IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 27, 2021, accepted September 13, 2021, date of publication September 16, 2021, date of current version October 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3113464

Facial Expression Recognition: A Review of Trends and Techniques

OLUFISAYO S. EKUNDAYO AND SERESTINA VIRIRI, (Senior Member, IEEE)
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban 4041, South Africa

Corresponding author: Serestina Viriri (viriris@ukzn.ac.za)

This work was supported by the University of KwaZulu-Natal, South Africa.

ABSTRACT Facial Expression Recognition (FER) is presently the aspect of cognitive and affective computing with the most attention and popularity, aided by its vast application areas. Several studies have been conducted on FER, and many review works are also available. The existing FER review works only give an account of FER models capable of predicting the basic expressions. None of the works considers intensity estimation of an emotion; neither do they include studies that address data annotation inconsistencies and correlation among labels in their works. This work first introduces some identified FER application areas and provides a discussion on recognised FER challenges. We proceed to provide a comprehensive FER review in three different machine learning problem definitions: Single Label Learning (SLL)- which presents FER as a multiclass problem, Multilabel Learning (MLL)- that resolves the ambiguity nature of FER, and Label Distribution Learning- that recovers the distribution of emotion in FER data annotation. We also include studies on expression intensity estimation from the face. Furthermore, popularly employed FER models are thoroughly and carefully discussed in handcrafted, conventional machine learning and deep learning models. We finally itemise some recognise unresolved issues and also suggest future research areas in the field.

INDEX TERMS Facial expression recognition, single label learning, multilabel label learning, label distribution learning.

I. INTRODUCTION

Facial Expression Recognition (FER) has gained remarkable attention in computing, which is not limited to Computer Vision (CV) and Human-Computer Interaction (HCI). The advancement in technology and the aim to achieve machine-human communication encourage many researchers to explore the field in more than two decades. FER is about detecting human affective states due to responses observed in a face through facial muscles movement due to involuntary action triggered by changes in human emotional states. From the psychological point of view, the categories of human emotional states are into six basic emotions; sad, happy, fear, surprise, anger and disgust [1]. According to the study conducted by [2], facial expression carried a larger percentage of communication information in man than any other non-verbal medium like hand gesture, body gesture, and text [3], [4]. A man without difficulty can easily interpret expression display in the face, but the automation of this task in the machine remains a challenge [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

FER is a combination of two significant fields or disciplines (Psychology and technology). In Psychology [5], [6], facts about facial responses to emotional changes are thoroughly studied and established. Likewise, applying technology employed image processing concepts (Computer Vision) and machine learning techniques to achieve automation. FER's general architecture comprises three major phases; pre-processing, feature extraction, and classification or recognition. These phases carry out their respective tasks sequentially on a particular FER database to establish ground truth for the system to achieve its goal. Details of the FER architecture description is available in Figure 6.

FER's automation comes in two main procedures; feature extraction methods and feature classification methods. However, it is advisable to carry out some data engineering techniques before applying these methods or both accordingly. Achieving a robust system is the goal of FER. Nevertheless, FER's automation is challenged with some factors like; intensity, occlusion, facial tribal mark or accidental facial mark, face morphology, age, to mention a few. FER's emotion recognition has various applications: medicine, psychology,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

security, clinical investigation of neuropsychiatric disorders (affective disorder or schizophrenia).

The quest for adequate recognition of man affects state led to the evolution of several approaches in developing a FER system. Existing FER review works [7]–[10] have diversely presented comprehensive studies on the traditional FER implementation methods, including the handcrafted techniques and the machine learning algorithms. Likewise, different overview studies of deep learning methods approach to FER have been presented in FER literature. The works concentrated mainly on different methods for a robust and efficient FER model. Virtually all the works provided information about the databases in the field, but studies on FER data annotations have not been given adequate consideration, which is the motivation for this work. This study considers studies that proposed methods to resolve FER data annotation inconsistency and label ambiguity. This work presents FER in three different machine learning problem definitions, which include: Single Label Learning (SLL) (Multiclass problem), Multilabel Learning (MLL): where a FER image contains one or more basic emotions. Another approach is Label Distribution Learning (LDL), which proportionally estimate all the basic emotions present in facial expression image. SLL also consider estimation of the intensity of a recognised emotion available in the expression image. No review literature in the field includes studies that consider expression intensity estimation, label discrepancies and ambiguity, and correlation among labels in their work to the best of our knowledge. The uniqueness of this work include:

- The review of existing FER application areas and suggestions of possible FER application environments to explore. This information is necessary as a quick guide or enlightenment for interested researchers in the field.
- Review of identified problems in the field that affect system performance and provides new researchers with possible challenges to consider for efficient model development.
- Review of FER literature and their classification into three groups of machine learning problem definitions, SLL: contains methods that consider FER tasks a multiclass problem. MLL: FER approach that resolves the ambiguous nature of FER data. Lastly, LDL: FER methods for label annotation inconsistency and correlation among labels.
- Provide a thorough review study on traditional FER classification models and modern deep learning models. Although literature in the field considered these separately, reviews have been presented on conventional machine learning models or deep learning models. Nevertheless, the integration of both in a single work is one of the uniqueness of this work. We purposely include them for new and interested researchers to have a general overview of what has been done in the field.

This work is organised as follows; In Section II, we discuss some FER application areas. Section III illustrates some of the challenges to be considered while developing the system to achieve a robust system with excellent performance. Section IV presents information about the available FER databases. Comprehensive FER literature studies that present FER in a different category of problem definitions are thoroughly and carefully presented in Section V. Likewise, Section VI illustrates some popularly used FER techniques in their group of handcrafted, Machine learning algorithms and state-of-the-art deep learning methods. Section VII presents a general discussion and opens up some unresolved research issues and future research areas. The last section, which is section VIII is the conclusion of this review work.

II. APPLICATION OF FACIAL EEXPRESSION RECOGNITION

There is still no limit to FER's application, and it spans through every facet in which natural interaction between man and machine is achievable. This section considers some of the areas of FER applications.

A. SOFTWARE DEVELOPMENT

The goal of every software is to meet or satisfying the requirement elicitations of end-users. Software usability is one of the means of determining the degree of satisfaction through feedback from end-users. The traditional way of measuring user satisfaction is by administering a questionnaire, but Kolakowaska *et al.* [11] believe that a questionnaire may be biased and misleading. They introduce FER as part of multimodal inputs for software usability testing and research on finding the relationship between software developers and Job's quality delivery within a particular time frame. The study's outcome shows that developers' emotions affect software productivity and quality, and they suggested incorporating an emotion detection mechanism in the HCI system.

B. EDUCATION

Education is one of the backbones of a country's economic sectors. Therefore, practical knowledge dissemination and appropriate learning are inevitable. Every institution's learning process requires thorough monitoring and proper feedback from both the learners and the instructors. The traditional methods of using surveys via questionnaire and interview have their limitations. Some factors inhibit knowledge transfer in the learning system, according to the emotional state of an individual involved [12]. These factors should be investigated regarding the assessment of learners' emotional state, evaluation of educational resources in a virtual institution and distance learning environment, and usability testing of educational tools [11]. The most appropriate means of achieving excellent results from the listed experiments would be via FER. Lisetti *et al.* [13] suggested adopting FER feedback-like mechanism into a tele-teaching assistant system in a distance learning environment and claimed that this would ensure class dynamism.

Zhou *et al.* [12] proposed an e-learning FER to capture the real-time students' emotional states for timely adjustment of teaching strategies. The recent pandemic that befalls the whole world transformed teaching and learning environments from physical contact to virtual. Most of the applications employed like zoom and the likes have the challenges of capturing students affect, which is vital information in achieving class dynamism and effective teaching.

C. MEDICINE

FER is applicable to some medical fields like; neuro-psychiatric disorder, Patients treatment feedback, patient's emotion monitoring, rehabilitation, autism and music therapy [14]. Human Facial expression has been employed in investigating neuro-psychiatric disorder as it affects emotion perception, expression and recognition in affected patients [15]–[17]. The available method used by clinicians in the field is a qualitative manual method, which is more subjective and human-intensive [16]. This challenge requires an objective process that possibly reduces human-intensive efforts and provides a qualitative result. Wang *et al.* [18] proposed the FER framework that derived probabilistic expression profiles for video data, and in turn, automatically quantified emotional expression differences between neuropsychiatric disorders patients and healthy controls. The advent of telemedicine [15], [16] in the medical field gives more justifications for FER's application. With the dynamic evolution and advancement experienced in technology development of communication devices and mobile applications such as a computer, mobile devices, video chat applications, to mention a few, could be explored using FER technology that employs facial cues to determine users' emotions in real-time.

D. SECURITY

Application of FER into identity recognition system will strengthen and improves the functionalities of the system. Biometric systems (face recognition) designs for identity authentication, and its application to security, access control, forensic and so on had been successfully achieved. Likewise, a security surveillance system saddled with the responsibility of monitoring an environment has the capability of providing detailed information on events within a specified time frame. Security surveillance System and biometric Security inclined system has the limitation of not preventing the environment from experiencing imminent attack from enemies. Adding FER to these systems will incorporate a layer of security intelligence to detect enemies' intention [19] through the emotion displays and alert the security personnel. [20] proposed improving surveillance systems by incorporating FER to make a system that would detect a person with malicious intentions from their facial expression and report to the securities before the perpetration of the intended evil. There is a need for this type of intelligent surveillance in public places like Shopping malls, Sports arenas, airports, and other places where people's gathering is encouraged.

E. MARKETING

The heartbeat of any company or business organisation is marketing, and it includes market research and advertising. The market research department could either use an interview or questionnaire, a traditional means of collecting information about users' opinions. This conventional means, according to [21], is facing out of effectiveness. Another method is to capture a user's behaviour using a sample of the product [22]. The later approach needs to carry out video analysis by experts. The method is capital and human-intensive. The cost of a behavioural approach could be minimised by employing a FER system for video analysis tasks. Yolcu *et al.* [23] developed a non-invasive deep learning-based system for monitoring customers' interest and advertisement acceptance rating. This method is more objective and reliable for adequate decision-making than the traditional way users formulate their preferences, which often mislead the research team. The advertising department could also incorporate FER into the analysis of public opinion towards various advertisement approaches. With FER, they could concentrate on the advertisement that captures more attention with positive responses.

F. ROBOTICS AND GAMES

Personal assistant robot tasks could be extended to exhibit a human-like interaction, and in most cases, they discharge their respective duties accordingly if they are embedded with a sensor that could interpret the boss's facial expression. Games or computer games should explore automatic FER thoroughly and develop game applications with characters that display affective states applicably and accordingly. It would also be of more interest if a game application could capitalise on FER for its dynamism. It should be from the user's facial expression to detect the user's feelings and trigger an action to meet the user's satisfaction.

Other areas of FER's application include image and video information retrieval, forensic investigation (Lie detector) [24], stress and depression management [25], Driver's monitoring agent in automobile [26], a fear detector at real-time in the critical mission, real-time expression recognition in mobile digital devices, temperament detection, a job interview and many more. Some of the suggested application areas have been deployed already by companies like Affestiva, EmoVu, Kairos, Nviso, Sightcorp and many more.

III. FACIAL EXPRESSION RECOGNITION CHALLENGES

FER is like many recognition systems, where intraclass variation minimisation and interclass variation maximisation are critical together with system robustness. Individual differences in a class majorly cause Intraclass variations in FER as a result of the following;

A. OCCLUSION

This is a form of the challenge posed due to disturbance or hindrances that obscure the characteristic feature from

the expression image. This problem is limited to natural occurrences like moustache and beard, and self-made like wearing glasses, cosmetics headscarf, or hijab.

B. AGEING

Age categories contribute to variations in how people express emotion through the face. For example, emotional states are observed in children's faces, obviously noticed in adults and mildly displayed in elders. Cohn *et al.* [27] in their investigation on the performance of optical flow and high gradient detection algorithm on infants, the algorithm had less performance on infants compared to its performance on adults. The degradation in performance was assumed to be due to infant skin texture, more fatty tissue, facial conformation, and the absence of transient furrows. More emphasis is given by [28], [29] that different physical appearance like skin texture affects the analysis of facial expression intensity. Tian *et al.* [30] claimed that the variation in the way people express emotion could be attributed to the degree of facial plasticity, face morphology, rate of expression and frequency of intense expression.

C. POSE AND ILLUMINATION VARIATION

the location of a face at the time of data collection could also be a challenge, in a 2D morphology; the head should be positioned in frontal view, using 2D image to reduce the computational cost, but determination of appropriate facial features is extremely difficult. However, the reverse is the case for a 3D image. A side view position could affect the performance of the system. Non-frontal view and rigid head motion are challenges peculiar to spontaneous data. Illumination variation in light direction often leads to changes in light intensity and causes a cluttered background for expression images.

Aside from the intraclass problem, interclass challenges are experienced when the differences between emotion classes are less conspicuous. For instance, the same subjects are used in each of the expression classes. Interclass variation implies that the expression classes would have more similar information than unique information in the representative features.

Nature of database: Most Facial expression databases are collected in a controlled environment; the expression images are static, acted by either professional or non-professional actors. FER developed from a monitored environment are found to degrade in performance in a real-world where spontaneous, and sequence images are available.

IV. DATABASES

Facial Expression Database is a cogent and essential aspect of the FER system; like feature extraction and classifiers, facial expression database is one factor that contributes immensely to the robustness of the FER system. The early facial expression databases were posed database collected in a controlled environment [31], [32]. The choice of database for FER development depends on the type of its application. Apart from posed databases, there are also spontaneous databases captured at the real scene- a naturally expressed facial expression database. Recently, the quest to take FER beyond the laboratory to real-world applications requires facial expression databases in an unconstrained and uncontrolled environment, also termed In-the-wild databases. Figure 1 shows some selected samples of six basic image expressions from different databases. The widely employed FER databases include;

FIGURE 1. Facial expression samples of six basic emotion from different databases.

A. BOSPHORUS DATABASE

This database is one of the prevalent 3D face databases introduced by [33] and is composed of multi-expression and multi-pose facial images together with several occlusions captured in a more realistic scene. Enriched in AU and basic recognised emotional expression, adequate ground truth head pose, incorporation of different occlusion types, and employment of skilful subjects are the benefits of the Bosphorus database. Some of the compositions of this database are summarised in Table 1. The database was developed with 105 subjects altogether under different head poses, expression display and occlusion. Sixty of the subjects were men, and 45 were women, 18 wore beard/moustache, and 15 had short hair. At the point of data collection, each of 71 members of the subject had 54 face scans, and the remaining 34 Subjects had 31 face scans for each of the subjects. Despite the thoroughness exercise in capturing the AUs and the facial expression, it was not still far from the fact that they were not natural. Also, screening the AUs and the facial expressions means that not all the AUs and facial expressions will be present for all the subjects. The stated challenges are the limitations of the Bosphorus database.

B. REAL WORLD AFFECTIVE DATABASE (RAF-DB)

Raf-DB is a crowd-sourcing face data for facial expression database, categorised into basic emotions with single modal distribution and compound emotion with a bimodal distribution. According to [34] who introduced it, the database recognised it as the first of its kind, having a large scale that provided the labels of common expression perception and compound emotion in an unconstrained environment. This

database's main advantages are; availability of sufficient data, no constrained or controlled environment for data capturing and group perceiving on facial expressions and data labels with the least noise. Raf-Db contains almost 30000 facial images collected with an image search API called Flickr, and the search is by using keywords relevant to each of the emotions. The extracted images were downloaded in batches using an automatic open-source downloader.

