive data is required for training purpose to harness its complex functions solving ability to its fullest [137]. CNN uses convolution, min-max pooling, and fully connected as layers than the conventional fully connected deep neural network [53], [138], [138], [139]. When all these layers are stacked together, the complete architecture is formed. The complete architecture of CNN is shown in FIGURE 8.

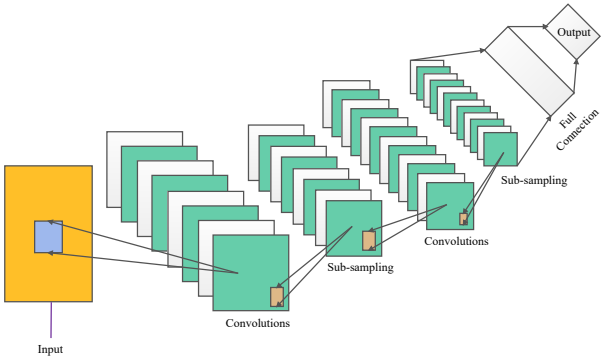

FIGURE 8: Traditional CNN architecture

- *The input layer* of CNN contains the image pixel values.
- *The Convolutional layer* convolves the $l \times l$ kernels with x feature maps of its preceding layer. If the next layer has feature maps, then $n \times m$ convolutions are performed and $n \times m \times (w \times h \times l \times l)$ Multiply-Accumulate (MAC) operations are needed, where h and w represents the feature map height and width of the next layer [138]. The important function of Convolutional layer is to calculate the output of all the neurons which are connected to the input layer. The activation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

TABLE 7: Comparison of various State-of-art emotion classification techniques.

Author	Year	Objective	Dataset			Algorithm	Accuracy	Merits	Demerits
			Name	Classes	Size				
Luo *et al.* [90]	2012	Proposed a model which uses PCA and LBP for feature extraction and used SVM for classification	–	7	350	PCA, LBP, SVM	93.75%	Performed better than traditional approach	Low accuracy
Lui *et al.* [122]	2014	Proposed a novel BDBN framework for overall process of facial emotion recognition	CK+; JAFFE	8	1308; 213	Boosted DBN, BTD-SFS	96.7% (CK+), 91.8% (JAFFE); 93% (CK+ and JAFFE)	Model results were far better than the previous ones	High computational complexity
Lv *et al.* [123]	2014	Studied the FER for face parsing using DBN by removing the redundant information	JAFFE; CK+	6;7	213;593	DBN	91.11% (CK+), 90.47% (JAFFE)	Removes redundant information for expression recognition	Low accuracy
Mollahosseini *et al.* [96]	2015	Implemented the NN to perform cross-database classification	MultiPIE; MMI; CK+; DISFA; FERA; SFEW; FER2013	7	20,000; 11,500; 309; 89,000; 7000; 663; 35887;	CNN	94.8% (MultiPIE); 77.9% (MMI); 55.0% (DISFA); 93.2% (CK+); 66.0% (FER2013)	Increased classification accuracy over AlexNet	Works well with only MMI and FER2013
Adeyanju *et al.* [124]	2015	Analyzed the performance of SVM kernels for face emotion recognition	–	7	714	4 SVM Kernels	99.32%	Increased Performance	High computation time
Talele *et al.* [125]	2016	To propose the neural network approach for classification	JAFFE; TFED; CK; Indian Face database	7	213; 7200; 1308; 1,23,213	Regression NN	95.48%; 94.25%; 94.91%; 94.81%	Better performance on (64 X 64) block size	Low accuracy over advanced approaches
Khorrami *et al.* [126]	2017	Demonstrated how zero-bias CNN achieves recognition accuracy	CK+, TFD	6	1308; 4187	Zero-bias CNN, data augmentation	95.1%	No biasness, quick training, reduced learning parameters	Not good for real-time application
Ding *et al.* [82]	2017	Proposed a two-stage training algorithm for expression identification	CK+, Oulu-CASIA, TFD, SFEW	7	1308(CK+), 1444(Oulu-CASIA), 4178(TFD), 1322(SFEW)	Fine tune CNN	98.2% (CK+), 87.71% (Oulu-CASIA), 88.9% (TFD), 48.19% (SFEW)	Improved visual representation and outperforms with four public datasets	–
Wen *et al.* [127]	2017	Implemented an ensemble approach to overcome the CNN limitations	JAFFE, CK+, Emotiw2015, FER2013	7	213, 593, 1238, 35887	Ensemble CNN	50.70%(JAFFE); 76.05%(CK+); 34.09%(Emotiw2015)	Addressed CNN cons	Highly inefficient with few training samples
Zhang *et al.* [36]	2017	Proposed an Hierarchical Bidirectional RNN for analyzing the facial expressions	CK+, Oulu-CASIA, MMI	7,7,8	593, 1440, 203	PHRNN, MSCNN	98.50% (CK+), 86.25% (Oulu-CASIA), 81.18% (MMI)	Outperforms over state-of-the-art approaches	–
Datta *et al.* [128]	2017	Proposed an advanced SVM to improve the performance of FER	CK+	7	593	SVM	91.85% (One vs One), 89.26% (DAGSVMs)	Improved performance with geometric and texture features	Poor accuracy compared to other approaches
Lopes *et al.* [83]	2018	Classified the facial expressions of elderly people with other age group people	Lifespan dataset	3	778	Multiclass SVM	87.93% for 19-59 age group; 80.51% for 60-93 age group	Successful in classifying emotions in elders that have wrinkles that are hard to detect	Accuracy detecting sad emotion was 66% which is very low
Jadhav *et al.* [54]	2018	Implemented general CNN for emotion recognition	FER2013	7	9000	Haar featured cascaded classifier, CNN	63%	Reviewed different approaches and proposed a model using CNN	Low accuracy compared to other approaches
Cai *et al.* [129]	2018	Implement a novel Island loss CNN and VGG16 for expression recognition	CK+, MMI, Oulu-CASA, SFEW	7	981 (CK+), 624 (MMI), 1440 (Oulu-CASA), 1776 (SFEW)	Loss Layered VGG16 and CNN	94.39% (CK+), 74.68% (MMI), 84.58% (Oulu-CASA), 52.52% (SFEW)	Improved accuracy than naive VGG and naive CNN	Low accuracy
Dhankhar [130]	2019	Implement an ensemble approach to overcome the limitations of CNN	FER2013; KDEF	7	35887, 4900	VGG16, RESNET50	67.2% (FER2013), 71.2% (KDEF)	Addressed the barriers of CNN	Less accuracy with FER2013 dataset
Renda *et al.* [131]	2019	Provided the indications to build an effective ensembles of CNN	FER2013, CK+, SFEW	7	35887, 593, 1321	Ensemble approach	$71.236 \pm 0.013\%$(SE), $71.728 \pm 0.220\%$(PS), $72.044 \pm 0.160\%$(PT), $70.558 \pm 0.387\%(BA)$	Improved Performance in FER classification	Low performance
Gan *et al.* [132]	2019	Improved the ensemble classifier performance with novel label-level perturbation strategy	FER2013, SFEW, RAF	7	35887, 1766, 15339	Ensemble Approach	73.73%, 86.31%, 55.73%	Enhanced discrimination ability of ensemble classifier	Not emphasized on all face points
Zadeh *et al.* [133]	2019	Proposed a DL approach for FER to improve its performance	JAFFE	7	213	CNN with 2 gabor filters	91.16% (CNN)97.16% (CNN with 2 gabor filter)	Increased system learning rate performance and by using two gabor filters	Poor performance
Kurup *et al.* [134]	2019	Proposed a semi-supervised emotion recognition algorithm with reduced features	CK+, MMI, RAFD	7, 6, 8	327, 205, 536	DBN	98.57% (CK+), 91.95% (RaFD 135°), 94.50% (RaFD 90°), 92.75% (RaFD 45°), 98.75% (MMI)	Improved efficiency	–
Ravi *et al.* [100]	2020	To provide fair comparison between commonly used FER techniques	CK+, JAFFE and YALEFACE	7	213 images of 10 subjects	CNN	JAFFE-73.81%, YALEFACE-72.18%, CK+-89.62%	High accuracy compared to previous state-of-the-art approaches	Low accuracy on JAFFE dataset compared to others
Pranav *et al.* [135]	2020	Implemented Deep CNN to accurately classify the emotions	Self-made dataset	5	2550 images	CNN	78.04%	Proposed Deep CNN and self-made difficult dataset	Low-accuracy compared to previous state-of-the-art techniques

functions such as ReLu, sigmoid, tanh etc. aim to apply element wise activation and to add the non-linearity into the output of neuron