C. COHN KANADE AND COHN KANADE EXTENSION (CK AND CK+) DATABASE

Cohn *et al.* [32] released a facial expression database in 2000; the database contains 97 subjects between the ages of 18 and 30; 65% were female, and the remaining 35% were male. The subjects were chosen from multicultural people and races. There were 486 sequences collected from the subjects, and each sequence started from neutral expression and ended at the peak of the expression. The expressions' peak was fully FACS coded and emotion labelled, but the label was not validated. Luecy *et al.* (2010) itemised three challenges with CK databases; invalidation of emotion labels, Unavailable standard performance metrics for algorithm performance evaluation and lack of standard protocol for a standard database. Cohn *et al.* [35] identified the challenges with the CK database and proposed its extension, termed extended Cohn Kanade (CK+) database. In CK+, the number of subjects increased by 27%, the sequence by 22%. Also, there were slight changes in the metadata. The age group of the subject ranged between 18 and 50years. The percentage of the male and the female population is 31% and 69%, respectively. The emotion labels were revised and validated using the FACS investigator guide as a reference and confirmed by appropriate expert researchers. Leave-one-out subject cross-validation and area underneath the Receiver Operator Characteristics curve were proposed for Algorithm performance evaluation metrics.

D. JAPANESE FEMALE FACIAL EXPRESSION (JAFFE) DATABASE

Lyon *et al.* [31] introduced a database for facial expression called the JAFFE database; the database is one of the popularly used databases as a FER system benchmark. It contains ten subjects, which are all Japanese females. Each of the subjects produced 3 or 4 images for each of the six basic facial expressions. The corresponding subject images were captured while looking at the camera via a semi-reflective plastic sheet. The environment was controlled from occlusion, illumination variation, and head poses.

E. BINGHAMTON UNIVERSITY 3D FACIAL EXPRESSION (BU-3DFE)

This database was introduced at Binghamton University by [36] contains 100 subjects with 2500 facial expression models. Fifty-six of the subjects were female, and 44 were male. The age group ranges from 18 to 70 years old, with various ethnic/racial ancestries, including White, Black, East-Asian, Middle-east Asian, Indian, and Hispanic Latino. A 3D face scanner was used to capture seven expressions from each subject; in the process, four intensity levels were captured alongside each of the six basic prototypical expressions. Each expression shape model is associated with a corresponding facial texture image captured at two views (about +45° and -45°). As a result, the database consists of 2,500 two views' texture images and 2,500 geometric shape models.

F. BINGHAMTON UNIVERSITY 3D DYNAMIC FACIAL EXPRESSION (BU-4DFE)

BU-4DFE is a 3D dynamic facial expression database. The 3D facial expressions are captured at a video rate of 25 frames per second for each subject, and six model sequences showing six prototypic facial expressions. Each expression sequence contains about 100 frames. The database contains 606 3D facial expression sequences collected from 101 subjects, with approximately 60,600 frame models. Each of the 3D models of a 3D video sequence has a resolution of approximately 35,000 vertices. The texture video has a resolution of about 1040×1329 pixels per frame. The resulting database consists of 58 female and 43 male subjects, with various ethnic/racial ancestries, including Asian, Black, Hispanic/Latino, and White.

G. BINGHAMTON-PITTSBURGH 3D DYNAMIC SPONTANEOUS FACIAL EXPRESSION DATABASE (BP4D)

Posed and spontaneous 3D facial expressions differ along several dimensions, including complexity and timing, well-annotated 3D video of spontaneous facial behaviour is necessary. BP4D was presented by [37] as a newly developed 3D video database of spontaneous facial expressions in a different age group. The database includes forty-one subjects of 23 women and 18 men. The age ranges between 18 and 29 years; the database's cultural races are 11 Asian, 6 African-American, 4 Hispanic, and 20 Euro-American. Emotions were educed from each of the subjects using a protocol called emotion elicitation, where eight different tasks were conducted along with the interview process to deduce eight emotions.

FER databases are not limited to those discussed in this section. Information about others are briefly summarised in Table 1. [38] presented detailed information on FER databases, it's available for any interested reader.

V. FACIAL EXPRESSION RECOGNITION RESEARCH TRENDS

FER can appropriately predict individuals' emotional state from the deformation displays in the face as one of the cognitive and affective research fields. Many works have been attempted in the field to make it an achievable task. FER research has produced several models and different FER databases together with their annotations. The successes recorded so far in the literature are about FER models

TABLE 1. The summary of some FER benchmark databases.

Database	No of Actor	Population	Nature	Environment	Application	Available Expression
CK [32]	97	486	Sequence	Controlled (Lab)	Posed or Spontaneous	6 BExp + Neutral
CK+ [35]	123	593	Sequence	Controlled (Lab)	Posed or Spontaneous	6 Basic Emotion + Neutral
JAFFE [31]	10	213	Posed	Controlled (Lab)	Posed	6 Basic Emotion + Neutral
Bosphorus [33]	105	4888	Posed	Controlled (Lab)	Posed	6 Basic Emotion + Neutral
BU-3DFE [36]	100	2500	Posed	controlled (Lab)	Posed	6 Basic Emotion + Neutral
BU-4DFE [39]	101	606	Sequence	Controlled (Lab)	Posed and Sequence	6 Basic Emotion + Neutral
RafD [40]	67	1608	Posed	controlled (Lab)	Spontaneous and static	6 Basic Emotion + Neutral
FER2013 [41]	NA	35,887	Spontaneous	Uncontrolled (Internet)	Spontaneous and Static	6 Basic Emotion + Neutral
SFEW [42]	NA	1766	Vidoe clips	Uncontrolled (Movies)	Spontaneous	6 Basic Emotion + Neutral
AFEW [42]	NA	1809	video	Uncontrolled (Movies)	Spontaneous	6BExp + Neutral
BP4D [37]	41	NA	video	Uncontrolled	Spontaneous & Dynamic	6BExp + Neutral
Oulu-CASIA [43]	80	2880	Posed	Controlled(Lab)	sequence	6 Basic Emotion + Neutral
AffectNet [44]	NA	450,000	Wild	Uncontrolled (Internet)	Static or Sequence	6 Basic Emotion + Neutral
MMI [45]	NA	2900	Video	Controlled (Lab)	Sequence	6 Basic Emotion + Neutral
MMI [45]	25	740	possed	Controlled(Lab)	static	6 Basic Emotion
ExpW [46]	NA	91,793	Wild	Uncontrolled (Internet)	Static or Sequence	6 Basic Emotion + Neutral
RAF-DB [34]	NA	29,672	Wild	Uncontrolled (Internet)	Static or Sequence	Basic Emotion + Neutral + Compd
4DFAB [47]	180	1.8 Million	Posed	Controlled (Lab)	Static or sequence	6 Basic Emotions + Neutral
EmotioNet [48]	NA	450,000	Wild	Uncontrolled (Internet)	Static or Sequence	Basic + Compd Emotions

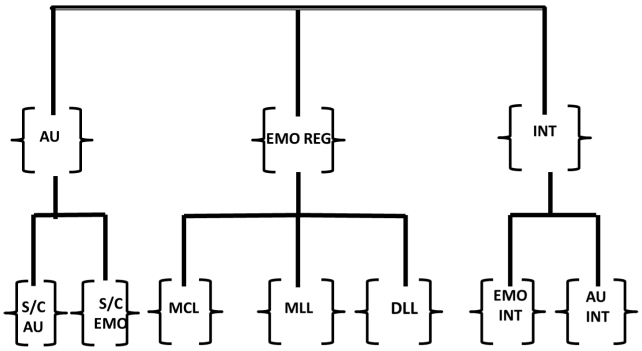

FIGURE 2. The tree structure representing the main trends of research in FER. AU stands for Action Unit, S/C for single or compound AU, S/C EMO for single or compound emotion, EMO REG for Emotion Recognition, MCL for Multi-class Learning, MLL for Multilabel Learning, LDL for Label Distribution Learning, INT EST for intensity Estimation, AU INT for AU intensity estimation, EMO INT for emotion intensity estimation.

that could predict the basic emotion from facial expression images. No consideration is given to other aspects of FER research that considered the intensity estimation of the emotion, Facial expression ambiguity, and the label inconsistency and correlation among labels. This section will present research diversities in FER as we categorise them based on machine learning problem definitions; SLL, SLL extension (FER and intensity estimation), MLL, and LDL. The trend in FER approaches to emotion recognition is pictorially presented in Figure 2. Table 2 presents the categories of emotion recognition research in FER with the associate limitations.

A. SINGLE LABEL LEARNING (MULTICLASS)

Early studies on the human cognitive and affective aspect of computer vision were pilots by the established work of [6], which introduced the six basic classes of emotion. Classifying an instance of face expression image into any of the six basic emotion states is identified as a multiclass task termed single label learning. Figure 5A illustrates how SLL reports only one emotion out of all the possible outcomes. Methods that attempt facial expression multiclass tasks are considerably presented in FER literature. These methods revolve around the handcrafted, conventional machine learning and the deep learning models, which we discuss in section 6. [7] is a comprehensive study of early methods on FER, [8]–[10] presented a review studies of the state-of-the-art (deep learning) methods. FER's scope as a multiclass task spreads across emotion recognition in various environments like; 1) static environment [49]–[52]. (2) Temporal and dynamic environment [53]–[55] and (3) In-the-wild [48], [56]. Several promising performances have been reported in the literature. Despite the SLL approach to FER's achievement, its simplification of assigning a single emotion to an expression instance limits its application in the real world. SLL fails to account for the inconsistency and ambiguity in FER data annotations and does not provide information about the intensity of the possible available emotions in an expression instance.

B. FACIAL EXPRESSION RECOGNITION AND INTENSITY ESTIMATION

Facial expression intensity estimation is the observable differences between facial expression images of the same

TABLE 2. Summary of diverse approaches to emotion recognition.

FER Task	Description	Limitation	Database	Metrics
SLL	A multiclass Problem that report emotion with the highest predicting value	No consideration for data annotation inconsistency, and correlation among labels	static (Bu-3DFE, JAFFE), in-the-wild, Dynamic data	Accuracy, F1, Precision, ROC
MLL	predict one or more possible emotion from the expression face	Capable of depicting the subjectivity in expression label but fails to illustrate the intensity to which each label describes the expression face.	RAF-ML, ML-JAFFE, HAPPEI, ML-BU-3DFE	RAkEL, ML-KNN, CLM, LIFT, ML-LOC
LDL	predict the possible emotion in expression image with appropriate proportion of their occurrence.	Although LDL has attempted the issue of label inconsistency and ambiguity in FER annotations to an appreciable length, yet research still open for better methods.	s-BU-3DFE, S-JAFFE	Kullback-Leibler, Euclidean distance, Sorensen, Fidelity, Intersection.

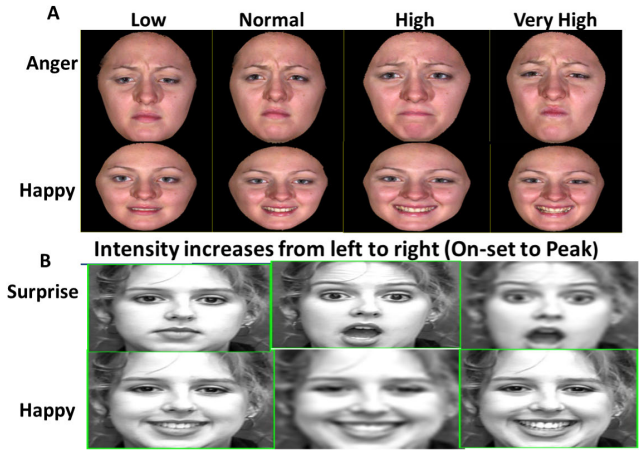

FIGURE 3. Sample A is BU-3DFE extraction (Anger, Happy) that indicates intensity displays with ordinal metrics (low, normal, high and Very_high). Sample label B is an extraction from CK+ (Surprise, Happy) showing intensity rises from ON-set to PEAK.

expression or the degree of dissimilarities of facial expression image from its reference base. One of the facial expression analysis tasks is facial expression intensity estimation; expression intensity is estimated in emotion and AUs quantifications. Figure 3 is the sample of expression intensity from static data (Figure 3A) and sequence data (Figure 3B). Some methods for FER intensity estimation have been explored in the field. Khairunmi [29] grouped these methods into; distance-based, cluster-based, regression-based, and probabilistic graphical-based.

Verma *et al.* [28] approach is a distance-based emotion intensity estimation model that uses shape transformation to capture the deformation between a template face and emotion reflected face. The deformations caused by the expansions and contractions in face regions and boundaries are quantified through elastic interpolation between the template face and expression face. The vector value generated in shape transformation is used to define a Regional Volumetric Difference (RVD) function that provides a numeric value for each of the face pixels representing the quantities of emotion displayed. Le and Xu [57] estimated facial expression intensity using isometric feature mapping. The resultant 1D manifold and facial feature trajectories are used by SVM and Cascade Neural Network (CNN) to model expression intensity. It requires that this method should conduct training for a different subject.

Observation showed that the distance-based approach quantified facial expression intensity before the recognition of the emotion. This model disagrees with how human expresses emotion.

Quan *et al.* [58] proposed a cluster-based method for expression intensity estimation. The unsupervised method employed a K-Means clustering algorithm to Haar-like features extracted from the CK+ dataset to get the K-order of the expression intensity and applied SVM classifier for the expression classification. Just like the distance-based, this approach also predicts the intensity before the expression class. Chang *et al.* [59] approach expression intensity estimation by considering the relative order information available in facial expression images. They argued that it is more appropriate and convenient to use relative order to distinguish between two expressions than considering their absolute difference. Their method employed a scattering transformation to extract discriminating and translation invariant features and used RED-SVM with Radial Basis Function (RBF) kernel for expression ranking. This method is single image-based and does not consider available temporal information.

The work of [60] is a regression-based approach, and they proposed an ensemble of naive Bayesian classifiers for expression classification and intensity estimation, respectively. They employed some naive Bayes classifiers to classify selected features weakly and generate a robust classifier from the weak classifiers' output for expression classification, and the normalised output scores are the class intensity estimation. Wu *et al.* [61] considered expression intensity estimation by quantifying energy variation of facial expression sequence. They were motivated by the possibility of quantifying energy value for each state of expression using facial landmarks. The model employed HMM to discriminate different expressions and used a linear regression algorithm to obtain intensity curves for each expression. [62] presented a regression-based model; their model utilised the ordinal information distributed in sequence image to annotate expression intensity. The proposed Ordinal Support Vector Regression model (OSVR) could generalise well in both supervised and unsupervised environments because OSVR is a combination of Support Vector Regression, which is

TABLE 3. Summary of various models for emotion and intensity recognition. NA:Not Applicable, MAE: Mean absolute error, PCC: Pearson correlation coefficient, ICC: Intraclass correlation, MAL: Mean absolute loss, HL: Hamming loss, RL: Ranking loss; AP: Average precision, CE: Coverage error.