- *The pooling layer* has the responsibility to achieve spatial invariance by minimizing the resolution of feature map. One feature map of the preceding CNN model layer is corresponding to the one pooling layer.
  1) *Max Pooling*: It has a function *u(x,y)* (i.e., window function) to the input data, and only picks the most active feature in a pooling region [140]. The max pooling function is as follows:

$$a_j = \max_{N \times N}(a_i^{n \times n} u(n,n)) \tag{11}$$

  2) *Average Pooling*: It has a function *u(x,y)* (i.e., window function) to the input data, and selects the average value for each input data on the preceding layer feature map [141], [142]

$$act_i = \frac{1}{M \times M} \sum_{J=1}^{M \times M} x_j \tag{12}$$

Mostly, 2×2 pooling can be used without overlapping. This means that M in the above equation is always 2. And the large pooling has M value as 4, 8, 16, which always have a dependency on input image size. So, [142], in their paper, proposed a Multi-activation pooling method in order to satisfy the need of a large

pooling region. This method allows top-p activations to pass through the pooling rate. Here p indicates the total number of picked activations. If p = $M \times M$, then it means that each and every activation through the computation contributes to the final output of neuron [142]. For the random pooling region X_i, we denote the *nth*-picked activation as act_n

$$act_n = max\left(X_i \Theta \sum_{j=1}^{n-1} act_j\right) \tag{13}$$

where the value of n ∈ [1,p]. The above pooling region can be expressed as below where symbol Θ represents the removal of elements from the assemblage. The summation character in Eq. 13 represents the set of elements that contains top1 (n-1) activation, but not adding the activation values numerically. After having the top-p activation value, we simply compute the average of each value. Then, a hyper-parameter σ is taken as a constraint factor which perform the multiplication of the top-p activations [142]. The final output refers to

$$output = \sigma * \sum_{j=1}^{p} act_j \tag{14}$$

Here, the summation symbol represents the addition operation, where $\sigma \in (0,1)$. Particularly,if σ = 1/p, the output is the average value. The constraint factor, i.e., σ can be used to adjust the output values [142].

- *Fully connected (FC) layer* is the last layer of CNN architecture. It is the most fundamental layer which is widely used in traditional CNN models [143], [144]. As it is the last layer, each node in it is directly connected to each and every node on both sides. As shown in FIGURE 8, it can be noted that all the nodes in the last frame of the pooling layer are converted into a vector and then are connected to the first layer of the fully-connected layer. There are many parameters used with CNN and need more time for training [145], [146]. The major limitation of FC layer, is that it contains a large number of parameters that need complex computational power for training purposes. Due to this, we try to reduce the number of connections and nodes in the FC layer. The nodes and connections which are removed can be retrieved again by adding the new technique named dropout technique.

In the past few years, CNN has emerged in Computer Vision, including the field of facial sentiment analysis. Researchers have modified the traditional CNN for better performance. A modified CNN was proposed by Mollahosseini *et al.* [96] in which an inception layer was introduced. Their network architecture consisted of two elements. Firstly, it had two traditional CNN architectures containing the Convolutional layer, followed by ReLu. Following these modules, they added two inception layers which consists of 1×1, 3×3 and 5×5 Convolutional layers with ReLu in parallel.

By slightly modifying the traditional CNN, Khorrami *et al.* [126] developed the model based on a classic feed-forward CNN. They introduced a modification in their model by ignoring the biases of the Convolutional layers. The network contained three convolutional layers having filters of size 64, 128, and 256, respectively, with 5×5 sized d filters, which were then followed by activation function ReLu. They placed max-pooling layers after each of the 1^{st} two Convolutional layers, and after the 3^{rd} convolutional layer, they placed the quadrant pooling layer. Then after convolutional layers, FC layer containing 300 hidden units followed by quadrant pooling was used. At last, the softmax layer was used for classification.

The idea of using zero-bias model was first introduced in [147] for the fully-connected layers in the CNN model and later was extended in [148]. They implemented this model on CK+ and TFD datasets. Their model was successful in recognizing the emotions with the rate of $88.6\% \pm 1.5\%$ on TFD with 7 classes and $95.1\% \pm 3.1\%$ on CK+ with 8 classes. They used Data Augmentation combined with dropout to boost the performance [126].

Further, the increase in accuracy and recognition performance of the computer vision algorithms researchers proposed advanced and deeper CNN architectures. Taking into consideration the work of Ding *et al.* [82] designed a new technique named FaceNet2ExpNet to train the model. They used a fine-tuned face net and proposed a unique distribution function to train neurons of expression net. To improve the discriminativeness of features that were learned, they used the conventional network to design the ession net. The training process was executed in the two levels. In 1^{st} level, the Convolutional layers were given training using loss function, and the output of the last pooling layer was used for supervision. In the $2nd$ stage, they added randomly initialized FC layer and then trained the network using labeled training data. The testing was done on constrained as well as unconstrained datasets and achieved better results than the previous approaches [82].

In 2018, Jadhav *et al.* [54] investigated the previous approaches and applied modifications to increase performance. They investigated three different popular networks that were quite successful in classifying emotions. The first network was proposed by Krizhevsky and Hintion [149]. The second network for investigation was inspired by AlexNet [144] Convolutional network and the last one was proposed from work done by Gudi [150]. It consisted of an input layer of 48×48, one Convolutional layer, normalization layer to reduce normalize dimensions, and then a max-pooling layer. Then again, 2 Convolutional layers and then finally 1 FC layer, linked with the softmax layer. To decrease the number of parameters, [54] applied one more max-pooling layer. They trained and tested the model on FER2013 and RaFD respectively and got better results than [150].

Cai *et al.* in [129], introduced a new island loss layer in CNN to minimize intra-class variations. Their network includes 3 Convolutional layers, each followed by the PReLU and batch normalization (BN) layer. Pooling was used with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

the two initial BN layers. After the third Convolutional layer, 2 FC layers and island loss was calculated at the second FC layer. And then, at last, the softmax layer was used. Their architecture was named IL-CNN. They also employed VGG-16 [151] network as their backbone network. This approach achieved good performance in comparison to the state-of-art methods.

In 2012, Simonyan *et al.* [151] proposed VGG16 architecture for object recognition and classification task. VGG16 has replicative structure of convolution, ReLu and pooling layer. The network architecture of VGG16 is shown in FIGURE 9.

The invention of Residual Networks has created a revolution in the image recognition field. Residual Networks (ResNet) [152] is a classic network used as a backbone for the many computer vision tasks. ResNet50 is the current state of art convolutional neural network, which has the identity mapping capability. It introduced the concept of skip connections. Traditional CNN was successful in FER to some extent, but there were various limitations too. Those limitations were:

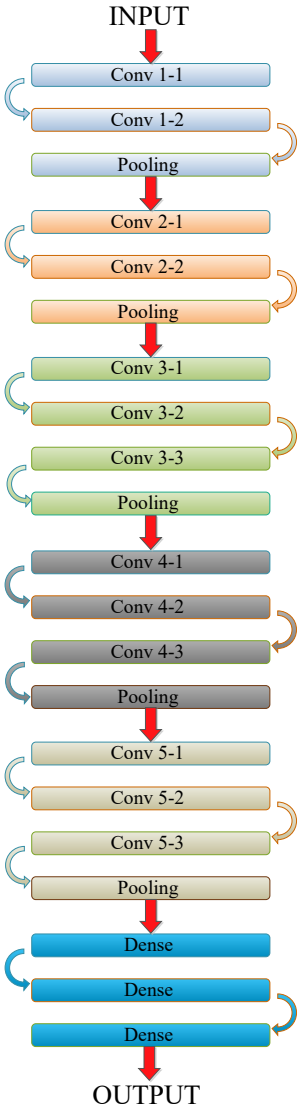

FIGURE 9: A conventional VGG16 architecture.

- It requires a considerable skills and high experience in selecting the appropriate hyper-parameters values for CNN [127].
- It used a stochastic gradient descent method, which caused trouble in mounting to enormous data and also in learning NN with multiple layers. It is because the gradients tended to decrease and also because of the problem of "vanishing gradients" [127] [153].
- The last one was for real-time environments, CNNs for human facial appearance was easily affected by various parameters like age, gender, face length, hair, mustache, and ethnicity. [154]. Due to this, facial expressions had overlapping features, which makes its implementation difficult and complex [127] [155].