Method	Model	DB	Performance	Limitation
Lee and Xu [57]	Optical flow tracking algorithm (Distance)	Real-time data:	88.32% (accuracy)	Need for each subject to be trained differently, not generalise, predicting intensity before emotion
Verma et al. [28]	Distance based	Primary source	NA	only few emotions are considered, method not generalise, emotion intensity before emotion recognition, computationally expensive.
Kim et al. [64]	HCORF (Prob)	CMU	89.05% (accuracy)	Intrisic topology of FER data is linearly model.
Rudovic et al. [67]	LSM-CORF (Prob)	(BU-4DFE, CK+)	(MER:19.0, 12.0), (MAL:0.36, 0.32),	Latent states are not considered in the modeling of sequences across and within the classes
Chang et al. [59]	Scartering transform + SVM (Cluster)	CK+	Mean Error: 0.313, MAE:0.318	Emotion recognition task is omitted.
Quan et al. [58]	K-Means (Cluster)	CK+	accuracy: 88.32%	Predict intensity before emotion, intensity estimation based on graphical difference is not logical
Walecki et al. [68]	VSL-CRF (Prob)	CK+, AFEW	(F1:96.7%, 28.1%), (accuracy:94.5%, 32.2%)	Result of emotion intensity is not accounted for.
Zhao et al. [62]	SVOR (Regression)	Pain DB	PCC: 0.6014, ICC: 0.5593, MAE:0.8095	Correlations between emotion classes are not modelled.
Khairuni et al. [29]	weighted vote	CK+	(Exp. acc: 82.4%), Exp. F1:69.7%, Intensity acc: 82.3%, Intensity F1: 81.8%	Emotion and emotion intensity not concurrently predicted.
Ekundayo and Viriri [69]	ML-CNN (Multi-Label)	BU-3DFE	HL:0.0628, RL:0.1561, AP:0.7637, CE:4.3140	Assume temporal information among sequence data as ordinal metrics.

responsible for intensity labels in the annotated frame and Ordinal Regression, a baseline for temporal order for frame sequence and not the label intensity values.

Probabilistic graphical-based model for emotion and intensity estimation have been thoroughly reported in [63]–[66]. [63], [65] used HMM and CRF to successfully recognise the emotion or the intensity of the target expression. [64] identified the limitation of the existing models and enhanced the discriminative ability of CRF. They proposed a Hidden Conditional Ordinal Random Field (HCORF) model to simultaneously capture multiple emotions and their respective intensities. Despite this improvement, HCORF is limited to the variations in facial expression and their respective intensities, which a simple linear model could not adequately express. Rudovic *et al.* [67] enhanced the capability of HCORF; they used ordinal manifold, a low dimensional manifold to model facial affective data topology and incorporated it into the HCORF model. The ordinal manifold preserves facial expression discriminative information and the ordinal relationship of the corresponding intensity. Walecki *et al.* [68] complimented laplacian shared parameter Multi-output CRF and HCORF and proposed Variable-state Conditional Random field method, which considered both nominal and ordinal latent state in the model of expression sequence both within and across the expression classes. They reported that the proposed method outperformed HCORF and LSM-CRF but failed to state the intensity estimation result categorically. Khairuni [29] introduced a method that employed weight voting and Hidden Markov Model for expression recognition and intensity estimation. HMM is saddled with detecting the input frame's emotion in the method, and change-point detection captured the temporal segment. The result showed that the proposed method performed better than any existing probabilistic graphical methods in accuracy

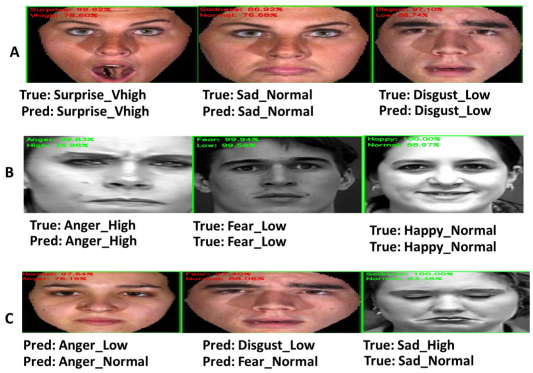

FIGURE 4. ML-CNN concurrent predictions of emotion with the associated intensity of BU-3DFE and CK+ test datasets. A and B are samples of correct predictions, and C is the samples where one of either the emotion or the intensity is incorrectly predicted.

and computation time. Our approach to FER and intensity estimation is presented in [69]. We considered FER and intensity estimation a multilabel task with the motivation that an instance of a facial expression image contains information about emotion displays and the corresponding intensity. We proposed ML-CNN (Multilabel Convolution Neural Network) that uses CNN as a binary classifier for an enhanced binary relevance model. We optimised the model with a VGG-16 pre-trained network and employed island loss to minimise intraclass and interclass variations. Our model concurrently predicts emotion and its intensity using ordinal information available in the data. The predictions of our model are presented in Figure 4. The experiments conducted on BU-3DFE and CK+ datasets produced an optimal result.

The summary of the models for emotion and intensity estimation and their corresponding evaluation are presented in Table 3.

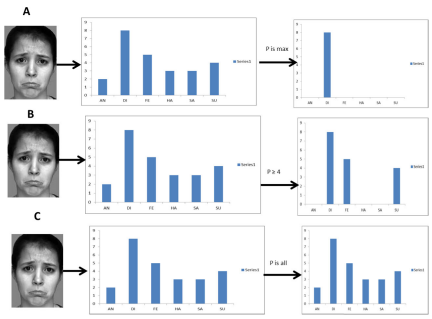

FIGURE 5. A is the description of FER Multi-class learning, where only the class with the highest prediction value becomes the identified expression. B is a FER multi-label learning scenario where more than a class with prediction value equal to or greater than a certain threshold. In C (FER distribution learning), all the expression classes are identified along with their respective prediction values.

C. MULTILABEL LEARNING

Ekman *et al.* [1], and Plutchik *et al.* [70], [71] reported that facial expression is more of a mixture of basic emotions and that a single basic expression is only displayed on a rare occasion. The argument defines the FER task as a Multilabel (ML) problem. Figure 5B shows multilabel prediction's possible output. An instance of expression image could contain one or more basic emotion information in facial expression multilabel tasks. There are few FER literature with a multilabel approach; this resulted from the few available multilabel datasets. The datasets list include; JAFFE [31] BU-3DFE [36] HAPPEI [72], EmotioNet [48] and the most recent RAF-ML [34]. One of the multilabel methods applied to FER is Group Laso Regularised Maximum Margin classifier (GLMM) proposed by [73], GLMM considered the fact that the AU at different affective states is triggered in the same region of the face. GLMM used the feature extracted for different expressions at the same region to classify them into a zero or non-zero, making it possible for a group to contain different expressions. The global solution of the model was achieved by a function called Maximum Margin Hinge loss. GLMM was later enhanced to Adaptive Group Lasso Regression [74] to assign a continuous value to the distribution of expression present in a non-zero group. GLMM shows its superior performance compares with some existing ML methods from the experiment conducted on s-JAFFE. The work of [34] is also a multilabel approach to FER, Li and Deng [34] introduced a multilabel deep learning model termed Deep Bi-Manifold CNN (DBM-CMM). The model preserves the local affinity of deep emotion features and the manifold structure of emotion labels, while learning the discriminating feature of multilabel expression. The deep network training is jointly supervised by softmax cross-entropy loss with the bi-manifold loss for feature discriminating enhancement. This model learned emotion distribution properly from RAF-ML data and generalised well with existing multilabel data through the incorporated adaptive mechanism.

D. LABEL DISTRIBUTION LEARNING

The extension of the multilabel approach is the Label Distribution Learning (LDL). The main reason that triggers the introduction of the LDL approach to FER is the inconsistencies in FER datasets annotations, which might be due to human annotators' subjectivity, and the subtlety and ambiguous nature of FER data [75]. These challenges adequately justify the need for LDL because LDL could assign multiple labels in different proportions to an expression image. One of the LDL application studies to FER is Emotion Distribution Learning (EDL) [71]. Ying *et al.* [71] resolve the challenge of emotion intensity information loss in the SLL and MLL approach and propose the EDL method to eliminate the threshold constraint. The EDL method describes emotion intensity as a probability distribution of basic emotions present in facial expression, and finally assigns each emotion to the computed degree of intensity. EDL outperform some existing LDL methods and MLL methods when evaluated on s-JAFFE and s-BU-3DFE datasets. In the same manner, [76] proposed two LDL models, which are LDLogitBoost that employs weighted regression tree as the base learner and AOSO-LDLogitBoost that uses vector as base learner. These algorithms are Logistic Boosting Regression (LBR) Based formed from additive weighted function regression. Both LDLogitBoost and AOSO-LDLogitBoost show a promising performance when evaluated on s-BU-3DFE. This method only considers data with distribution scores.

Similarly, [77] proposed an EDL method based on surface Electromyography (sEMG) that uses PCA as feature selection and Jeffery's divergence to find similarities between basic emotions. The sEMG based distribution learning system gains from the robustness of EMG features to head pose variation, the possible influence of external factors, and their unbias information. Nevertheless, EDL is only applicable to datasets with emotion distribution scores.

Most FER databases do not come with distribution scores; applying LDL to these datasets requires methods to recover or relabel the data with distribution scores. The few techniques that consider this challenge include; label enhancement based on fuzzy clustering algorithms [78], which employs C-means clustering to cluster feature vectors and iteratively minimise the objective function to achieve label distribution from logical labels. Another group of Label enhancement is graph-based label enhancement, which includes enhancement algorithm based on label propagation [79] and manifold learning algorithms [80]. The motive behind manifold learning-based label enhancement could achieve label distribution by reconstructing every data point from its neighbour through graphical representation of the topological feature space. In comparison, label propagation-based label enhancement depends solely on iterative propagation techniques to generate label distribution from the logical label. These methods create the distribution labels, but they fail to consider the correlation among the labels. [81] approach distribution label recovery with Graph Laplacian Label Enhancement (GLLE)

TABLE 4. Analysis of MLL, LDL and label enhancement models.

Author	Method	Data	Performance Evaluation	Contribution	Limitation
Zhao et al. [74]	GLMM	ML-JAFFE	Average Precision: 0.9143, Coverage error:2.9381, hamming loss:0.2035, One error:0.1071, ranking loss:0.1466	Model the relationship among FER labels	Model not capture the intensity estimation of the available emotions in FER data.
Ying et al. [71]	EDL + JD	(sJAFFE, sBU-3DFE):	(Kullback leibler:0.0346,0.0402), (Euclidean:0.0957,0.1005), (intersection:0.8998,0.8939), (fidelity:0.9914,0.9898)	present information about emotion intensity in an expression instance	limited to label distribution data.
Xing et al. [76]	LDLogitBoost	s-BU-3DFE	Kullback Leibler:0.0491, Euclidean:0.1263, Fidelity:0.9886, intersection:0.8800	more general entropy model for modeling information distribution in facial images	Not generalised to in-the wild and logical label data, the performance may degrade with large volume data
Xing et al. [76]	AOSO-LDLogitBoost	s-BU3DFE	Kullback Leibler:0.0515, Euclidean:0.1297, Fidelity:0.9874, inersection:0.8764	more general entropy model for modeling information distribution in facial images	Not generalised to in-the-wild and logical label data, the performance may degrade with large volume data
Li and Deng [34]	DBM-CNN	RAF-ML	CLM-Hamming: 0.217, RAkEL-Hamming:0.177, ML-KNN-Hamming:0.168, ML-LOC:0.173, LIFT-Hamming:0.167	Preserves both local affinity and manifold structure of emotion label. Introduction of Adaptation mechanism for data generalisation	computational complexity and resource consumption
Xu et al. [81]	Label enhancement with GLLE (manifold learning)	Bu-3DFE	cheb:1.00, clark:1.13, canb:1.13, cosine:1.00, Interception: 1.07	Label enhancement with consideration given to correlation among labels. Could be applied to data with no distribution label	Not advisable to use on large data size or in-the-wild data. It is Computationally expensive due to implementation of KNN search.
Jia et al. [82]	EDL-LRL + ADMM optimiser	(s-JAFFE, S-BU-3DFE	(cheb: 0.0806,0.0951), (clark:0.3008,0.3556), (cand: 0.6134, 0.7463), (Kullback Leibler:0.0361,0.0694), (cosine: 0.9660,0.9626), (intersection: 0.8970:0.8686)	Preserve correlation among data label locally.	Not generalised to in-the-wild data and data with logical label
Abeere et al. [84]	EDL-LBCNN (CNN + LBC features) KL loss	s-JAFFE	Kullback Leibler:0.0168, CS:0.9842	system performance increases via hybrid convolutional features.	Not generalised to in-the-wild data and data with logical label
Zhang et al. [83]	Cosine similarity + Deep CNN	Oulu-CASIA NIR FER	(Accuracy 81.97% in weak light), (Accuracy: 82.67% in the Dark), (Accuracy in strong light: 84.40%)	the model is immune to illumination variations	not applicable to data with logical label and not generalises to in-the-wild data.
Chen et al. [75]	Auxiliary label (manifold learning) + Deep CNN	posed data(CK+, Oulu-CASIA, CFEE, MMI), wild data (AFFNET, RAF, SFEW)	(Avg. Accuracy: 76.25), (Avg. Accuracy:66.64)	resolve label inconsistency using label enhancement with correlation among labels. Applicable to data without distribution label, not affected by data volume, minimises searching with approximate kNN. Generalises to in-the-wild data and logical data	Time complexity and resource consumption due to auxiliary label space construction.

method. This method successfully generates distribution labels by leveraging topological information of the feature space and adequate consideration of the correlation among labels with appropriate optimisation. GLLE outperforms almost 11 different ML methods, based on the experiments' report on the BU-3DFE dataset as one of the datasets considered. GLLE application to a large dataset and FER data in the wild fails because of its profound assumption of topological space and K- Nearest Neighbour (KNN) search implementation.

Recently, there has been a considerable increase in the quantity and number of FER databases, encouraging the state-of-the-art method, deep learning, for emotion recognition. Using deep networks for distribution learning in FER is evident in [75], [82]–[84]. Jia *et al.* [82] in their quest to preserve the correlation among FER data label locally, they proposed EDL-LRL (Emotion Distribution Label-Low Ranking label correlation Locally), which forms a low-rank structure that alleviates the complexity in emotion correlation, with an assumption that low-rank structure represents the label space. The experiment conducted on label distribution datasets (s-JAFFE and s-BU3DFE) shows the proposed model's prominence. The model considers the correlation among the label locally on data with a distribution label. A generalisation of the method to in-the-wild data and data with a logical label is a challenge. [75] generate an auxiliary label space from two different tasks with intimate correlation with facial expression recognition. The auxiliary tasks employed are facial landmark detection and action unit recognition, which depend on facial structure and movement. This method's motivation is the possibility of two expression images in the auxiliary label space having close expression distribution and consistency in their annotations. This method minimises the problem encountered in GLLE for label enhancement by using approximate KNN for building the approximate KNN (akNN) graphs that generate the auxiliary

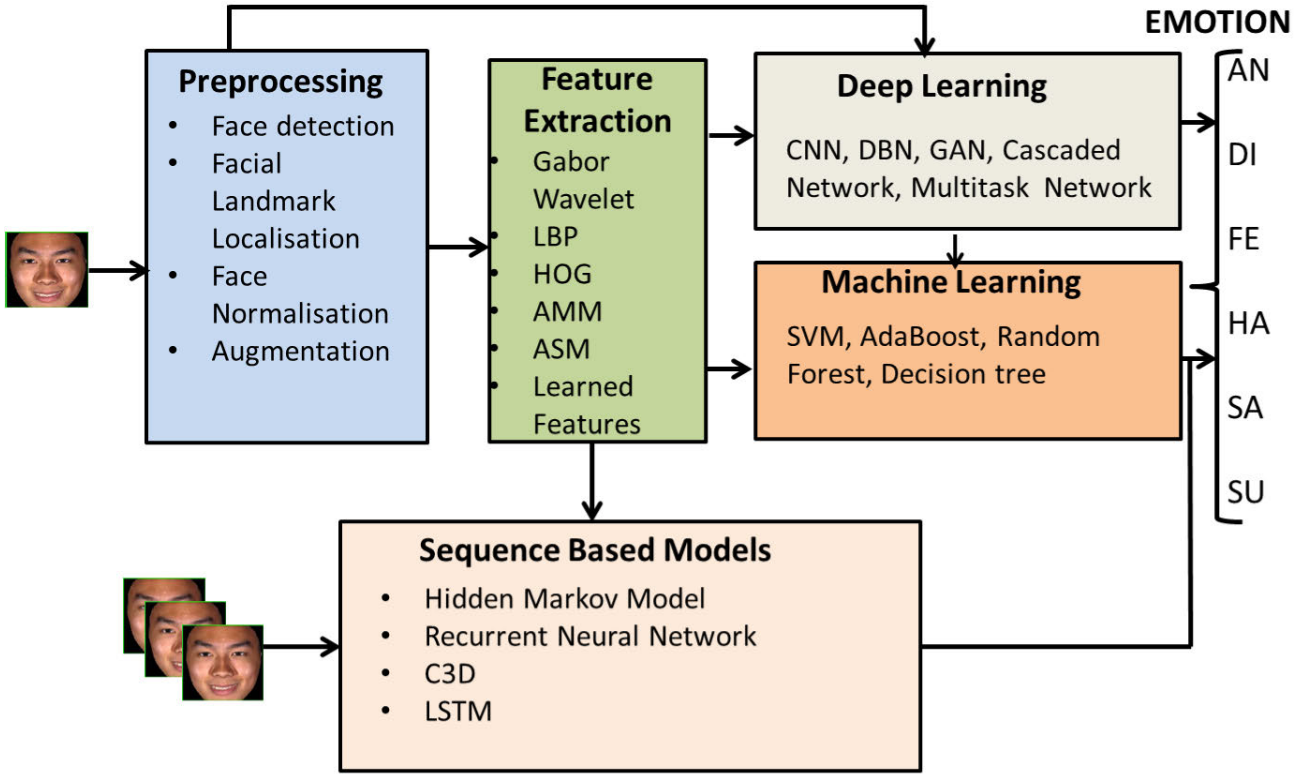

FIGURE 6. Facial expression recognition framework.

labels. Deep CNN was used as the backbone of the proposed system. An experiment conducted on laboratory-controlled data (CK+, Oulu-CASIA, CFEE, MMI) and in-the-wild (AFFNET, RAF, SFEW) proved the system's efficiency over existing methods with an assurance of label consistency and removal of label ambiguity. Zhang *et al.* [83] proposed a Correlated Emotion Label Distribution Learning (CELDL) model for Infrared facial expression recognition. The model initially computes the correlation between expression images using cosine similarities and finally learns the basic emotion in infrared expression with deep CNN. [84] proposed a feature hybrid based model called EDL-LBCNN, which hybridised Local Binary Convolution (LBC) features and Convolution Neural Network (CNN) features train with Kullback-Leibler loss and optimise with ADMM (Alternating Direction Method of Multipliers). The outcome of the experiment on the s-JAFFE dataset shows its promising performance. Figure 5C represents the LDL approach to FER, and Table 4 provides information about the MLL and LDL FER models.

VI. FER ARCHITECTURE

Although FER architecture contains two significant phases, the feature extraction phase and the classification or recognition phase, in most cases, the preprocessing stage is a crucial phase that should not be left out. Automatic FER architecture most time begins with the preprocessing phase. Figure 6 contains major preprocessing algorithms employ in FER. The next phase is feature extraction, where the discriminating features are extracted using feature extraction techniques. The last phase is the classification phase, where each expression image belongs to one of the six basic classes of emotion.

A. PRE-PROCESSING PHASE

Facial feature preprocessing is a vital phase in FER. It assists in preserving relevant features by limiting the infiltration of redundant information during data extraction. It has been observed that data preprocessing has a significant influence on the performance of both conventional machine learning methods and deep learning models. Several algorithms have been proposed in FER, and the list is not limited to face localisation, facial landmark localisation, face normalisation, and data augmentation. We shall elaborate briefly on each of the listed preprocessing methods in the subsections below.

1) FACE LOCALIZATION

Face localisation algorithms help detect the region and the size of a human face in an image or frame of images. It removes the possible background information that may influence the prediction of FER. One of the most popularly used methods for face detection is the algorithm proposed by [85]. They employed Haar-like features and used the AdaBoost classifier to learn a strong classifier from weak

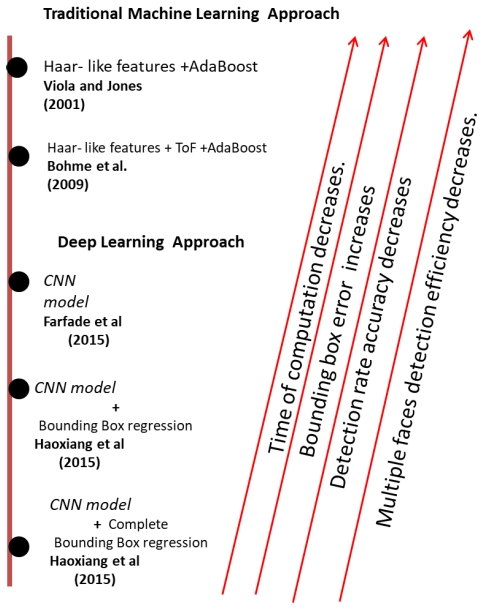

FIGURE 7. Analysis of face detection algorithms.

cascaded classifiers. The algorithm was optimised for speed with integral images.

The study conducted by [86] showed that the model proposed by [85] outperformed LBP-AdaBoost, GF-SVM and GF-NN methods in both speeds of computation and detection accuracy. Despite the excellent detection rate achieved by this algorithm, the training cost is considered expensive. Other identified shortcomings of the Viola and Jones method include non-robustness to partial occlusion and limitation to angular face position. Bohme *et al.* [87] enhanced the Viola and Jones algorithm with range and intensity data from the Time of Flight (ToF) camera. The report showed that the improved method gained a better detection rate at a reduced training time. [88] proposed a CNN model to minimise face detection algorithms' limitation to face angular position, which efficiently detects multiple faces in diverse poses, illumination, and occlusions.

Nevertheless, the method fails to implement bounding box regression. Haoxiang *et al.* [89] worked on the deficiency of [88] model and introduced a cascaded CNN based model that employed bounding box regression. This method failed to fully utilise bounding box regression because it did not evaluate the bounding box for possible reuse. Luo *et al.* [90] fully explore bounding box regression in their CNN model for face detection to determine if the bounding box is fit for a face. They iteratively applied bounding box regression until achieving the appropriate fit and face localisation begins the preprocessing stage of FER architecture. Figure 7 presents an overview of the face detection algorithm discussed.

2) FACIAL LANDMARKS LOCALISATION

Facial landmark localisation (facial alignment) has gained remarkable popularity in Computer Vision and Biometrics. Facial landmarking requires a face detection algorithm before its implementation. The available facial components coordinates (eye-brows, mouth corners, nose ridge, eyes, and lips) in facial landmarks could improve a FER system due to their tendency to minimise in-plane rotation variation. Before the dominance of deep learning methods, most literature employs facial alignment for feature extraction enhancement. Happy *et al.* [91] reconstruct facial patches position in the face using a facial landmark detection model and edge detection algorithm. Happy *et al.* [91] used the method for the extraction of distinctive active patches for expression recognition. [92] before extracting feature patches with HOG, they first located 68 facial landmarks using ensembles of regression trees; some of the points generated formed the patches extracted. A comprehensive study conducted on facial landmark localisation is available in [93], for any interested reader. In recent years, deep learning-based models have been frequently adopted for facial landmark detection. The models have proved their superiority over other models in every facial landmarking detection competition [54], [73], [94]. The authors implemented a combination of cascading CNN modules with specific modifications to the network proposed by [95] that employed three different cascading modules to predict five landmarks. Bodini *et al.* [96] contain more information on deep learning-based methods for face landmark localisation. This paper will consider some literature that applied facial landmark detection at the preprocessing phase in a FER model.

Zhu *et al.* [97] introduced a CovNet model that incorporates the face landmark detection method proposed by [98], which produces 68 fiducial points in the face. The landmark detection aid in the creation of images with eye-brows and mouth locations. The model's performance ascertained the claim that facial landmarks position and shape representation learning could improve expression recognition from images. [99] considered AAM to generate a transformed face region of bidirectional warping facial landmarks for face registration, and that precedes the CNN and Conditional Random Field (CRF) model in solving FER task in a Spatio-temporal environment. [100] employ the supervised descent method to track 49 facial landmarks on facial expression frames in the wild, which could be used by both handcrafted methods and the DNN models for facial expression classification. Many deep learning models supported the prospect of facial alignment in FER, especially in the Spatio-temporal environment.

3) NORMALIZATION

Face normalisation algorithms tend to compliment the effort of face localisation and face alignment. It is expedient to use face normalisation algorithms after facial alignment so that problems that are feature independent (rotation, brightness, background and occlusion) could be minimised. The types of available face normalisation include; geometric, lighting, head rotation (Head Pose), face expression and occlusion. The application of the list depends on the challenges involved. The lighting and the pose normalisation

are necessary for FER in an uncontrolled environment. Variation in the illumination of faces is a significant problem in FER because there is a high tendency for images of a particular subject to differing in brightness and contrast. The lighting normalisation approach minimises intraclass variation that arises from the lighting condition. Li *et al.* [101] use homomorphic filtering normalisation, a photometric normalisation algorithm and histogram equalisation for face preprocessing and claimed that the combination of the two techniques produced effective performance. Shin *et al.* [102] conduct experiments on four different lighting normalisation algorithms, histogram equalisation, isotropic diffusion-based normalisation, DCT-based normalisation, and Difference of Gaussian (DoG). Results showed that the deep network that employs Histogram equalisation at the preprocessing phase has outstanding performance compared to the same network that implements other methods. Bargal *et al.* [103], and Pitaloka *et al.* [104] are among other works that employ histogram equalisation at the preprocessing stage with promising performance. Histogram equalisation normalisation works best when the face foreground and the background are nearly uniform in brightness. Otherwise, local contrast emphasis is possible to occur [105]. Kuo *et al.* [105] proposed combining histogram equalisation and linear mapping to solve the problem of local contrast emphasis. Another hindrance to FER optimal performance in an uncontrolled environment is head pose variation. Pose normalisation has been used severally in the literature to neutralise the pose variation effect.

Most approaches to pose variation correction involve 2D and 3D model fitting that incorporates facial alignment [106], [107]. The motive behind the 2D model fitting for pose variation is that desire pose could be achieved by warping the face with 2D geometrical transformation, the methods that use 2D model fitting techniques capitalised on the working of the facial landmarking method and the warping algorithm.

Sagonas *et al.* [108] generated a frontal face image by applying a Robust Statistical based method. [109] enhanced AMM-based approach for facial landmarks, which, in turn, enhances the fitting process initialisation. The use of the Discriminating Appearance Model (DAM) for pose normalisation is considered in [110]. [111] addressed pose normalisation using Gaussian Process Regression (GPR) and affine transformation. The 3D model fitting achieves normalisation in three procedures; (i) fitting a 3D model on a located facial landmark. (ii) Mapping of face texture to the landmarked 3D model. (iii) Generation of the desired facial pose image from the 3D model texture. 3D model fitting for pose normalisation has been explored diversely in the literature. [112] used a five landmark-based 3D model and quotient image symmetry to develop a lighting aware pose normalisation. [107] introduced a homographic-based pose normalisation technique from the dense grid-based 3D landmark. [113] proposed a method that employed 3D Morphable Model (3DMM) and an interpolation method for frontal view reconstruction. Likewise, [114] synthesised frontal face using 3D Generic Elastic Model (3DDEM) with texture mapping. [115] generate a frontal face from five facial landmarks 3D mesh in a single reference. Deep learning models also explore 2D and 3D model fitting for pose normalisation. The deep learning model was able to synthesis frontal faces from the training of several multi-posed data. [115] used the deep learning method to achieve pose and illumination normalisation, and they trained a deep neural network with face images generated from 3DGEM. [116] introduced the Face Frontalization Generative Adversarial model (FF-GAM) using 3DMM. Model fitting approach for pose normalisation is expensive in terms of time and computational resources. Instead of fixing pose variation with a model fitting method, Obaydy [117] presented a technique that fully utilised facial landmarking and thin-plane spline warping technique for face normalisation, and they were able to efficiently produce a frontal face image from pose variation image in a video. Table 5 contains some normalisation models for FER and their target challenges.

TABLE 5. Presentation of normalisation models application to FER and the target challenges.

Model	Author	Year	Robustness
GPR +Affine transformation	Yang et al. [111]	2010	Pose variation
Histogram Equalization + Hormomorphic filtering	Li et al. [101]	2015	Light variation
DAM (Discrimination Appearance Model)	Gao et al. [110]	2015	Pose variation
Histogramm Equalization	Bargal et al [103]	2016	Light variation
Facial landmark& enhanced AAM	Haghigat et al. [109]	2016	Pose variation
Homomorphic Normalization + 3D landmark	Yao et al. [107]	2016	Pose variation
3DMM + interpolation frontal view	Ferari et al. [113]	2016	Pose variation
3D Mesh facial landmark	Wu et al. [115]	2016	Pose variation
Histogramm Equalization	Pitaloka et al. [104]	2016	Light variation
Satistical based model	Sagonas et al. [108]	2017	Pose variation
FF-GAN	Tran et al. [116]	2017	Pose variation
landmark + 3D + quotient Image symbol	Deng et al. [112]	2017	Pose variation
3D Generic Elastic model	Mendoza et al. [114]	2018	Pose variation
Histogram Equalization + linear mapping	Kuo et al. [105]	2019	Local contrast emphasis
Facial Landmarking + Warpping	Obaydy et al. [117]	2019	Pose variation

4) AUGMENTATION

Data augmentation is a policy adopted in computer vision to improvise for data limitation, a long-time challenge in the field. Data augmentation alleviates data challenges in deep learning through computational manipulations like flipping, cropping, scaling, rotation, and many more. Data augmentation has a significant contribution to machine learning models' performance, especially the deep learning models. Implementation of data augmentation could be done by

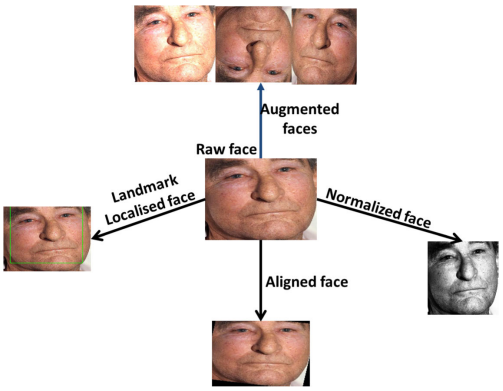

FIGURE 8. The display of some preprocessing algorithms output (augmentation, face localisation, landmark localisation and normalisation) on the raw image.

offline approach or by online approach. The offline method is employed when training data is of few hundreds, while the online approach augments data on the fly. Data augmentation has been widely explored in works of literature [118] and notable in FER [119] where there is a need for large data size. Some works consider the automatic augmentation policy learning approach because of possible biases introduced into the dataset due to the wrong augmentation policy.

Among the existing augmentation policies learning approaches [120]–[124] the work of [125] is the state-of-the-art. [125] introduce AutoAugment using reinforcement learning as a searching technique for augmentation policy with an associated probability. The result guides the system to decide the required policy that is appropriate for the dataset. Cubuk method achieved a significant efficiency, but the reinforcement searching algorithm makes the method computationally expensive. The augmentation policies learning method proposed by [126] is called Population-Based Augmentation (PBA) schedule. This approach generates an augmentation schedule from the Population-Based Training (PBT)- algorithm introduced by [127]. The method is both time and computationally cost-effective compare to the state-of-the-art. In computer vision, a robust augmentation policies learning method is still open research. Figure 8 presents the transformation that occurs after the application of any of the preprocessing algorithms discussed.