To overcome these barriers, ensemble learning was applied because it helped in more accurate classification and prediction by concatenating the outputs of the base learning approaches. So, Dhankhar [130], [156] developed an Assemble model by combining the state-of-the-art DL approaches such as VGG16 and ResNet50. He attempted to obtain a vector of weights from the second to last layers, which can be treated as feature vectors, which represented the latent representations of the input image. Then, he combined above mentioned representations by joining the feature vectors, then it can be taken as input to logistic regression models to calculate the final emotion prediction [130]. After applying transfer learning, he trained and tested his model on Karolinska Directed Emotional Faces (KDEF) dataset and got better accuracy compared to individual models.

Another Ensemble network was proposed by Wen *et al.* [127] where they ensembled CNN with probability-based fusion for FER [157]. For each ensemble method, the multiplicity and diversity in the classifiers is considered as a major concern in achieving comparable performance [158]. The CNN architecture was implemented using ReLU and multiple hidden layers (maxout layer) with random values for them to overcome the problem of calculating the stochastic gradient descents. They used Softmax classifier at the last in order to roughly calculate the possibility to test sample to each class [127]. They trained and tested the model on FER2013 and CK+ datasets, respectively and the accuracy was 76.05%, which was better than other methods. Thus it can be concluded that ECNN consistently outperformed traditional CNN.

Alessandro Renda *et al.* [131] proposed a feed-forward CNN, which was inspired by Kim *et al.* [159], in which three convolutional and max-pooling layers with 32, 32 and 64 feature maps respectively were placed after input layer. They used an ensemble learning approach to increase performance. These max-pooling layers consist of an overlapping kernel with size 3×3 and stride of 2×2, which results in size halving. They added a dropout layer after FC hidden layer,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

having 0.15 as drop probability. A FC layer with almost 1024 neurons was proposed to yield 7 classes of emotion in the FER2013 dataset. Their model has a network depth of 5 with 2,436,007 trainable parameters. They used ReLU (Non-linear function) as an activation function for both the convolutional and FC layers to remove non-linearity and the softmax function at the output layer. They used the batch normalization [160] with each convolutional layer as well as the FC layers. To preserve the data, they used zero-padding in the convolutional layers. They achieved an accuracy of approx. 72.249 % with the ensemble of 9 networks.

To solve the problem of poor performance in real applications caused because of the stored facial images that most of the time show expression not as a single emotion but represents a multiple emotions, Gan *et al.* [132] designed an approach using CNN and soft label which associates with multiple emotions and expressions. They obtained the soft labels using constructor involving 2 step scheme:

1) The initial step is to prepare a CNN model with hard data labels for supervision and the softmax function for optimization.
2) The second step is to fuse the possibility of prediction to get soft labels from the pre-trained models. [132].

Their architecture is similar to VGG16, however, the last FC layer is adjusted as C-way yields, where quantity C is the number of emotion classes.

Zadeh *et al.* [133] proposed a DL model having a CNN layer and 2 Gabor Filters to classify different human sentiment. This model uses a feature selection method called Gabor Filter, which is commonly applied for texture outline. It returns where there is any texture change in the image. Then these features are fed to a CNN (Convolutional Neural Network) for the classification of human sentiment. Their model has the following stages- Input Images, resize, 1st Gabor Filter, 2nd Gabor Filter, CNN layer, and classification of sentiments. They tested their model on the JAFFE dataset. They also compared the dataset classification using simple CNN and its model (CNN with 2 Gabor Filters). They trained them for 30 epoch and got an accuracy of 91.16% on simple CNN and 97.16% on their model [133].

2) Support Vector Machine (SVM)

SVM is a classifier [161] was designed for classifying out of two classes. If the SVM has more than two classes, then more than one SVMs is to be implemented. There are three methods by which we can implement SVM for more than two classes.

- One versus all: It was proposed in [162]. It constructs k SVM models for training data having k number of classes. If there are three classes, then SVM can be performed three times for every class. [163]
- One versus one: It was introduced in [162]. This method constructs $k(k-1)/2$ classifiers, where two classes at a time are taken to train the model. In this method, SVM is performed between every class that is to be classified. [163]
- Directed Acyclic Graph SVM (DAGSVM): It was proposed in [164]. Its training phase is similar to one-vs-one method. The testing phase makes use of a rooted binary DAG having at the most $k(k-1)/2$ internal nodes and the maximum k leaves. An advantage of using a DAGSVM is that is to generalize the analysis [164].

The aim of SVM is to identify the maximum margin plane between the classes. The maximum margin plane can be obtained from the maximum distance between the positive and the negative margin plane, respectively of the two classes. The distance between the separating plane and the positive margin plane should be equal on both sides.

To solve the problem of recognizing emotions from facial expressions in a simple and speeded manner, Datta *et al.* [128] presented a classification system that used the concatenation of geometric as well as texture-based features to classify the emotions using SVMs. They have used the hierarchical SVM architecture to leverage the benefits of multi-class binary classification. CK+ dataset was used for classification. They have achieved the significant enhancements in the accuracy using hybrid SVM features compared to LBP features.

Nuno Lopes et al. [83] in 2018 given a classification model for FER in the elderly and also present the differences of FER in the elderly and other age people. They used the Support Vector Machine with a multi-class classification for classifying the emotions [165]. They proposed two architectures, the first approach removes the wrinkles, nasolabial fold, and other facial features, using edge-preserving smoothing techniques. While in the second architecture, they introduced an algorithm from API Microsoft, which detects the age of the person. The lifespan dataset was used to train and test the multi-class SVM. They used 80% images to test the accuracy of the SVM and 20% to test the accuracy of the application. They got an accuracy of 95.24% in the young age group and accuracy of 90.32% in the elderly age group.

SVM is a linear classifier that can be applied for linearly separable data. But SVM can take high dimensional data as input also which most of the time is non-linear data. So a mapping function is applied to the SVM training, which is non-linear and converts the data into linearly separable but in a higher dimension. This function is called a kernel function. There are various kernel functions, but [90], in his paper, used Radial Basis Kernel Function(RBF). They used one versus one approach in this paper.

Ibrahim Adeyanju *et al.* [124] proposed a method in which he used four SVM kernels to classify different emotions of faces. They used a Radial Basis, Polynomial, Linear, and Quadratic functions as SVM kernels. They tested their model on 467 training and 238 test sets to classify 7 emotions. They got a maximum average accuracy of 86.4% on RBF kernel, 99.33% on Quadratic function, 97.65% on Polynomial, and 97.86% on Linear.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

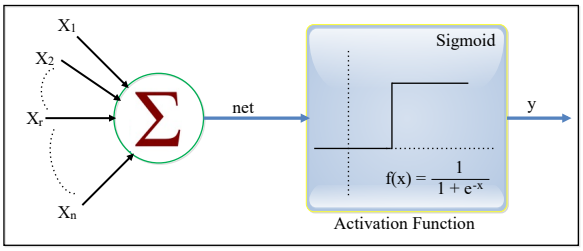

FIGURE 10: W.S McCulloch and W. Pitts proposed a single neuron model as a mathematical model for an artificial neuron [166]

3) Artificial Neural Network (ANN)

ANN is inspired by the biological neural networks that constitute the brain [166]. Our brain consists of millions of neurons that form a neural network. These neurons are interconnected with each other and process the signals to/from the brain to the other parts of our body [167]. This type of link is called synapses. There are approximately 100 billion neurons and are interconnected by thousands or more synapses. In ANN, the signal is a real or binary number and the output of these neurons is obtained by some activation function of the sum of all inputs to that neuron [166]. The connection between two neurons/nodes is called an edge. This edge has a weight, which describes how intensive the signal is. Normally, neurons are aggregated into layers. The first and last layer is the input and output layer, respectively. The in-between layers are called the hidden layers. A simple perceptron model given by W.S McCulloch and W. Pitts is shown in the FIGURE 10.

$$net = \left(\sum_{j=1}^{n} w_j x_j - u \right) \tag{15}$$

$$y = \theta\,(net) \tag{16}$$

$$\theta(x) = \frac{1}{1 + \exp\,(-x)} \tag{17}$$

where n inputs are given to the neuron x1, x2, âĂę , xn, weights are assigned to edges as w1, w2, ... , wn, θ is the step function before calculating the output O. The positive weights correspond to excitatory synapses and negative as inhibitory synapses. The activation function is the sigmoid function. This model does not possess the actual behavior of biological neurons. Talele *et al.* [125] proposed a model using the General Regression Neural Network (GRNN) based on ANN to classify emotions from the image. The proposed model has the following features- Input layer goes about as feed to the subsequent layers. The pattern layer decides the Euclidean distance and activation function. The summation layer comprises of the numerator and the denominator part took care of by the output layer. The fundamental principle on which the system works is the joint likelihood estimate of the input and the output as given below

$$f(x,y) = \frac{1}{(2\pi)^{\frac{d+1}{2}}} \times \frac{1}{n} \sum \exp \frac{(x - x_i)^r (x - x_i) + (y - y_i)^2}{2\sigma^2} \tag{18}$$

Where n is the number of watched tests, σ is the spread parameter, x_i is the i^{th} training vector, x_i is the corresponding yield esteem. The physical interpretation of the likelihood estimate is that it assigns test likelihood of width σ for each input and output test.

4) Deep Belief Network (DBN)

A probabilistic and unsupervised DL algorithm which comprises of numerous stochastic dormant factors. These Idle factors are likewise called as feature locators. It is a hybrid graphical model that has two undirected upper layers, while lower layers have directed connections [168]. DBN consists stack of Restricted Boltzmann Machine(RBM) or Auto-encoders. They represent a data vector. The two most important properties of DBN are:

- It utilizes layer by layer learning approach that decides how the loads rely on the layer above it, a top-down approach.
- A single bottom-up pass layer which begins with observed data vector and furthermore that utilizations loads in separate layers give the estimation of latent variables [169].

Deep Belief Network is pre-trained using the Greedy algorithm. In greedy algorithms, we train each layer, in turn, in unsupervised learning. The multi-layer DBN is divided into various RBMs, which are learned sequentially. Pre-training is done for better optimization. Fine Tuning is done because the features are modified so that we can get the category boundaries right [170].

Lui *et al.* [122] proposed a novel model named Boosted Deep Belief Network to implement FER with performance enhancement. Their Framework has three main contributions. First, they build the model, which consists of three-stage training of feature learning, feature selection, and classifier construction. Secondly, their proposed work facilitated the part-based representation and not the whole facial region as input, which is highly suitable for expression analysis. At last, they proposed a discriminative DL framework where multiple DBNs are integrated and a boosting technique is also applied. They used an experimental dataset named CK-DB prepared from the first and last three frames of the famous CK+ dataset with a total of 1308 images. The accuracy of the model was found to be 96.7%.

Kurup [134] used five layers. The input layer has two nodes. All classes are represented as 4-bit codes. All other layers have 3,3 and 4 nodes and have a sigmoid activation function. An unsupervised approach is used to train the first layer using contrastive divergence(CD) and then a softmax activation function is applied. At the end of the DBN, a fine-tuning backpropagation procedure was applied. RBMs are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

trained layer by layer and each RBM was trained individually five times. 5 times k fold cross-validation was used and the training data was divided into 5 groups. With this model, they got an accuracy of 98.57% on the CK+ dataset while 98.75% on the MMI database.

For the FER, Yadan Lv *et al.* [123] had proposed an approach via DL. Unconventional training of component detectors was done with DBN and was adjusted by logistic regression. After that, the parsed component features, including eyes, mouth are concentrated for expression recognition. The main contributions of their work are, they were the first to use only facial components to recognize emotion, treated every single feature of parsed component equal and parse the face via DBN so that the images need not to be pre-processed before extracting features. In simple words, their approach at first detected the face, and then the nose, eyes, and mouth are used for expression recognition. Emotion classification was done by a stacked autoencoder classifier.

5) Recurrent Neural Network (RNN)

They are an exciting twist to basic neural networks. RNNs can take a series of inputs with no initial limit on it. They remember the past and make decisions based on past learning. RNNs remember the prior inputs while generating outputs. RNNs take at least one input vectors that produce output vectors impacted by hidden state vectors dependent on earlier sources of input and output, as shown in FIGURE 11. They provide a smart method for managing the sequential data that gives co-relations between data points, which are close in the sequence [171]. The information captured by the RNN relies upon the structure and training algorithm it implements. [172] in his work used RNN by assuming

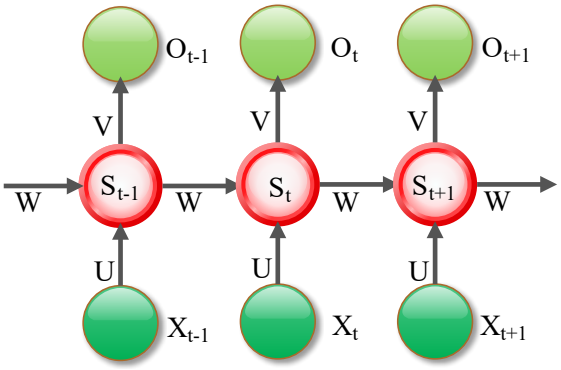

FIGURE 11: A typical RNN architecture.

the eucledian metric which Arecords the distance between two frame sequences. They used RNN with one hidden layer consisting of 150 unidirectional Long-Short Term Memory fully interconnected cells. They give 1500 frame vectors to the input layer. They trained the RNN with Adam Optimizer with learning rate=0.0001.

Zhang *et al.* [36] proposed a PHRNN (Part-based Hierarchical bidirectional Recurrent Neural Network) to classify the different facial emotions. Their model extracts facial features using PHRNN, which extracts the temporal features and used a multi-signal CNN (MSCNN) to extract facial emotion features from the still frames. Their model, Spatial-Temporal Network, mainly consists of PHRNN and MSCNN. Their model has the following stages of interest- PHRNN, MSCNN, and Model Fusion. They tested their model on Oulu-CASIA, MMI and CK+. Their model performs well on Oulu-CASIA, MMI and CK+ and diminishes the error rates from the early trial. They also compared different models- CNN, MSCNN, PHRNN-MSCNN (without sorting item), and PHRNN-MSCNN (with sorting item) with the assessed accuracy of 93.4, 95.7, 96.7, and 98.5 respectively. The results of their model were better than most of the other strategies.

FIGURE 12 shows the comparative analysis of the accuracies of various state-of-the-art approaches on different datasets.

VI. OPEN ISSUES AND RESEARCH CHALLENGES

FER has been an active research area in recent years. There are various works that have shown tremendous results and classified the emotions accurately. Yet there are various challenges and issues which are faced during facial sentiment analysis. In this section, we will discuss about various issues and challenges faced by FER. We analyzed various survey papers and understood the issues.

A. OCCLUSION AND OCCLUDED DATA COLLECTION

It is the major obstacle that comes in the way of automatic facial expression. Most of the current works are on the JAFFE, CK+ datasets without occlusion, and also with artificially occluded faces. There is a lack of datasets that include natural facial occlusion. So, there should be a creation of databases that has occlusion, which is a time-consuming and difficult task to do. Datasets should be prepared by deciding what or where the face should be occluded. Certain crucial parts of the image should not have an occluded region. The effective training and testing of the occluded dataset still remain a big challenge [4].

It is still a challenging task to collect spontaneous expressions under occlusion. The everyday human emotions such as happiness, surprise, and sadness can be easily evoked, but emotions such as curiosity, attentiveness are still difficult to evoke, particularly under occlusion. Therefore several strategies should be considered which induce emotions that are precise and contextual dependent [26]. These strategies might bring challenges in implementation on selection and limit the types of occluded data collected.