B. FEATURE EXTRACTION

As mentioned earlier, the human face is an embodiment of information. A facial image is represented with vast, complex numeric data void of human understanding. The subject information in the image data is termed feature, and extracting useful features from image data correctly to preserve accuracy is called feature extraction. Feature extraction is usually downsized or causes a dimensional reduction in a dataset because it removes redundant attributes from the data to prevent computational complexity, overfitting, and non-generality of the feature model. Every feature extraction technique's main goal is to achieve a feature representation with minimum intraclass variation and maximum interclass variation of high discriminating features. For a FER task, the popular feature extraction techniques include the appearance-based method, geometric-based method, learning features-based method and the hybrid-based method. Each of these methods contains different algorithms for feature descriptors. Both the appearance-based models and geometric based models are classified as handcrafted feature models.

1) HANDCRAFTED FEATURE MODELS

Appearance-based models could describe facial expression features as either a global feature or a local feature. This method extricates changes in the facial image by convolving either the whole image or some region of interest in the image with an image filter or filter bank [128]. Global feature descriptor algorithms translate image features into a single multidimensional feature vector of either colour, shape or texture. While the local feature descriptor algorithm is more concerned about interest points (key points), the number of interest points N forms the N-dimensional feature vectors. The following are the famous appearance-based feature extraction algorithms for FER.

a: GABOR WAVELET

This descriptor was named after the man called Denis Gabor by 1946 [129]. It is a local descriptor. Gabor's image analysis finds a region in the image with a specific frequency in a particular direction; the frequency and orientation description made Gabor appropriate for image texture representation and discrimination. A Gabor filter is a function obtained from amplitude modulation of a sinusoid with Gaussian function in a spatial domain and captures the relevant frequency spectrum. The strength of the Gabor wavelet transform algorithm for feature extraction is its adequate directional selectivity, spatial and frequency maximisation of information, and sensitivity to a slight shift in direction. Equation (1) is the formal definition of the Gabor filter. Assuming the following parameters: (x,y) to be the pixel position in the spatial domain, α to be the wavelength in pixel, θ to be the orientation of the Gabor filter and Sx, Sy to be the standard deviation along the x and y direction; then:

$$G(X, Y) = \frac{1}{2\pi S_x S_y} \exp^{\left[-\frac{1}{2}\left(\frac{x'^2}{S_x^2}\right)+\left(\frac{y'^2}{S_y^2}\right)\right]} \exp^{\left[j\frac{2\pi X'}{\lambda}\right]} \quad (1)$$

where $X' = x\cos\theta + y\sin\theta$ and $Y' = -\sin\theta + y\cos\theta$

Lajevardi [130] argued that the whole Gabor feature extraction method both consume time and yield highly dimensional feature vectors. They considered proposing an average Gabor filter feature method that reduced the feature samples for each facial image from 491520 samples to 12288 samples before downsampling and applying PCA for dimensionality reduction. They achieved this by decreasing 40 feature images of size 128 × 96 pixels each to an average feature image of size 128 × 96 pixels. They used 64 samplings and PCA for dimensionality reduction and applied the K-means classifier on the finally extracted fea-

FIGURE 9. The raw image is placed at the left, follows by the Garbo filter version.

ture. The experiment's result on the JAFFE database showed that the method almost has the same output as the fully Gabor filter method and a gain of minimum time and space consumption. The work of [131] was similar. Still, instead of averaging as in [130], they proposed superimposition of eight images generated from each facial expression image when eight orientation Gabor filters were applied to obtain a single Gabor filter transformation image. Sisodia *et al.* [132] in an approach to minimise computation complexity and dimensionality reduction, selected the best representative number of significant Gabor features to represent each image's expression. Recently, Verma *et al.* [133] follow after the work of [134] however, the significant difference is that [133] employed Gabor filters for feature extraction. They used a Gabor filter bank of five frequencies and eight orientations to convolve each expression image, producing 40 Gabor magnitude images as the required Gabor feature. Harit *et al.* [135] used a Gabor filter to extract features from a normalised face for fiducial points detection. He initially created a Gabor filter bank of six orientations and three spatial frequencies, and later convolved each point in the image with the created filter bank. They reported that 1224 features extracted from each image were generated from 18 magnitudes from 68 fiducial points in the expression image. The classifier used is ANN; the experiment was conducted on two different datasets; JAFFE and Yale. The results showed that the method performed better on the JAFFE database having 81% accuracy, than Yale with 57% accuracy. The Gabor filter transformation of a happy expression image is available in Figure 9.

b: LOCAL BINARY PATTERN (LBP)

Ojala *et al.* [136] proposed LBP as an image texture algorithm suitable for texture analysis. The motive behind the LBP descriptor is that image texture can be represented by the local spatial, with a high tendency to benefit from the grayscale contrast [136]. LBP operates on each of 3×3 pixels of grayscale values of an image and thresholding every neighbour pixel P(0, . . . ,7) with the centre pixel R(1) to generate a binary sequence using a binary threshold function S(x) and then compute the decimal equivalent for the centre pixel with (2).

$$LBP_{R,P} = \sum_{p=0}^{p=1} S(x)2^p \tag{2}$$

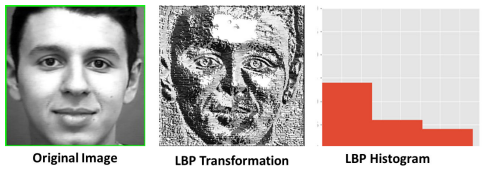

FIGURE 10. The raw image on the left, the LBP version of the image is at the centre, and the histogram of LBP image is on the right side.

Both the LBP equivalent and LBP histogram of a happy image are presented in Figure 10.

The histogram of the LBP encoded region is computed and use as a texture descriptor of that region. The strength of LBP for texture analysis lies in its tolerance for monotonic illumination changes, pose variation and computational simplicity. LBP and its variants have been widely explored in FER. LBP was used as representing feature for FER by [137]–[140]. However, the LBP feature descriptor results in poor performance in the presence of noisy data. This is because it concentrates only on the signs of the difference between the gray values and considers the magnitude relevant texture information as irrelevant. [4], [128] enhanced LBP with a feature selection algorithm. LBP pattern was extracted by dividing a facial image into regions, then the histogram of each region was calculated and later concatenated to form a single face image vector. The feature selection algorithm is further applied, which derived LBP images for all available images and groups all the images into their respective expression classes. Pixel's variance is then computed for each image in the expression class. A threshold called average variance was set to capture high variance code and low variance code. The binary image was formed from the high and low variance code matrix union and became the reference feature selection for LBPs. The experiment conducted on BU-3DFE showed better performance as reported in [4].

Ahmed *et al.* [128] understood the challenges with the original LBP and proposed a Compound Local Binary Pattern (CLBP)- a variant of LBP that uses 2P bits instead of a single P bit employed in LBP with the motives of improving LPB robustness and complement it with other important texture information. The 2P bits captured both the sign differences between the centre and the neighbour gray values and their respective magnitude information. The experiments conducted on the CK and JAFFE datasets using the SVM classifier showed that CLBP outperformed some other feature representation techniques. Another variant of LBP called uniform LBP (uLBP) has also been considered for the FER task. [141] stated that uLBP is a suitable and reliable image descriptor because of its fundamental image texture properties, with its high percentage in texture image that encourages considerable dimensionality reduction without losing texture context significance, and its tendency to ensure statistical robustness by identifying important local texture pattern. [142] extracted feature from the face using uLBP and reduce the high dimensionality of feature data, utilising the firefly and Great-Deluge algorithm to select an optimal representative subset of the extracted feature. The experiment

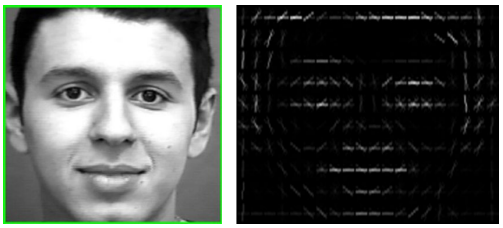

FIGURE 11. The raw image is placed at the left follow by the HOG version of the image.

conducted on the JAFFE dataset using the proposed feature showed that the result produced based on accuracy outperformed the state-of-the-art methods. [141] used Significant Non-Uniform LBP combined with uLBP features to improve the FER recognition rate. He was motivated by the fact that useful micro pattern structural features in facial expression images might be lost if all the non-uniform patterns in the expression image are treated as miscellaneous. He generated features with significant patterns extracted from a non-uniform LBP by considering the transitions from two or more consecutive zeros to two or more consecutive ones, combined with uLBP as FER features.

c: HISTOGRAM OF ORIENTED GRADIENT (HOG)

Dalal *et al.* [143] introduced the histogram of Oriented Gradients (HOG). It is a feature descriptor employed in several fields where objects' characterisation is essential through their shapes and appearance. The histogram of oriented gradients descriptor's motivation is that the distribution of intensity gradients can describe local object appearance and shape within an image and corresponding edge directions [144]. HOG is prominent in object detection as a feature descriptor for image region description. HOG transformation of the happy expression image is shown in Figure 11. HOG starts by dividing an image into blocks and further divides each block into cells. The overlapping blocks made the cell a subcell of many blocks, and then the vertical and horizontal gradient is obtained for each cell's pixel. If $G_y(Y, X)$ is the vertical gradient and $G_x(Y, X)$ is the horizontal gradient, then the magnitude of the gradients are obtained as specified in (3).

$$G(X, Y) = \sqrt{G_x(Y, X)^2 + G_y(Y, X)^2} \tag{3}$$

$$\theta(Y, X) = \arctan \frac{G_y(Y, X)}{G_x(Y, Z)} \tag{4}$$

For each cell, HOG is created. The number of bins with the descriptor is the concatenation of these histograms. Since different images may have different contrast, contrast normalisation is necessary to improve performance. This normalisation results in invariance to changes in illumination and shadowing. Another advantage of HOG is attributed to its operation on local cells, making it invariant to geometric and photometric transformations. HOG was initially utilised for pedestrian detection in static images [143]. [145]–[148] employed HOG as feature descriptor in face detection and recognition. Recently, HOG has been one of the promising feature descriptors for FER. [149] used HOG to encode the deformed components from the detected face and then performed system recognition with linear SVM. The facial parts encoded were the eye-brow and the nose-mouth of the JAFFE database. [92] employed HOG descriptor to describe the representative feature vector for a real-time facial expression system, the feature for each of the patches containing cells concatenated, and the resultant feature vector classified with multiclass SVM. Many other works on FER have also considered HOG mostly as the feature descriptor.

d: PRINCIPAL COMPONENT ANALYSIS (PCA)

Image data is a high dimensional data in which deriving a pattern from it is not easy, but PCA can achieve pattern identification and degree of variabilities in data. PCA uses the dependency between variables of high dimensional data and projects it without losing a significant amount of information into a more tractable lower-dimensional version. PCA tends to find an axis system in data, pointing to maximum covariance in the giving data. The reconstruction of image data results in high dimensionality reduction by using only the significant Eigenfaces responsible for apparent variability. [150] is a thorough survey of FER on PCA. Most of the recent studies in emotion detection used PCA for dimension reduction. PCA is used as a global feature by [3], [151] for expression recognition. [150] in a comprehensive study of facial expression with PCA reported from their research that PCA conducted on facial shape information produced a better result and a better method for FER than the PCA uses facial identities. [152] also enhanced the performance of PCA with Singular Value Decomposition (PCA-SVD) to extract unique features, which provided better performance than both ordinary PCA and LBP + Adaboost. PCA has shown an impressive performance in expression recognition when compared with other Appearance-based features. [153] in their experiment on the JAFFE database, examined the performance of PCA and LDA separately with Euclidean distance as the classifier. Their observation showed that PCA outperformed LDA in terms of recognition rate. Liu *et al.* [154] employed PCA to reduce the hybrid feature dimension of a gray pixel value and extracted LBP from active facial patches of the CK+ database. Then softmax regression classified the dimensionally reduced data space into six basic emotion states under the leave-one-out validation technique.

e: SCALE INVARIANT FOURIER TRANSFORM (SIFT)

SIFT is a detection algorithm introduced by [155], it has four main processing steps; scale-space extrema detection, keypoint localisation, keypoint orientation assignment and keypoint description generation. Scale-space extrema detection deals with keypoint detection, which is achieved by Gaussian (DoG) difference by blurring an image using two different scaling parameters at different octaves of the image Gaussian pyramid. The keypoint obtain at the local extrema

is by comparing a pixel with its eight neighbours, nine pixels of the scale above it, and nine pixels of the scale below it. Keypoint localisation ensures a better keypoint by removing low-contrast keypoint and edge keypoint. It can be referred to as a keypoint refiner. Keypoint orientation assignment goal is to make the keypoint robust or invariant to image rotation. Keypoint orientation is achieved by assigning orientation to keypoint. The orientation is computed from the orientation histogram's peak created from the gradient magnitude and direction, calculated from the surrounding keypoint location neighbourhood. The last stage is Keypoint descriptor generation. At this stage, a neighbourhood of 16×16 blocks around keypoint is divided into a 4×4 size of 16 sub-blocks sub-block creates eight bins orientation histogram that produces a vector of 128 bins value to form the required keypoint descriptor.

Barreti *et al.* [156] extracted SIFT descriptor from the depth of face landmark as a feature for the SVM classifier to addressed person independent problems in 3D expression data. [157] approached emotion recognition from non-frontal facial images by generating super-vectors from the extraction of SIFT features and trained with Edergogic Hidden Markov Model (EHMM). The resultant super-vector was finally classified with Linear Discriminant Analysis (LDA). In [158] keypoints descriptors of SIFT was used as Discriminative SIFT (D-SIFT) features for expression recognition. The investigation conducted by [159] is evidence of the discriminative prowess of SIFT, the result of the experiments on three appearance features; SIFT, LBP and HOG, in a multi-view facial expression analysis showed that SIFT had the best performance. The deep learning approach proposed to solve FER in multi-view images challenge takes a matrix of SIFT features extracted from facial landmarks of images as input feature vector [160]. The model was able to characterise the SIFT feature vectors and their respective high-level semantic information using the corresponding relationship. [161] minimised the small FER data challenge in CNN with dense SIFT feature descriptors and reported that the hybrid of CNN and dense SIFT results in a better performance than using either CNN or CNN with SIFT.

Generally, the strength of appearance-based features lies in capturing transient differences in facial characteristics such as furrows, wrinkles, bulges and many more. However, these features are susceptible to illumination changes and variations in image qualities.

2) GEOMETRIC FEATURE

Geometric features are features extracted statistically from facial landmark displacement. The theory behind this approach is that subsets of face components are more pronounced in facial expression analysis. Geometric feature extraction targets geometric information from facial deformation caused by different kinds of expressions. Geometric based approaches for feature extraction use the Active Shape Model (ASM) or Active Appearance Model (AAM) or their variants to track a dense set of facial points. ASM tends to match groups of model points to an image with a statistical model, and AAM matches an object's shape and texture to an image.

a: ACTIVE APPEARANCE MODEL (AAM)

Cootes [162] introduced the AAM model, which as an extension of the ASM model. AAM successfully forms both the shape and texture of an object. It is categorised as a generative, non-linear and parametric model. AAM has vast applications due to its modelling capability to fix any arising complexity likely to result from high dimensional texture representation. There are three main steps in forming AAM models; (i) connection of shape and texture vectors jointly to each AMM in the training set; (ii) Correlation coefficient matrix computed for the connected shape and the texture vectors in the training set. (iii) Analysis of the correlation coefficient matrix with PCA for each pattern in the training group.