After the collection of the occluded dataset, it's effective training, and testing remains a big issue. The occluded region, the level of occlusion, the type of occlusion, the components, materials that are present in the occluded region pose a challenge for the FER System. One way to counter this challenge is to use raw pixels of the occluded region, but

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993803, IEEE Access

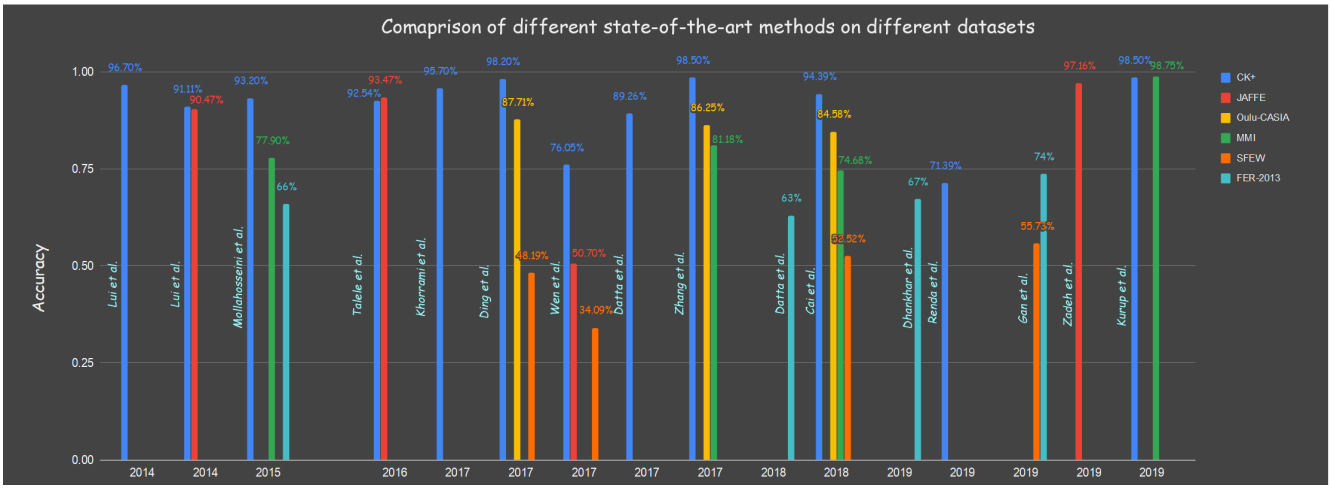

FIGURE 12: Accuracy of proposed models on different datasets vs year

enough information on the specific features of the image may not be recorded. Reliably determining the special parameters such as type, materials, components, location is a critical component of FER.

There are many ways to detect occlusion in an image. Still, the most current feature extraction techniques extract features from the face directly without passing through a pre-processing layer of occlusion detection. Huang et al. [173] in his work, showed that the accuracy is improved if we include a pre-processing layer for occlusion detection.

B. DATASETS IN FER

The other challenge in FER is the lack of proper training dataset in terms of both quality and quantity. The dataset should include images of people of all age groups as different age groups exhibit emotions differently. There are datasets that have images of a particular age group, but no dataset has a mixture of all the age groups [7]. This dataset, if developed, would assist in developing research on cross-age, cross-culture, and cross-gender.

C. FER ON 3D DATA

The current research mainly focuses on 2D FER data, which faces challenges to illumination factors and pose variations [174]. 3D face shape models are naturally robust to pose variations and illumination factors. [175] in his work, proposed CNN without facial landmark detection, which estimates expression coefficients from image intensities. Recently, many works have been proposed which combines both 2D and 3D data to improve the accuracy.

D. DIFFERENT MODALITIES IN FER

Facial Expression is only modality that can be used to recognize human behavior. The combination of other patterns like infrared images, capturing the information of 3D models, and physiological data is trending research area due to large complementary expressions. [176] employed various multi-modal affect recognition techniques.

E. FER ON INFRARED DATA

At present, the gray-scale and RGB colors are at the trend in the deep FER, but are more vulnerable to light effects. However, the infrared images records the emotions produced by skin distribution which are not subtle to the illumination variations. In 2017, Wu *et al.* [177] given a 3D CNN architecture to fuse spatial and temporal features in FER images.

F. VISUALIZATION TECHNIQUES

Adding visualization techniques [178] over the CNN model results in a quantitative analysis of how it contributes to the visualization-based rule of FER and also figures out which part of the face has more discerning information. Its results indicate the activations of filters with strong correlation to the face mark regions which correspond to a particular Action Unit. In 2016, Mousavi *et al.* [179] used the concept of visualization techniques and proposed a new visualization technique LIPNet.

G. OTHER ISSUES

Various other issues have risen based on the prototypical expression categories, namely real versus fake emotion recognition challenge and complementary emotion recognition problem. Also, the apps for real-time FER is still a challenging task [180]. Many DL techniques have been applied regarding the above problems.

VII. CONCLUSION

This paper presents a detailed systematic survey to analyze current state-of-the-art approaches for facial emotion recognition in static images and various parameters that influence the results of these approaches. We have developed a taxonomy based on different methods used for face detection, feature extraction, and emotion classification. Various facial

expression databases used as input for the FER are discussed. We have reviewed previous works on this field and concluded that much of the work had been done in this field. We have compared various detection, extraction, and classification approaches and concluded that which approach is more prominent in achieving better performance in available computation power. By discussing current issues and research challenges in the future, we concluded that there is still much research needed in this field, such as FER in 3D face shape models, recognizing emotion in images under occlusion, etc. Real-time FER is still a challenging task. In the future, we would like to survey the FER problem in videos using more advanced DL techniques.

REFERENCES

[1] A. Kumari, S. Tanwar, S. Tyagi, and N. Kumar, "Fog computing for healthcare 4.0 environment: Opportunities and challenges," Computers and Electrical Engineering, vol. 72, pp. 1 – 13, 2018.

[2] J. Hathaliya, P. Sharma, S. Tanwar, and R. Gupta, "Blockchain-based remote patient monitoring in healthcare 4.0," in 2019 IEEE 9th International Conference on Advanced Computing (IACC), pp. 87–91, Dec 2019.

[3] J. Vora, P. DevMurari, S. Tanwar, S. Tyagi, N. Kumar, and M. S. Obaidat, "Blind signatures based secured e-healthcare system," in 2018 International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 1–5, July 2018.

[4] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, "Facial expression analysis under partial occlusion: A survey," ACM Comput. Surv., vol. 51, Apr. 2018.

[5] D. Matsumoto, "More evidence for the universality of a contempt expression," Motivation and Emotion, vol. 16, no. 4, pp. 363–368, 1992.

[6] T. Amano, "Coded facial expression," in SIGGRAPH ASIA 2016 Emerging Technologies, SA '16, (New York, NY, USA), Association for Computing Machinery, 2016.

[7] S. Li and W. Deng, "Deep facial expression recognition: A survey," arXiv preprint arXiv:1804.08348, 2018.

[8] T. Abhishree, J. Latha, K. Manikantan, and S. Ramachandran, "Face recognition using gabor filter based feature extraction with anisotropic diffusion as a pre-processing technique," Procedia Computer Science, vol. 45, pp. 312–321, 2015.

[9] R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, "Tactile internet and its applications in 5g era: A comprehensive review," International Journal of Communication Systems, vol. 32, no. 14, p. e3981, 2019. e3981 dac.3981.

[10] R. Gupta, S. Tanwar, S. Tyagi, N. Kumar, M. S. Obaidat, and B. Sadoun, "Habits: Blockchain-based telesurgery framework for healthcare 4.0," in 2019 International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 1–5, Aug 2019.

[11] J. Vora, A. Nayyar, S. Tanwar, S. Tyagi, N. Kumar, M. S. Obaidat, and J. J. P. C. Rodrigues, "Bheem: A blockchain-based framework for securing electronic health records," in 2018 IEEE Globecom Workshops (GC Wkshps), pp. 1–6, Dec 2018.

[12] J. J. Hathaliya, S. Tanwar, S. Tyagi, and N. Kumar, "Securing electronics healthcare records in healthcare 4.0 : A biometric-based approach," Computers & Electrical Engineering, vol. 76, pp. 398 – 410, 2019.

[13] S. Tanwar, K. Parekh, and R. Evans, "Blockchain-based electronic healthcare record system for healthcare 4.0 applications," Journal of Information Security and Applications, vol. 50, p. 102407, 2020.

[14] R. Gupta, S. Tanwar, F. Al-Turjman, P. Italiya, A. Nauman, and S. W. Kim, "Smart contract privacy protection using ai in cyber-physical systems: Tools, techniques and challenges," IEEE Access, pp. 1–1, 2020.

[15] G. Hemalatha and C. Sumathi, "A study of techniques for facial detection and expression classification," International Journal of Computer Science and Engineering Survey, vol. 5, no. 2, p. 27, 2014.

[16] D. Deodhare, Facial Expressions to Emotions: A Study of Computational Paradigms for Facial Emotion Recognition, pp. 173–198. New Delhi: Springer India, 2015.

[17] U. Asad, N. Kashyap, and S. N. Singh, "Recent advancements in facial expression recognition systems: A survey," in 2017 International Conference on Computing, Communication and Automation (ICCCA), pp. 1203–1208, May 2017.

[18] A. Baskar and T. G. Kumar, "Facial expression classification using machine learning approach: A review," in Data Engineering and Intelligent Computing, pp. 337–345, Springer, 2018.

[19] K. Chengeta and S. Viriri, "Facial expression recognition: A survey on local binary and local directional patterns," in International Conference on Computational Collective Intelligence, pp. 513–522, Springer, 2018.

[20] G. Rajeswari and P. IthayaRani, "Literature survey on facial expression recognition techniques," in 2018 3rd International Conference on Communication and Electronics Systems (ICCES), pp. 137–142, Oct 2018.

[21] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," IEEE Transactions on Affective Computing, vol. 10, pp. 325–347, July 2019.

[22] S. Bhattacharya and M. Gupta, "A survey on: Facial emotion recognition invariant to pose, illumination and age," in 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), pp. 1–6, Feb 2019.

[23] A. S. Vyas, H. B. Prajapati, and V. K. Dabhi, "Survey on face expression recognition using cnn," in 2019 5th International Conference on Advanced Computing Communication Systems (ICACCS), pp. 102–106, March 2019.

[24] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, pp. 1–1, 2020.

[25] A. Fathima and K. Vaidehi, "Review on facial expression recognition system using machine learning techniques," in Advances in Decision Sciences, Image Processing, Security and Computer Vision, pp. 608–618, Springer, 2020.

[26] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 1, pp. 39–58, 2008.

[27] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 6, pp. 1113–1133, 2014.

[28] P. Bhattacharya, S. Tanwar, U. Bodke, S. Tyagi, and N. Kumar, "Bindaas: Blockchain-based deep-learning as-a-service in healthcare 4.0 applications," IEEE Transactions on Network Science and Engineering, pp. 1–1, 2019.

[29] J. Vora, D. Vekaria, S. Tanwar, and S. Tyagi, "Machine learning-based voltage dip measurement of smart energy meter," in 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 828–832, Dec 2018.

[30] J. N. Bassili, "Facial motion in the perception of faces and of emotional expression.," Journal of experimental psychology: human perception and performance, vol. 4, no. 3, p. 373, 1978.

[31] C. Padgett and G. W. Cottrell, "Representing face images for emotion classification," in NIPS, 1996.

[32] Guodong Guo, S. Z. Li, and Kapluk Chan, "Face recognition by support vector machines," in Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), pp. 196–201, March 2000.

[33] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," Neural Networks, vol. 16, no. 5, pp. 555 – 559, 2003.

[34] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," IEEE Transactions on Image Processing, vol. 16, pp. 172–187, Jan 2007.

[35] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI âĂŹ15, (New York, NY, USA), p. 467âĂŞ474, Association for Computing Machinery, 2015.

[36] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," IEEE Transactions on Image Processing, vol. 26, pp. 4193–4203, Sep. 2017.

[37] W. Zheng, X. Zhou, C. Zou, and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (kcca)," IEEE transactions on neural networks, vol. 17, no. 1, pp. 233–238, 2006.

[38] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in 2010 ieee computer soci-

ety conference on computer vision and pattern recognition-workshops, pp. 94–101, IEEE, 2010.

[39] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in 2005 IEEE international conference on multimedia and Expo, pp. 5–pp, IEEE, 2005.

[40] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," Image and Vision Computing, vol. 29, no. 9, pp. 607–619, 2011.

[41] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," Image and Vision Computing, vol. 28, no. 5, pp. 807–813, 2010.

[42] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," Cognition and emotion, vol. 24, no. 8, pp. 1377–1388, 2010.

[43] N. Aifanti, C. Papachristou, and A. Delopoulos, "The mug facial expression database," in 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10, pp. 1–4, April 2010.

[44] J. M. Susskind, A. K. Anderson, and G. E. Hinton, "The toronto face database," Department of Computer Science, University of Toronto, Toronto, ON, Canada, Tech. Rep, vol. 3, 2010.

[45] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. C. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. E. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. T. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, C. Zhang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," Neural networks : the official journal of the International Neural Network Society, vol. 64, pp. 59–63, 2013.

[46] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, pp. I–I, Dec 2001.

[47] B. Xiao, "Principal component analysis for feature extraction of image sequence," in 2010 International Conference on Computer and Communication Technologies in Agriculture Engineering, vol. 1, pp. 250–253, IEEE, 2010.

[48] S. Tanwar, T. Ramani, and S. Tyagi, "Dimensionality reduction using pca and svd in big data: A comparative case study," in Future Internet Technologies and Trends (Z. Patel and S. Gupta, eds.), (Cham), pp. 116–125, Springer International Publishing, 2018.

[49] G. Kumar and P. K. Bhatia, "A detailed review of feature extraction in image processing systems," in 2014 Fourth international conference on advanced computing & communication technologies, pp. 5–12, IEEE, 2014.

[50] Neha, P. Mathur, and S. K. Gupta, "Performance analysis of feature extraction techniques for facial expression recognition," International Journal of Computer Applications, vol. 166, pp. 1–3, 2017.

[51] K. Cho and S. M. Dunn, "Learning shape classes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 9, pp. 882–888, 1994.

[52] S. Ke-Chen, Y. Yun-Hui, C. Wen-Hui, and X. Zhang, "Research and perspective on local binary pattern," Acta Automatica Sinica, vol. 39, no. 6, pp. 730–744, 2013.

[53] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," 1998.

[54] R. S. Jadhav and P. Ghadekar, "Content based facial emotion recognition model using machine learning algorithm," in 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), pp. 1–5, Dec 2018.

[55] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering âĂŞ a systematic literature review," Information and Software Technology, vol. 51, no. 1, pp. 7 – 15, 2009.

[56] P. Mehta, R. Gupta, and S. Tanwar, "Blockchain envisioned uav networks: Challenges, solutions, and comparisons," Computer Communications, vol. 151, pp. 518 – 538, 2020.

[57] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[58] P. Lucey, J. F. Cohn, T. Kanade, J. M. Saragih, Z. Ambadar, and I. A. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94–101, 2010.

[59] S. Li and W. Deng, "Deep facial expression recognition: A survey," ArXiv, vol. abs/1804.08348, 2018.

[60] "Japanese female facial expressions (jaffe)." https://zenodo.org/record/3451524.

[61] "Extended cohn-kanade (ck+)." http://www.consortium.ri.cmu.edu/ckagree/.

[62] "Mmi." https://mmifacedb.eu/.

[63] "Oulu-casia." http://www.cse.oulu.fi/CMV/Downloads/Oulu-CASIA/.

[64] "Multi-pie." http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html.

[65] "Multimedia understanding group (mug)." https://mug.ee.auth.gr/fed/.

[66] "Toronto faces dataset (tfd)." https://josh@mplab.ucsd.edu.

[67] "Radbound faces database (rafd)." http://www.socsci.ru.nl:8180/RaFD2/RaFD.

[68] "Fer-2013." https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data.

[69] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2106–2112, IEEE, 2011.

[70] "Sfew (emotiw)." https://josh@mplab.ucsd.edu.

[71] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, 2017.

[72] "Affectnet." http://mohammadmahoor.com/affectnet/.

[73] R. Kosti, J. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[74] "Context is important to recognize emotions." http://sunai.uoc.edu/emotic/.

[75] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205, April 1998.

[76] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," Cognition and Emotion, vol. 24, no. 8, pp. 1377–1388, 2010.

[77] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," 2005 IEEE International Conference on Multimedia and Expo, pp. 5 pp.–, 2005.

[78] M. F. Valstar and M. Pantic, "Induced disgust , happiness and surprise : an addition to the mmi facial expression database," 2010.

[79] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikÃď inen, "Facial expression recognition from near-infrared videos," Image and Vision Computing, vol. 29, no. 9, pp. 607 – 619, 2011.

[80] J. Jayalekshmi and T. Mathew, "Facial expression recognition and emotion classification system for sentiment analysis," in 2017 International Conference on Networks Advances in Computational Technologies (NetACT), pp. 1–8, July 2017.

[81] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5325–5334, 2015.

[82] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), pp. 118–126, May 2017.