Both ASM and AAM prove to be relevant in facial affective computing, and their application reported severally in literature [163]–[165]. AMM application to FER is frequent in a sequence or Spatio-temporal data. The system proposed in [166] is a real-time system and extracted independent AAM with the aid of the Inverse Compositional Image Alignment (ICIA) method for expression recognition. [167] enhanced the shape produced from the extraction of AAM with second-order minimisation to mitigate large FER errors, to develop FER robust to real-world challenges [168] extracted AAM from edge images rather than gray images, and the report showed that the system is robust against lighting variation. In the pain detection system proposed in [169], AAM was used to decouple shape feature from appearance feature for proper detection of pain through facial expression analysis. [170] used fuzzy logic to monitor the emotion in the shape and texture feature in the facial expression model with AAM. In [171] AAM served as a detector of fiducial point location on facial expression images at the synthesis of feature extraction in the wild. [172] achieved a system that considered ambiguity in the expression displacement for emotion classification by using AAM for face point specification before applying fuzzy C-means for clustering of the emotions. Geometric features are not affected by the lighting condition, and they are not difficult to register and perform well for some Action Units. Nevertheless, they are not suitable to represent an action unit that does not cause landmark displacement.

3) LEARNED-BASED FEATURE

Learned features are attributed to Artificial Neural networks (ANN) and deep learning. Here, ANN learned the direct representative features from the input without feature extraction mathematical models. [51] use visualisation techniques in deep learning to see the kind of feature that Convolution Neural Network (CNN) is using for classification, they observed that the features at the low level resembled low-level Gabor filters. [173] showed that CNN learned features correspond to Facial Action Units (FACs).

TABLE 6. Feature extraction algorithm summary that include strength, limitation and variants. RIFT (Rotational invariant feature transform), GLOH (Gradient location and orientation histogram).

Feature	Strength	Limitation	Variants
LBP	Simple computation with high discriminating power and invariant to grayscale changes	Affected by image rotation and capture limited structural information	Uniform LBP, LBP rotation invariance, Rotated LBP, and Complete LBP
Gabor Filter	maximize spatial and frequency information and possess high sensitivity to small changes in direction	Computationally intensive, and susceptible to high dimensional complexity	Garbor wavelet and log polar Gabor filter
HOG	Capacity to provide global information in large scales and fine-grained details in small scales. invariant to illumination changes and photometric transformation	Extraction takes more time as final descriptor vector grows larger.	Circular HOG, Rectangular HOG
SIFT	invariant to affine rotation and illumination changes	Affected by image rotation, computationally intensive, susceptible to high dimensional complexity	RIFT, PCA-SIFT, GLOH, Gauss-SIFT
AAM	Efficient speed of computation and tendency to reduce high dimensional complexity, it is robust against lighting condition	Not effective for emotion when their is no obvious change in facial landmark displacement, suffers from generalization problem	Locality Constrained AAM, combination with Appearance-based features.
Learned	Efficient descriptor	computaionally intensive, require high computing resources and large volume of data	possible Neural Networks
Hybrid	Feature compliments each other.	Prone to computational complexity	possible combinations of Appearance and/Geometric features

However, the major problem with applying learned-based features to FER is the lack of sufficient data for a network to learn, resulting in overfitting. Another high performing learning-based feature technique is transfer learning, and transfer learning is very efficient in FER where there is limited data for model training. [174]. Apart from CNN based transfer learning, [175] employed an inductive boosting based transfer learning approach to implementing a person-specific model for AUs detection and pain recognition and aimed at achieving generalisation with available minimum data. Learned features have shown promising results, especially in FER, because of its robustness to illumination, rotation, translation, and head pose challenges.

4) HYBRID-BASED FEATURE

Hybrid features give room for the research question of how best to combine features to achieve ultimate performance. [176] proposed an algorithm that fused LBP and HOG features extracted from CK+ and JAFFE database and reduced the extracted features dimension with PCA after permutated the fusion on several classifiers. He found that the fused features on the softmax classifier produced 98.3% on CK+ and 90% on the JAFFE database. The result is evidence that proper hybrid features could significantly improve the system. [177], in their investigation on the best combination of features for optimum performance of the FER system, discovered that the combination of SIFT and geometric features gave better performance compared to either of the features. Also, the experiment showed that LBP and Gabor filter is better in their combination. Table 6 is the concise information of hand-crafted feature extraction algorithms discussed in this work.

C. MACHINE LEARNING MODELS

The feature classification phase ensures the arrangement of features into their respective classes. Classification or regression is achieved chiefly with machine learning classifiers like; Adaboost, SVM, Artificial Neural Network (ANN), deep learning models, or machine learning regression algorithms like Support Vector Regression, Linear Regression and Regression Tree. This section will consider only the popularly used algorithms for classification in FER.

1) SUPPORT VECTOR MACHINE (SVM)

SVM was introduced by [178] as a supervised learning algorithm. It is a binary classifier to find a separating hyperplane of the maximal distance between two trained support vectors.

$$f(x) = W \cdot x + b \tag{5}$$

W in (5) is given as $\mathrm{W} = \sum_i \alpha_i t_i x_i$

SVM kernel was modified and adopted for solving the multiclass problem, Figure 12. shows how multiclass SVM is applied to classified six basic emotions. The application of SVM to multiclass tasks is in two categories; direct and indirect. Direct multi-class SVM is discussed in [179]. [180] used one single optimisation process to distinguish all classes. This approach is possible by designing one objective function for training all K-binary SVMs simultaneously and maximise the margins from each category to the remaining levels. Other multi-class SVM direct approaches include; Simplified Multi-class SVM (SimMSVM) [181], Crammer and Singer's multi-class SVM [182]. The dominant indirect multiclass SVM approach is one versus one and one versus rest. In one Versus one, all possible pairwise classifiers are evaluated and therefore induces K(K-1) individual binary classifier [181]. A new feature is applied to each classifier and categorised them using the classifier with the highest vote. K's separate binary classifiers for K class classification are constructed in the one versus rest SVM multiclass approach. This is possible by first training a classifier using the samples from

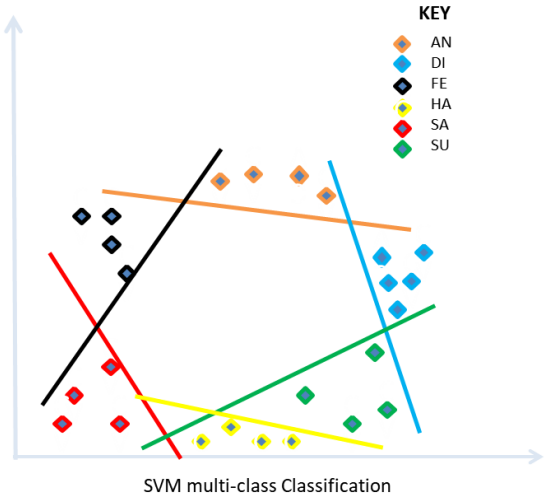

FIGURE 12. Classification of FER with Multi-class SVM.

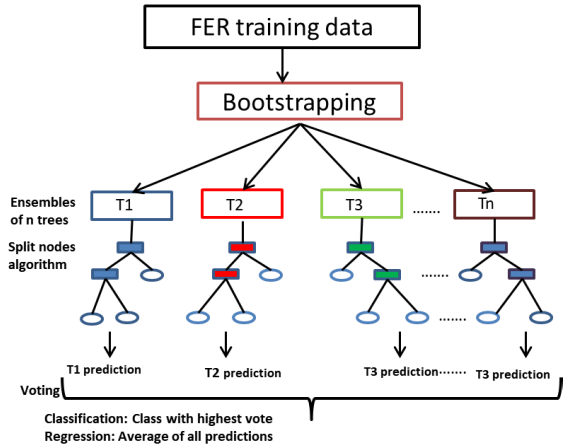

FIGURE 13. Description of how random forest classifies basic emotions.

the class as positive samples and regards others as negative. There is an iteration of the process until all the classes have their classifier. SVM is characterised with high performance in terms of accuracy and data size flexibility, and it has proved to be successful in recognising facial expression, based on its generality, more often when the labels are adequately defined [183]. SVM is mostly employed at the classification phase of FER [184] reported that PCA and SVM give better performance on both JAFFE and MUFE databases to individual performances of LPB and PCA. [175] showed that the system achieved appreciable performance when SVM was used to classified boosted geometric features. SVM has also been employed in micro and macro feature classification [50]. It proved so efficient at recognising Facial expressions in real-time [185], [186].

2) ADAPTIVE BOOSTING (ADABOOST)

Adaboost was coined from Adaptive boosting, a boosting algorithm introduced in 1996 by [187]. It builds a robust classifier from a week classifier that vaguely performed better than random guessing. Adaptive boosting is a development over the existing boosting algorithms, and the word boosting came from adapting the new weak classifier to the mis-classified data by the previous weak classifier [188]. The robust classifier constructed is a linear combination of weak classifiers. The design of AdaBoost was initially for binary classification problems but recently modified and adapted to various multiclass tasks like FER.

Multiclass Adaboost is achievable by boosting a multiclass classifier. For instance, Allwein *et al.* [189], and Benbouzid *et al.* [190] developed Adaboost with Multi-class Hamming loss (Adaboost.MH). [191] implement Adaboost hypothesis Margin (Adaboost.HM) with the aid of ANN. Also, [192] proposed Adaboost with Binary Decision Tree (Adaboost.BDT) for a multiclass task. SAMME (Stagewise Adaptive Modeling Using Multi-class Exponential Loss Function) was proposed in [193], this version resembles the binary version in a combination of a weak classifier, here the combination was a success by using the stage-wisely forward fitting adaptive model for Multi-class Adaboost. In FER, AdaBoost has been employed as a feature selection and as a classifier. [194] combined AdaBoost with the LBP feature to select the most representative feature for FER called AdaboostLBP. [192] approach the multiclass challenges of FER by incorporating ensembles of Binary Tree Adaboost (BTA), the experiment conducted by [195] established that a multiclass Adaboost that followed the adoption of Classification and Regression Tree (CART) performed better than SVM and MLP in terms of accuracy and speed of computation.

A similar classifier like AdaBoost is Random Forest. Random Forest was introduced by [196], it is an ensemble of trees with bootstrapping and bagging implementation. Its efficiency, computation speed, scalability and easy implementation made it a favourite for many classification tasks. The random forest has been employed mostly as a classifier for facial expression features; Figure 13 illustrates the application of random forest to FER. Random Forest is recently used to classify the facial expression feature, selected by Extreme Learning Auto-Encoder (ELAE) from a complete doubled-LBP features [197]. [198] proposed an extension of random forest termed Pair-wise Condition Random Forest (PCRF). The modified Random Forest learned Spatio-temporal pattern from the fiducial points and facial expression frames' appearance features. PCRF shows a significant result comparable with the existing methods. [49] introduce a cascade of forests model which learns in layers for emotion classification. The result shows that the proposed deep forest showed promising results in a wild environment with sparsely distributed and unbalanced data. Table 7 contains some conventional machine learning classifiers models and their various performances on different feature extraction algorithms.

Generally, the conventional machine learning models are binary classifiers (linear), and adapting them to a non-linear and high dimensional feature-based task, like FER, is a great challenge. This is the major limitation to the performance of

TABLE 7. Some conventional machine learning models analysis and performance.

Model	Features	Classifier	Data	Performance
PU et al. [199]	AAM	Random Forest	CK+	96.38%
Muzammil et al [200]	LBP + PCA	SVM	JAFFE MUFE	87% 77%
Kumar et al. [92]	HOG	SVM	CK+	95%
Zhang and Zheng [160]	SIFT	CNN	BU-3DFE Multi-Pie	80.1% 85.2%
Lilana et al. [172]	SIFT +AMM	Fuzzy C-means	CK+	80.71%
Kauser et al. [201]	LBP	ANN	CK	95.83%
Verma et al. [133]	Gabor	ANN	JAFFE	85.7%
Harit et al. [135]	Gabor Gabor	ANN	JAFFE Yale	81% 57%
Elmadhoun et al. [142]	Gabor	SVM	JAFFE	97.6%
Kumar et al. [92]	HOG	SVM	CK+	95%
Ben et al. [202]	LBP	SVM	MMI CK+ JAFFE	73.3% 97.3 86.7%
Ekweariri et al. [4]	LBP+ Feature selection	NA	BU-3DFE	55%
Wang et al. [203]	CNN feature	Random Forest	JAFFE CK+ FER2013 RAFDB	98.9% 99.9% 84.3% 92.3%

traditional machine learning algorithms applied to FER. Also, conventional machine learning models are shallow learners, and their performance depends on the feature extraction models' output. Nevertheless, research still opens to find the appropriate way of incorporating them into the state-of-the-art method.

The table comprises the model evaluation of some traditional machine learning algorithms for FER.

a: DEEP LEARNING MODELS

Deep learning contains some algorithms which are stacked in a hierarchy of increasing complexity and abstraction. Each of the algorithms applies a non-linear transformation to its input and then uses what it learns to create a statistical model as output. This process is iterative until a detectable level of accuracy is reached. The popularly used deep learning Neural Networks in computer vision is the Convolutional Neural Networks (CNN) and the Recursive Neural Networks (RNN). In FER, CNN is used as a supervised classification task, while RNN is used as an unsupervised classification task, especially FER in real-time.

3) CNN

CNN is one of the deep learning algorithms whose concept evolved from the ANN. [204] introduced CNN in 1998. The design of CNN is purposely for image processing and Computer vision. CNN performs an end to end learning, and

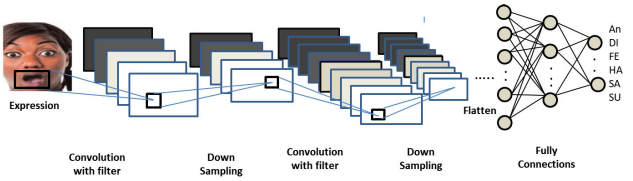

FIGURE 14. A convolution neural network architecture.

the procedure executes in a hierarchy of layers, as shown in Figure 14. Each CNN layer produces representative features ranging from low-level features of the image to a more abstract concept. The process at which CNN automatically learns its representative features emulates the vision mechanism of an animal. That is, the animal visual cortex inspires CNN architectural design. CNN models are self-sufficient in extracting their representative features; there is no need for any pre-calculated features extraction methods. Its high performance contributes immensely to its popularity. The main components of CNN architecture include; convolution layer, pooling layer, dense layer, and fully connected layer.

4) COMMON CNN ARCHITECTURES

There are quite some impressive number of convolution architectures which have contributed immensely to the field of computer vision, few of the networks include LeNet [204], GoogLeNet [205], ResNet [206], ZFNet [207], VGGNet [208] and AlexNet [209]. Most of the listed networks have been used as a deep base network for the training and classifying facial expression images into basic emotion classes. [174], employed GoogLeNet [205] as the deep base network with a different weight learning algorithm called Peak Gradient Suppression (PGS) for backpropagation. The PGS's essence is to strictly bring the feature representation of non-peak expression closer to their corresponding peak expression. CNN networks complexity varies with the increase in the number of the network components or parameters; this came with the belief that the deeper the network, the better the learning of the data's characteristic features, which improves the network's classification power. This capability makes CNN the most relevant tool in both the machine learning and AI world. Many of the networks are useful for the FER task. Most notably in transfer learning, where expression representative features are learned from a pre-trained network, to improvise for insufficient data challenge in FER. Data insufficiency is the major challenge of employing deep learning to FER tasks because; most of the benchmark datasets are just in their few hundreds or unit of thousands. Table 8 presents a summary of CNN deep architectures.

Application of CNN to FER continues to increase favourably with technology evolution and reducing CNN limitations to FER. Many works have been conducted on FER using CNN as a base classifier in different forms. [210] proposed an Action Unit based deep learning network called AU-inspired Deep Network (AUDN). The CNN network has

three phases. The first phase employed the convolution and the pooling operations that learned the representative features called Micro-Action-Pattern. The learned features are to contain information about the local appearance variation. The correlated learned features adaptively combined in the receptive field, which is the second phase. The third phase formed higher-level representations by constructing group-wise sub-networks by applying a multilayer learning process to each receptive field. [211] considered enhancing CNN feature learning capability with some pre-processing procedures so that the network could cope with insufficient data and maximise generalisation capacity. They reported that the system gave an optimal result compare with the state-of-the-art method. [212] proposed an implicit method of ensemble diversity for CNN. They generate different classifiers from a single classifier using parameter variation and fusion of the base classifiers' output. In this case, the classifier considered was CNN. The base classifiers independent CNNs are formed from a random selection of parameters and random selection of CNN architecture, the output generated by each of the base classifiers are fused using the probability-based fusion method. [52] argued that most of the research that automates facial expression considered only strong expressions while weak expressions were left out. The authors presented a CNN network called Deeper Cascaded Peak-piloted Network (DCPN) to join the few. The network design is a version of PPCN by [174], but instead of using GoogLeNet for Network pre-training and fine-tuning, a hybrid of inceptions network called Inception-w, which is a deeper CNN was designed along with a cascaded fine-tuning method used for Pre-training and fine-tuning.

5) RECURRENT NEURAL NETWORK

RNN is a form of Feedforward Neural Networks (FNN) with hidden nodes of memory. The term recurrent emanates from the mechanism of operation of RNN, in the sense that the output of the current input depends on the results obtained from the processing of the previous input(s), as indicated in Figure 15. The hidden nodes make RNN appropriate for many sequence-related tasks like; joined handwriting, voice and speech recognition, Natural Language Processing (NLP) and video processing. Equation (6) is an expression that the current h_t is a function of previous state h_{t-1} and the current input state X_t.

$$h_t = f(h_{t-1}, x_t) \tag{6}$$

Application of activation function to RNN modify (6) to (7)

$$y_t = tanh(Wh_{t-1} + Vx_t) \tag{7}$$

W is the weight of the previous hidden state, V is the weight of the current input state, and tanh is the activation function for non-linearity implementation. The output of RNN is expressed in (7), where y_t is the output state and W is the weight of the output state.

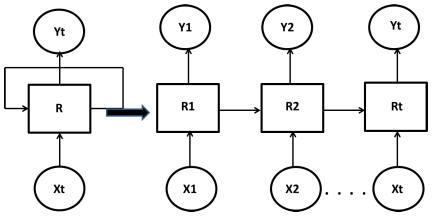

FIGURE 15. A recursive operation of recurrent neural network.

Application of FER to a dynamic or a Spatio-temporal environment is possible with the introduction of RNN in [222]–[224]. Nevertheless, the main challenge with RNN is gradient vanishing and exploding. [225] used IRNN (Identity Recurrent Neural Network) proposed in [226] that incorporated ReLus as activation function and Identity matrix as an initialiser to resolve gradient vanishing problem for learning video level representation and classification model in emotion detection in video. Most of the works that modelled FER in a Spatio-temporal environment used LSTM, a modified RNN that remembers past data in memory and overcame gradient vanishing problems. For instance, [53] proposed a model that used ConvLSTM to learn global features for emotion characterisation from the local features generated by 3D-CNN in a spatiotemporal environment. Likewise, in [227], a nested LSTM (T-LSTM and C-LSTM) generated a multilevel feature model from the collection of Spatio-temporal features produced by 3D-CNN for expression characterisation. T-LSTM is a stack of LSTM units purposely designed for temporal dynamics modelling of facial expression, and C-LSTM used the output of T-LSTM to generate the multilevel target features.

Apart from CNN and RNN, other forms of the deep networks also showed commendable performance in their application to facial affective computation. The groups include; Cascaded Networks, Multitask Networks and Generative Adversarial Network (GAN). In cascaded networks, different modules with different functions are sequentially stacked together in hierarchies of dependency. [228] stacked a module for Local Translation Invariant (LTI) using a Multiscale Contraction Convolution Network (MCCN)stacked with Autoencoder that eventually completes the classification task having distinct emotion features from other latent features such as pose and person identity. Similarly, [229] proposed a cascade of DBN and Autoencoder, whereby expression images were trained with DBN to detect the expression region in the face. The output of DBN becomes the input of Autoencoder for expression classification.

Researchers also engage the capability of the GAN network to propose a robust FER model. The strength of GAN is channelled towards removing variation caused by pose and person identity. [55], [230] develop a pose invariant GAN-based network, while [54], [231] works centred on person identity invariant GAN-based model called IA-GAN (Identity Adaptive-GAN) and PPRL-VGAN (Privacy-Preserving Representation learning-VariationalGAN) respectively. Another group of network types is multitasked networks. The motive

TABLE 8. Summary of popular deep convolution neural networks. FER APP: Application of the model to FER, LP: Model learning parameters.

Network	Author and Year	No of Layers	LP	FER APP.	Contribution	Drawback	Network Variants
LeNet	LeCum et al. [204]	5	51,050	[213] [214]	Form the base model for other deep networks	shallow networks	NA
AlexNet	Krizhevsky et al. [209]	8 (5Conv, 3FC)	60Million	[215]	Implement reLu at the convolution layers which aid the speed of computation and also avoid the propagation of negative value into the network.	loss of detail information due to the size of the kernel used	NA
ResNet-50	He and Zhang [206]	50	26 Millions	[216], [217]	Solve saturation problem incur by network depth increase. It adopts batch normalization and skip connection techniques.	Much time of computation is require.	ResNet-34, ResNet-101, ResNet-18.
GoogLeNet	Szegedy et al. [205]	22(9 inception modules, 4Conv, 5FC, 3 softmax)	7Million	Zhou et al. [52], [174]	The introduction of Inception module leads to great reduction of parameters and implemetation of average pooling subsampling techniques eliminates redundant parameters. It resolve overfitting in deep networks. It minimises computational complexity, and also has fast computational speed.	Inception implementation in GoogLeNet almost make the network of no noticeable limitation.	Inception_V3 [218], Inception_V4 and Inception-ResNet [219].
VGGNet	Simoyan et al. [208]	16 (13Conv, 3FC)	138millions	[220], [221]	Able to capture necessary information available with small kernel size. It encourages deeper learning.	computationally complex in time and space. Highly prone to overfitting expecially with small data size	VGG-19, VGGFace

behind multitask networks is to build a robust FER system by creating a network that could identify features that are not relevant and not related to expression so that the network would be able to concentrate only on the relevant information for expression classification. A method proposed in [31], [98] improved FER performance by extending the FER system to include facial landmark localisation. Another example of a FER multitask learning system is Identity Invariant FER introduced in [232] this makes FER robust against subject identity. The method employs two sub-networks (CNN); one of the networks uses expression sensitive loss to learn discriminating expression features, and the other learns discriminating identity features using identity-sensitive loss. The resultant IACNN is robust against subject identity. The work proposed by [233] is a multitask learning called Multi-signal CNN that introduced FER and face verification for network supervision in the FER development system.

Aside from the demand for a large volume of data for CNN to learn discriminating features for its prediction accuracy, other significant CNN challenges include expensive Hyper-parameter tuning. Selecting an adequate number of layers and the components at each layer depends on skills and experience acquired over time. Also, Gradient Vanishing Problem (GVP) is possible due to constant and consistent decreasing in the gradient at each backpropagation operation through multiple levels of non-linearity.

VII. COMPARATIVE STUDY OF FER METHODS

This section provides comparative information based on performance evaluation of some FER methods, categorised into traditional and deep learning methods. The traditional methods are the category of methods that employed handcrafted techniques for feature representation and used machine learning models for classification [128], [201]. While deep learning methods self-learned the representative feature [234], [235]. The study would be based on the experimental results presented in some literature in the field. Table 9 contains the summary of experiments and results of some of FER's traditional and deep learning methods. The experiment conducted by [128] using Compound Local Binary Pattern (CLBP) features and SVM classifier yielded an average accuracy rate of 90% on CK+ data. While the method of [201] using LBP features and ANN classifier give a better recognition rate of 95% also on CK+ data. The traditional methods accuracy performances are high in a controlled environment and very competitive with deep learning methods performances. However, deep learning models gave a recognition rate higher than the traditional methods. The CNN model proposed by [236] reported an average recognition rate of 98% on CK+ data. However, when [160] enhanced the CNN model with SIFT features, they recorded an accuracy of 99.1%. The deep learning model also shows outstanding performance on the JAFFE dataset with the CNN model proposed by [237], which gave an accuracy of 95.8%. Experiments conducted on FER2013, which is a more challenging FER dataset and large, indicate that the work of [238] performed better. [238] combined SIFT features with CNN model to achieve 75.2% accuracy on FER2013. The recognition rate is higher than any traditional methods or pure deep learning predictions on the FER2013 dataset. Experiment on CK+ as sequence dataset shows that Hidden Markov Model (HMM) provided recognition rate of 98.4% [239], which is a good result, but the deep learning model by [240] termed Expression Intensity Invariant Network (EIINet) showed better result with an accuracy of 99.6%. Deep learning Networks have been considered diversely on in-the-wild data and dynamic data. The experiments conducted on AFEW 7.0, the deep model proposed by [241], which hybridised CNN, RNN and c3D for expression recognition in a dynamic environment, provided the state-of-the-art result of 59.02%.

We cannot but also consider some recent experiments, which are graph-based methods discussed in Section V. The methods tend to recover the emotion distribution from

TABLE 9. Summary of some recent experimental results in FER.

Method	Description	Database	No of Classes	Accuracy	Environment	Category
[128]	CLBP+SVM	CK+	6	90.4%	Static	Traditional
[160]	**CNN-SIFT**	CK+	7	**99.1**%	Static	Traditional
[201]	LPB+ANN	CK+	6	95.83%	Static	Traditional
[234]	CNN	CK+	7	97.10%	Static	Deep learning
[235]	GAN	CK+	7	97.30%	Static	Deep learning
[236]	CNN	CK+	6	98.90%	Static	Deep learning
[128]	CLBP+SVM	JAFFE	6	87.5%	Static	Traditional
[237]	**CNN**	JAFFE	7	**95.80**%	Static	Traditional
[242]	LBP+HOG+PCA+SVM	JAFFE	7	94.42%	Static	Traditional
[135]	Gabor Filter + ANN	JAFFE	6	81%	static	Tradtional
[243]	HOG + SVM	JAFFE	7	76.19	Static	Traditional
[244]	GF + KNN	JAFFE	5	68%	statis	Traditional
[244]	HOG + KNN	JAFFE	5	69%	Static	Traditional
[245]	HOG + SVM	JAFFE	6	95.23%	Static	Traditional
[246]	CNN	FER2013	7	75.10%	Static	Deep Learning
[160]	CNN-SIFT	FER2013	7	73.4%	Static	Traditional
[247]	CNN	FER2013	7	75.2%	Static	Deep learning
[238]	**CNN+SVM**	FER2013	7	**75.42**%	Static	Deep Learning
[248]	C3D	CK+	7	91.44%	Sequence	Deep Learning
[239]	HMM	CK+	7	98.54%	Sequence	Traditional
[68]	CRF	CK+	7	93.90%	Sequence	Traditional
[239]	AAM + HMMRF	CK+	7	93.06%	Sequence	Traditional
[249]	Expression Intensity Invariant Network	CK+	7	97.93	Sequence	Deep Learning
[250]	Network Essemble	CK+	6	97.28	Sequence	Deep Learning
[52]	**EIINet**	CK+	7	**99.60**%	Sequence	Deep Learning
[240]	C3D	CK+	7	97.38%	Sequence	Deep Learning
[251]	VGG16-LSTM	AFEW 6.0	7	44.46%	In-the-wild/Dynamic	Deep Learning
[241]	**CNN-RNN-C3D**	AFEW 7.0	7	**59.02**%	In-the-wild/Dynamic	Deep learning
[252]	Cascaded Network	AFEW 7.0	7	47.40%	In-the-wild/Dynamic	Deep Learning
[253]	C3D	AFEW 7.0	7	48.6%	Dynamic	Deep learning

the logical labels of FER data. The graph base models could be semi-supervised (label propagation) or unsupervised (manifold learning). Although these methods are yet to be wildly explored, the manifold feature proposed in [75] with CNN backend gave an accuracy of 76.25% on static and posed data and 66.64% on in-the-wild data. Likewise, the Deep Bi-Manifold CNN (DBM-CNN) model proposed by [34] gave 96.46% on CK+, which is a competitive result in the field.

The high accuracy recorded for the traditional methods could be attributed to the data size. Traditional methods are very efficient in a static environment and with a small data size. In a more challenging environment with a large data size, traditional methods tend to degrade in performance. Although deep learning methods also give high performance, but perform better when there are enough data for the model to learn the representative feature. The more the data, the better the deep learning performance. The combination of deep learning (CNN) and SVM [238] also produced an encouraging performance. The choice of method for a FER task depends on the available data size, type of data (sequence, static, or dynamic), and computational resources' availability. Nevertheless, Deep learning is state-of-the-art because of its universal performance. Its performance with static [235], [236], sequence [52], in-the-wild data and dynamic data [241] is evident in Table 9. Moreover, its challenges with small data size has been alleviated with some optimisation algorithms like; pretrained networks, transfer learning, and the availability of high computing resources.

VIII. DISCUSSION

FER applications still have no limit; They keep evolving with technology. Emotion Recognition and intensity estimation are the significant areas of FER research focus, just as illustrated in Figure 2. The success in detecting AUs' combination from facial expression contributes to compound emotion recognition from facial expression images.

Research outputs in emotion recognition cannot be overemphasised. Facial emotion recognition challenge is an SLL problem. Here, the research goal is a robust model that could tag a basic emotion to a facial expression image. The early works embraced the traditional methods of combining handcrafted feature models and the conventional machine learning models. These models have been diversely considered in different combinations to achieve an optimal result.

Furthermore, the introduction of deep learning models, and the availability of resources that mitigate its application to FER, encourage more research outputs and successes in the field. Deep Learning is still the trending and the state-of-the-art approach to FER. Many methods have been deployed recently to enhance deep learning performance for FER. They include; Enhancement by using a combination of handcrafted features with deep learning feature [254], enhancement by employing a machine learning classifier like decision tree,

forest tree and SVM at the output layer of deep learning model [255], [256], Network cascading, use of generative networks, application of some optimisation techniques and others. Deep learning model enhancement for FER is still open research in the field. Recently, the SLL approach to FER has been challenged. The challenge considers that facial expression often reveals more than a single emotion at every display. The argument undermines assigning a logical label to facial expression in the SLL approach models. Using logical labels also denied the FER system of assessing the possible intensity information available in an expression image. Likewise, logical labels prevent models in SLL to consider the correlation among labels, label ambiguity and label inconsistency that are inevitably present in the FER datasets.

FER as regards Intensity estimation has been well studied and also gained noticeable attention in the field. Expression intensity estimation began when some sequence datasets respectively captured the intensity of emotion along with the emotion displayed. Virtually all the studies that considered emotion intensity estimation relied either on annotated sequence datasets or Spatio-temporal data. The analogy that emotion rises from face neutral position to the ON-set and continues to the PEAK before eventually dies as OFF-set is the modality used by many researchers to estimate emotion intensity. The approaches employed in the literature include; Distance-based, Cluster-based, Regression-based and Graphical-based. These methods assign numeric value as the intensity estimation of emotion. This process has been discredited [257], [258] because human intuition does not assign numeric value as a measure of emotional intensity. The only reported ordinal intensity estimation is our model [69], we considered FER and intensity estimation as a multilabel learning task and presented a deep multilabel model, which adequately predicts the emotion and its intensity concurrently, using ordinal metrics.