[83] N. Lopes, A. Silva, S. R. Khanal, A. Reis, J. Barroso, V. Filipe, and J. Sampaio, "Facial emotion recognition in the elderly using a svm classifier," in 2018 2nd International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW), pp. 1–5, June 2018.

[84] M. Chaudhari, M. Deshmukh, G. Ramrakhiani, and R. Parvatikar, "Face detection using viola jones algorithm and neural networks," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–6, 2018.

[85] N. B. Kar, K. S. Babu, A. K. Sangaiah, and S. Bakshi, "Face expression recognition system based on ripplet transform type ii and least square svm," Multimedia Tools and Applications, vol. 78, no. 4, pp. 4789–4812, 2019.

[86] H. M. Shah, A. Dinesh, and T. S. Sharmila, "Analysis of facial landmark features to determine the best subset for finding face orientation," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), pp. 1–4, 2019.

[87] Y.-Q. Wang, "An Analysis of the Viola-Jones Face Detection Algorithm," Image Processing On Line, vol. 4, pp. 128–148, 2014.

[88] W. LU and M. YANG, "Face detection based on viola-jones algorithm applying composite features," in 2019 International Conference on Robots Intelligent System (ICRIS), pp. 82–85, June 2019.

[89] B. Islam, F. Mahmud, and A. Hossain, "Facial expression region segmentation based approach to emotion recognition using 2d gabor filter and multiclass support vector machine," in 2018 21st International Conference of Computer and Information Technology (ICCIT), pp. 1–6, Dec 2018.

[90] Y. Luo, C. ming Wu, and Y. Zhang, "Facial expression recognition based on fusion feature of pca and lbp with svm," Optik - International Journal for Light and Electron Optics, vol. 124, no. 17, pp. 2767 – 2770, 2013.

[91] Carnap, Hilbert, Ackermann, Russell, and Whitehead, "A logical calculus of the ideas immanent in nervous activity," Jan 1970.

[92] S.-s. Liu and Y.-t. Tian, "Facial expression recognition method based on gabor wavelet features and fractional power polynomial kernel pca," in Advances in Neural Networks - ISNN 2010 (L. Zhang, B.-L. Lu, and J. Kwok, eds.), (Berlin, Heidelberg), pp. 144–151, Springer Berlin Heidelberg, 2010.

[93] H.-F. Huang and S.-C. Tai, "Facial expression recognition using new feature extraction algorithm," ELCVIA: electronic letters on computer vision and image analysis, vol. 11, no. 1, pp. 41–54, 2012.

[94] S. Biswas and J. Sil, "Facial expression recognition using modified local binary pattern," in Computational Intelligence in Data Mining - Volume 2 (L. C. Jain, H. S. Behera, J. K. Mandal, and D. P. Mohapatra, eds.), (New Delhi), pp. 595–604, Springer India, 2015.

[95] "Parallel scale invariant feature transform based approach for facial expression recognition," in Creativity in Intelligent Technologies and Data Science (A. Kravets, M. Shcherbakov, M. Kultsova, and O. Shabalina, eds.), (Cham), pp. 621–636, Springer International Publishing, 2015.

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

[97] N. Mehta and S. Jadhav, "Facial emotion recognition using log gabor filter and pca," in 2016 International Conference on Computing Communication Control and automation (ICCUBEA), pp. 1–5, IEEE, 2016.

[98] M. Sajjad, A. Shah, Z. Jan, S. I. Shah, S. W. Baik, and I. Mehmood, "Facial appearance and texture feature-based robust facial expression recognition framework for sentiment knowledge discovery," Cluster Computing, vol. 21, no. 1, pp. 549–567, 2018.

[99] A. Srivastava, S. Mane, A. Shah, N. Shrivastava, and B. Thakare, "A survey of face detection algorithms," in 2017 International Conference on Inventive Systems and Control (ICISC), pp. 1–4, Jan 2017.

[100] R. Ravi, S. Yadhukrishna, et al., "A face expression recognition using cnn & lbp," in 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 684–689, IEEE, 2020.

[101] A. L. A. Ramos, B. G. Dadiz, and A. B. G. Santos, "Classifying emotion based on facial expression analysis using gabor filter: A basis for adaptive effective teaching strategy," in Computational Science and Technology, pp. 469–479, Springer, 2020.

[102] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, pp. 971–987, 2002.

[103] V. Takala, T. Ahonen, and M. Pietikäinen, "Block-based methods for image retrieval using local binary patterns," in Image Analysis (H. Kalviainen, J. Parkkinen, and A. Kaarna, eds.), (Berlin, Heidelberg), pp. 882–891, Springer Berlin Heidelberg, 2005.

[104] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, pp. 2037–2041, Dec 2006.

[105] M. Guo, X. Hou, Y. Ma, and X. Wu, "Facial expression recognition using elbp based on covariance matrix transform in klt," Multimedia Tools Appl., vol. 76, p. 2995âĂŞ3010, Jan. 2017.

[106] d. Huang, C. Shan, M. Ardabilian, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," IEEE Transactions on Systems, Man, and Cybernetics, Part C, vol. 41, pp. 765–781, 11 2011.

[107] K. Verma and A. Khunteta, "Facial expression recognition using gabor filter and multi-layer artificial neural network," in 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), pp. 1–5, Aug 2017.

[108] J. Ilonen, J. Kämäräinen, and H. Kälviäinen, "Efficient computation of gabor,"

[109] K. Verma and A. Khunteta, "Facial expression recognition using gabor filter and multi-layer artificial neural network," in 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), pp. 1–5, IEEE, 2017.

[110] A. B. Watson, "Image compression using the discrete cosine transform," Mathematica journal, vol. 4, no. 1, p. 81, 1994.

[111] E. Feig and S. Winograd, "Fast algorithms for the discrete cosine transform," IEEE Transactions on Signal processing, vol. 40, no. 9, pp. 2174–2193, 1992.

[112] S. Dabbaghchian, A. Aghagolzadeh, and M.-S. Moin, "Feature extraction using discrete cosine transform for face recognition," pp. 1 – 4, 03 2007.

[113] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," IEEE transactions on Computers, vol. 100, no. 1, pp. 90–93, 1974.

[114] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.

[115] D. G. Lowe, "Object recognition from local scale-invariant features," in Proceedings of the seventh IEEE international conference on computer vision, vol. 2, pp. 1150–1157, Ieee, 1999.

[116] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893 vol. 1, June 2005.

[117] T. Nguyen, E.-A. Park, J. Han, D.-C. Park, and S.-Y. Min, "Object detection using scale invariant feature transform," in Genetic and Evolutionary Computing, pp. 65–72, Springer, 2014.

[118] S. Tanwar, J. Vora, S. Kaneriya, S. Tyagi, N. Kumar, V. Sharma, and I. You, "Human arthritis analysis in fog computing environment using bayesian network classifier and thread protocol," IEEE Consumer Electronics Magazine, vol. 9, no. 1, pp. 88–94, 2020.

[119] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 532–539, June 2013.

[120] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in European conference on computer vision, pp. 404–417, Springer, 2006.

[121] B. Islam, F. Mahmud, A. Hossain, P. B. Goala, and M. S. Mia, "A facial region segmentation based approach to recognize human emotion using fusion of hog lbp features and artificial neural network," in 2018 4th International Conference on Electrical Engineering and Information Communication Technology (iCEEiCT), pp. 642–646, Sep. 2018.

[122] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1805–1812, June 2014.

[123] Y. Lv, Z. Feng, and C. Xu, "Facial expression recognition via deep learning," in 2014 International Conference on Smart Computing, pp. 303–308, IEEE, 2014.

[124] I. A. Adeyanju, E. O. Omidiora, and O. F. Oyedokun, "Performance evaluation of different support vector machine kernels for face emotion recognition," in 2015 SAI Intelligent Systems Conference (IntelliSys), pp. 804–806, Nov 2015.

[125] K. Talele, A. Shirsat, T. Uplenchwar, and K. Tuckley, "Facial expression recognition using general regression neural network," in 2016 IEEE Bombay Section Symposium (IBSS), pp. 1–6, Dec 2016.

[126] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?," in 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 19–27, Dec 2015.

[127] G. Wen, Z. Hou, H. Li, D. Li, L. Jiang, and E. Xun, "Ensemble of deep neural networks with probability-based fusion for facial expression recognition," Cognitive Computation, vol. 9, pp. 597–610, 2017.

[128] S. Datta, D. Sen, and R. Balasubramanian, "Integrating geometric and textural features for facial emotion classification using svm frameworks," in CVIP, 2016.

[129] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), pp. 302–309, May 2018.

[130] P. Dhankhar, "Resnet-50 and vgg-16 for recognizing facial emotions," 2019.

[131] A. Renda, M. Barsacchi, A. Bechini, and F. Marcelloni, "Comparing ensemble strategies for deep learning: An application to facial expression recognition," Expert Systems with Applications, vol. 136, pp. 1 – 11, 2019.

[132] Y. Gan, J. Chen, and L. Xu, "Facial expression recognition boosted by soft label with a diverse ensemble," Pattern Recognition Letters, vol. 125, pp. 105 – 112, 2019.

[133] 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), pp. 577–581, Feb 2019.

[134] A. R. Kurup, M. Ajith, and M. M. RamÃşn, "Semi-supervised facial expression recognition using reduced spatial features and deep belief networks," Neurocomputing, vol. 367, pp. 188 – 197, 2019.

[135] E. Pranav, S. Kamal, C. S. Chandran, and M. Supriya, "Facial emotion recognition using deep convolutional neural network," in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 317–320, IEEE, 2020.

[136] R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, "Machine learning models for secure data analytics: A taxonomy and threat model," Computer Communications, vol. 153, pp. 406 – 440, 2020.

[137] M. Mathieu, M. Henaff, and Y. LeCun, "Fast training of convolutional networks through ffts," CoRR, vol. abs/1312.5851, 2013.

[138] S. Anwar, K. Hwang, and W. Sung, "Structured pruning of deep convolutional neural networks," J. Emerg. Technol. Comput. Syst., vol. 13, Feb. 2017.

[139] D. Mungra, A. Agrawal, P. Sharma, S. Tanwar, and M. S. Obaidat, "Pratit: a cnn-based emotion recognition system using histogram equalization and data augmentation," Multimedia Tools and Applications, vol. 79, no. 3, pp. 2285–2307, 2020.

[140] D. Scherer, A. C. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in ICANN, 2010.

[141] M. D. Zeiler and R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks," CoRR, vol. abs/1301.3557, 2013.

[142] Q. Zhao, S. Lyu, B. Zhang, and W. Feng, "Multiactivation pooling method in convolutional neural networks for image recognition," Wireless Communications and Mobile Computing, vol. 2018, pp. 8196906:1–8196906:15, 2018.

[143] C.-L. Zhang, J.-H. Luo, X.-S. Wei, and J. Wu, "In defense of fully connected layers in visual representation transfer," in PCM, 2017.

[144] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in NIPS, 2012.

[145] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in 2017 International Conference on Engineering and Technology (ICET), pp. 1–6, Aug 2017.

[146] Y. Sun, W. Zhang, H. Gu, C. Liu, S. Hong, W. Xu, J. Yang, and G. Gui, "Convolutional neural network based models for improving super-resolution imaging," IEEE Access, vol. 7, pp. 43042–43051, 2019.

[147] R. Memisevic, K. R. Konda, and D. Krueger, "Zero-bias autoencoders and the benefits of co-adapting features," CoRR, vol. abs/1402.3337, 2014.

[148] T. L. Paine, P. Khorrami, W. Han, and T. S. Huang, "An analysis of unsupervised pre-training in light of recent advances," CoRR, vol. abs/1412.6597, 2014.

[149] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Commun. ACM, vol. 60, p. 84âĂŞ90, May 2017.

[150] A. Gudi, "Recognizing semantic features in faces using deep learning," ArXiv, vol. abs/1512.00743, 2015.

[151] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2014.

[152] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition supplementary materials," 2016.

[153] D. Eigen, J. T. Rolfe, R. Fergus, and Y. LeCun, "Understanding deep architectures using a recursive convolutional network," CoRR, vol. abs/1312.1847, 2013.

[154] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3d facial expression recognition: A comprehensive survey," Image Vision Comput., vol. 30, p. 683âĂŞ697, Oct. 2012.

[155] S. Wan and J. K. Aggarwal, "Spontaneous facial expression recognition: A robust metric learning approach," Pattern Recogn., vol. 47, p. 1859âĂŞ1868, May 2014.

[156] P. Thakkar, K. Varma, V. Ukani, S. Mankad, and S. Tanwar, "Combining user-based and item-based collaborative filtering using machine learning," in Information and Communication Technology for Intelligent Systems (S. C. Satapathy and A. Joshi, eds.), (Singapore), pp. 173–180, Springer Singapore, 2019.

[157] S. Kaneriya, S. Tanwar, S. Buddhadev, J. P. Verma, S. Tyagi, N. Kumar, and S. Misra, "A range-based approach for long-term forecast of weather using probabilistic markov model," in 2018 IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1–6, 2018.

[158] R. Lysiak, M. Kurzynski, and T. Woloszynski, "Optimal selection of ensemble classifiers using measures of competence and diversity of base classifiers," Neurocomput., vol. 126, p. 29âĂŞ35, Feb. 2014.

[159] B. Kim, S. Dong, J. Roh, G. Kim, and S. Lee, "Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach," in 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1499–1508, June 2016.

[160] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," CoRR, vol. abs/1502.03167, 2015.

[161] V. Vapnik, The nature of statistical learning theory. Springer science & business media, 2013.

[162] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: A stepwise procedure for building and training a neural network," in Neurocomputing: Algorithms, Architectures and Applications (F. Fogelman Soulié and J. Hérault, eds.), vol. F68 of NATO ASI Series, pp. 41–50, Springer-Verlag, 1990.

[163] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," IEEE transactions on neural networks, vol. 13 2, pp. 415–25, 2002.

[164] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin dags for multiclass classification," in NIPS, 1999.

[165] S. Tanwar, Q. Bhatia, P. Patel, A. Kumari, P. K. Singh, and W. Hong, "Machine learning adoption in blockchain-based smart applications: The challenges, and a way forward," IEEE Access, vol. 8, pp. 474–488, 2020.

[166] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," The bulletin of mathematical biophysics, vol. 5, no. 4, pp. 115–133, 1943.

[167] R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, "Tactile-internet-based telesurgery system for healthcare 4.0: An architecture, research challenges, and future directions," IEEE Network, vol. 33, pp. 22–29, Nov 2019.

[168] H. Vachhani, M. S. Obiadat, A. Thakkar, V. Shah, R. Sojitra, J. Bhatia, and S. Tanwar, "Machine learning based stock market analysis: A short survey," in Innovative Data Communication Technologies and Application (J. S. Raj, A. Bashar, and S. R. J. Ramson, eds.), (Cham), pp. 12–26, Springer International Publishing, 2020.

[169] G. E. Hinton, "Deep belief networks," Scholarpedia, vol. 4, no. 5, p. 5947, 2009.

[170] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[171] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Transactions on Signal Processing, vol. 45, pp. 2673–2681, Nov 1997.

[172] A. Mostafa, M. I. Khalil, and H. Abbas, "Emotion recognition by facial features using recurrent neural networks," in 2018 13th International Conference on Computer Engineering and Systems (ICCES), pp. 417–422, Dec 2018.

[173] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Towards a dynamic expression recognition system under facial occlusion," Pattern Recognit. Lett., vol. 33, pp. 2181–2191, 2012.

[174] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 12, pp. 1424–1445, 2000.

[175] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "Expnet: Landmark-free, deep, 3d facial expressions," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 122–129, IEEE, 2018.

[176] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, pp. 3–9, 2017.

[177] a. C. T. Wu, Zhan, Y. Chen, Z. Zhang, and G. Liu, "Nirexpnet: Three-stream 3d convolutional neural network for near-infrared facial expression recognition," Applied Sciences, vol. 7, no. 11, p. 1184, 2017.

[178] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision, pp. 818–833, Springer, 2014.
[179] N. Mousavi, H. Siqueira, P. Barros, B. Fernandes, and S. Wermter, "Understanding how deep neural networks learn face expressions," in 2016 International Joint Conference on Neural Networks (IJCNN), pp. 227–234, IEEE, 2016.
[180] I. Song, H.-J. Kim, and P. B. Jeon, "Deep learning for real-time robust facial expression recognition on a smartphone," in 2014 IEEE International Conference on Consumer Electronics (ICCE), pp. 564–567, IEEE, 2014.