FER definition as a multilabel task addresses the ambiguity problem in the SLL approach to FER. Adopting a multilabel approach will encourage analysis and recognition of both compound and mixture emotions from facial expressions. Nevertheless, multilabel methods fail to provide information about the proportion of the recognised emotions, and also, emotion intensities are not considered. The multilabel approach to FER is still at the early stage in the field.

Modelling FER as an LDL task efficiently and conveniently resolves label ambiguity, label inconsistency, and correlation among labels in FER databases. Direct application of LDL is achieved in emotion distribution learning [76], [77] model, but direct application of LDL to FER is only possible in datasets with distribution labels. Most of the publicly available FER databases contain logical labels. This limitation is further resolved by label enhancement techniques using clustering [78] and graphical-based methods [78], [80], [223]. The label enhancement techniques encourage more LDL models to explore FER with appreciable results.

The MLL and LDL approaches are yet to gain more attention, unlike the SLL approach, which has been studied differently on static datasets, sequence datasets, spatiotemporal or Video data in controlled or uncontrolled environments.

The available databases for FER research are static, sequence, or Spatio-temporal databases collected in controlled or uncontrolled environments. FER's research using the databases provides promising results, but the results degrade in performance in the real world. This challenge leads to the creation of emotion in the wild databases, possibly collected via internet resources and annotated by experts or using some annotated expert software [48], [56]. Another challenge posed to FER is the unavailability of the FER database in large quantities. Deep learning, the state-of-the-art method in the field, needs a large volume of data to learn the deformation in the face caused by the subtle expression for a reliable prediction. Apart from data size, FER databases also need to consider diversity in cultures, races, age, gender, and degree of emotion intensity at collection and annotation. Also, creating FER datasets with consideration given to correlation among labels in data annotation is highly important for developing an efficient FER system.

A. UNRESOLVED FER CHALLENGES

Despite the achievement in FER, FER research still opens up some unresolved issues. There is a need for a FER robust against the long-existing challenges like; non-frontal head poses, light variation in expression images, data morphology and occlusion. Also, a search in the field is required for optimal ways of combining handcrafted features for FER tasks to achieve better performance. Multi-modal affect recognition is of high interest in the field. Multi-modal suggests how to enhance the FER task with some other affective components (Verbal or non-verbal). Data generability is another obvious challenge in the field; there is a need to explore domain adaptation techniques to ensure cross-database generability. FER applications are yet to explore, despite their broad areas of application. Also, identity specificity, which causes an influx of a person's identity information into different classes that leads to wide intraclass variation and small interclass variation, demands attention. FER database creation and annotations that give preferences to the label correlation and inconsistencies need thorough attention too.

IX. CONCLUSION

We have successfully presented a holistic review of FER that covers its possible research trends based on the machine learning approaches. FER as SLL is the most studied aspect, which is still trending in the field. The MLL and LDL approaches are just gaining attention. It suffices to indicate that both SLL and MLL are possible LDL instances; it is just a matter of threshold definition. Our discussion about some popularly employed models ranging from handcrafted feature models, conventional machine learning models to deep learning models identifies deep learning as the state-of-the-art method and discusses its enhancement with traditional

methods. We itemise the unresolved issues in FER together with some future research focus.

REFERENCES

[1] P. Ekman and W. V. Friesen, *Unmaskinh Face a Guide to Recognising Emotion From Facial Clue*. Malor Books, 2003.
[2] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, 1977.
[3] A. Dauda and N. Bhoi, "Facial expression recognition using PCA & distance classifier," *Int. J. Sci. Eng. Res.*, vol. 5, no. 5, pp. 570–573, 2014.
[4] A. N. Ekweariri and K. Yurtkan, "Facial expression recognition using enhanced local binary patterns," in *Proc. 9th Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Girne, Cyprus, Sep. 2017, pp. 43–47.
[5] C. Darwin, *Expression of the Emotions in Man and Animals*, 2nd ed., F. Darwin, Ed. New York, NY, USA: Cambridge Univ. Press 2009.
[6] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personality Social Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.
[7] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.
[8] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," 2019, *arXiv:1901.02884*. [Online]. Available: http://arxiv.org/abs/1901.02884
[9] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, early access, Mar. 17, 2020, doi: 10.1109/TAFFC.2020.2981446.
[10] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: Review and insights," *Proc. Comput. Sci.*, vol. 175, pp. 689–694, Jan. 2020.
[11] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. Wróbel, "Emotion recognition and its applications," in *Proc. Adv. Intell. Syst. Comput.*, vol. 300, 2014, pp. 51–62.
[12] Z. Sheng, L. Zhu-Ying, and D. Wan-Xin, "The model of E-learning based on affective computing," in *Proc. 3rd Int. Conf. Adv. Comput. Theory Eng. (ICACTE)*, vol. 3, Aug. 2010, pp. V3-269–V3-272.
[13] C. L. Lisetti and D. J. Schiano, "Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive science intersect," *Pragmatics Cognition Pragmatics Cognition*, vol. 8, no. 1, pp. 185–235, May 2000.
[14] C. Yang, A. Qi, H. Yu, X. Guan, J. Wang, N. Liu, T. Zhang, H. Li, H. Zhou, J. Zhu, N. Huang, Y. Tang, and Z. Lu, "Different levels of facial expression recognition in patients with first-episode schizophrenia: A functional MRI study," *Gen. Psychiatry*, vol. 31, no. 2, pp. 1–6, 2018.
[15] B. H. Stamm, "Clinical applications of telehealth in mental health care," *Prof. Psychol., Res. Pract.*, vol. 29, no. 6, pp. 536–542, 1998.
[16] S. Poria, A. Mondal, and P. Mukhopadhyay, "Evaluation of the intricacies of emotional facial expression of psychiatric patients using computational models," Tech. Rep., 2015, pp. 1–286.
[17] D. Joachim and I. Song, "Mental health informatics: Current approaches," *Stud. Comput. Intell.*, vol. 491, pp. 247–253, Nov. 2014.
[18] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," 2020, *arXiv:2002.10392*. [Online]. Available: http://arxiv.org/abs/2002.10392
[19] M. A. Butalia, M. Ingle, and P. Kulkarni, "Facial expression recognition for security," *Int. J. Modern Eng. Res.*, vol. 2, no. 4, pp. 1449–1453, 2012.
[20] A. A. A. Al-modwahi, O. Sebetela, L. N. Batleng, B. Parhizkar, and A. H. Lashkari, "Facial expression recognition intelligent security system for real time surveillance," in *Proc. World Congr. Comput. Sci., Comput. Eng., Appl. Comput. (WORLDCOMP)*, 2012, pp. 1–8. [Online]. Available: http://elrond.informatik.tu-freiberg.de/papers/WorldComp2012/CGV2255.pdf
[21] A. M. Barreto, "Application of facial expression studies on the field of marketing," *Emotional Expression, Brain Face*, pp. 163–189, Jun. 2017.
[22] J.-U. Garbas, T. Ruf, M. Unfried, and A. Dieckmann, "Towards robust real-time valence recognition from facial expressions for market research applications," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 570–575.
[23] G. Yolcu, I. Oztel, S. Kazan, C. Oz, and F. Bunyak, "Deep learning-based face analysis system for monitoring customer interest," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 1, pp. 237–248, Jan. 2020, doi: 10.1007/s12652-019-01310-5.
[24] M. Owayjan, A. Kashour, N. Al Haddad, M. Fadel, and G. Al Souki, "The design and development of a lie detection system using facial micro-expressions," in *Proc. 2nd Int. Conf. Adv. Comput. Tools Eng. Appl. (ACTEA)*, Dec. 2012, pp. 33–38.
[25] N. L. Lopez-Duran, K. R. Kuhlman, C. George, and M. Kovacs, "Facial emotion expression recognition by children at familial risk for depression: High-risk boys are oversensitive to sadness," *J. Child Psychol. Psychiatry*, vol. 54, no. 5, pp. 565–574, May 2013.
[26] M. Jeong and B. C. Ko, "Driver's facial expression recognition in real-time for safe driving," *Sensors*, vol. 18, no. 12, p. 4270, Dec. 2018.
[27] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade, "Feature-point tracking by optical flow discriminates subtle differences in facial expression," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Nara, Japan, Apr. 1998, pp. 396–401.
[28] R. Verma, C. Davatzikos, J. Loughead, T. Indersmitten, R. Hu, C. Kohler, R. E. Gur, and R. C. Gur, "Quantification of facial expressions using high-dimensional shape transformations," *J. Neurosci. Methods*, vol. 141, no. 1, pp. 61–73, Jan. 2005.
[29] S. K. A. Kamarol, M. H. Jaward, H. Kälviäinen, J. Parkkinen, and R. Parthiban, "Joint facial expression recognition and intensity estimation based on weighted votes of image sequences," *Pattern Recognit. Lett.*, vol. 92, pp. 25–32, Jun. 2017.
[30] Y.-l. Tian, T. Kanade, and J. F. Cohn, "Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Washington, DC, USA, USA, May 2002, pp. 2–7.
[31] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Nara, Japan, Apr. 1998, pp. 2–7.
[32] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Grenoble, France, Mar. 2000, pp. 46–53.
[33] A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, *Bosphorus Database for 3D Face Analysis* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5372. 2008, pp. 47–56.
[34] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 884–906, Jun. 2019, doi: 10.1007/s11263-018-1131-1.
[35] 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 *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 94–101.
[36] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, 2006, pp. 211–216.
[37] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, Oct. 2014.
[38] G. Anbarjafari, R. Haamer, E. Rusadze, I. Lsi, and S. Escalera, *Review on Emotion Recognition Databases*. Jan. 2017.
[39] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3D dynamic facial expression database," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Shanghai, China, Apr. 2013, pp. 1–5.
[40] 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 Emotion*, vol. 24, no. 8, pp. 1377–1388, Dec. 2010.
[41] D. Erhan, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. H. Lee, and Y. Zhou, "Challenges in representation learning: A report on three machine learning contests," *Neural Netw.*, vol. 64, pp. 59–63, Apr. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608014002159
[42] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE MultiMedia*, vol. 19, no. 3, pp. 34–41, Jul./Sep. 2012.
[43] M. Taini, G. Zhao, S. Z. Li, and M. Pietikainen, "Facial expression recognition from near-infrared video sequences," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[44] A. Mollahosseini, B. Hassani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *CoRR*, vol. abs/1708.03985, Aug. 2017. [Online]. Available: http://arxiv.org/abs/1708.03985

[45] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the MMI facial expression database," Tech. Rep., 2010.

[46] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *CoRR*, vol. abs/1609.06426, May 2016. [Online]. Available: http://arxiv.org/abs/1609.06426

[47] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4DFAB: A large scale 4D database for facial expression analysis and biometric applications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

[48] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5562–5570.

[49] O. Ekundayo and S. Viriri, "Deep forest approach for facial expression recognition," in *Proc. Int. Workshops (PSIVT)*, Sydney, NSW, Australia, 2019, pp. 149–160.

[50] H. Khalifa, B. Babiker, R. Goebel, and I. Cheng, "Facial expression recognition using SVM classification on mic-macro patterns," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 6–10.

[51] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," 2017, *arXiv:1705.01842*. [Online]. Available: http://arxiv.org/abs/1705.01842

[52] Z. Yu, Q. Liu, and G. Liu, "Deeper cascaded peak-piloted network for weak expression recognition," *Vis. Comput.*, vol. 34, no. 12, pp. 1691–1699, Dec. 2018.

[53] D. A. A. Chanti and A. Caplier, "Deep learning for spatio-temporal modeling of dynamic spontaneous emotions," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 363–376, Jun. 2018.

[54] H. Yang, Z. Zhang, and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 294–301.

[55] Y.-H. Lai and S.-H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 265–270.

[56] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2584–2593.

[57] K. K. Lee and Y. Xu, "Real-time estimation of facial expression intensity," in *Proc. IEEE Int. Conf. Robot. Autom.*, Taipei, Taiwan, Sep. 2003, pp. 2567–2572.

[58] C. Quan, Y. Qian, and F. Ren, "Dynamic facial expression recognition based on K-order emotional intensity model," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2014, pp. 1164–1168.

[59] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Intensity rank estimation of facial expressions based on a single image," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Manchester, U.K., Oct. 2013, pp. 3152–3157.

[60] H. Nomiya, S. Sakaue, and T. Hochin, "Recognition and intensity estimation of facial expression using ensemble classifiers," *Int. J. Netw. Distrib. Comput.*, vol. 4, no. 4, pp. 203–211, 2016.

[61] J. Wu and S. Xiao, "Quantitative intensity analysis of facial expressions using HMM and linear regression," in *Proc. 13th ACM SIGGRAPH Int. Conf. Virtual-Reality Continuum Appl. Ind. (VRCAI)*, 2014, pp. 247–250.

[62] R. Zhao, Q. Gan, S. Wang, and Q. Ji, "Facial expression intensity estimation using ordinal information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3466–3474.

[63] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 1, pp. 64–84, Feb. 2009.

[64] M. Kim and V. Pavlovic, "Hidden conditional ordinal random fields for sequence classification," in *Proc. ECML PKDD*, in Lecture Notes in Computer Science, vol. 6322, 2010, pp. 51–65.

[65] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1940–1954, Nov. 2010.

[66] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 1, pp. 28–43, Feb. 2012.

[67] O. Rudovic, V. Pavlovic, and M. Pantic, "Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012.

[68] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015.

[69] O. Ekundayo and S. Viriris, *Facial Expression Recognition and Ordinal Intensity Estimation: A Multilabel Learning Approach*. Cham, Switzerland: Springer, 2020.

[70] R. Plutchik, J. M. Bering, and B. Descriptions, *Contents A Phylogenetic Approach to Religious Origins on the Subcortical Sources of Basic Human Emotions and Emergence of a Unified Mind Science*, vol. 8191. 2001.

[71] Y. Zhou, H. Xue, and X. Geng, "Emotion distribution recognition from facial expressions," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1247–1250.

[72] A. Dhall, R. Goecke, and T. Gedeon, "Automatic group happiness intensity analysis," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 13–26, Jan. 2015.

[73] K. Zhao, H. Zhang, M. Dong, J. Guo, Y. Qi, and Y.-Z. Song, "A multi-labelclassification aprroach for facial expression recognition," in *Proc. Vis. Commun. Image Process.*, Kuching, Malaysia, 2013.

[74] K. Zhao, H. Zhang, and J. Guo, "An adaptive group lasso based multi-label regression approach for facial expression analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 1435–1439.

[75] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13981–13990.

[76] C. Xing, X. Geng, and H. Xue, "Logistic boosting regression for label distribution learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4489–4497.

[77] X. Xi, Y. Zhang, X. Hua, S. M. Miran, Y.-B. Zhao, and Z. Luo, "Facial expression distribution prediction based on surface electromyography," *Expert Syst. Appl.*, vol. 161, Dec. 2020, Art. no. 113683. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417420305078

[78] N. El Gayar, F. Schwenker, and G. Palm, *A Study of the Robustness of KNN Classifiers Trained Using Soft Labels*. Berlin, Germany: Springer-Verlag, 2006, pp. 67–80.

[79] M.-L. Zhang, Q.-W. Zhang, J.-P. Fang, Y.-K. Li, and X. Geng, "Leveraging implicit relative labeling-importance information for effective multi-label learning," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 5, pp. 2057–2070, May 2019.

[80] X. Geng, Q. Wang, and Y. Xia, "Facial age estimation by adaptive label distribution learning," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4465–4470.

[81] N. Xu, Y.-P. Liu, and X. Geng, "Label enhancement for label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1632–1643, Apr. 2021.

[82] X. Jia, X. Zheng, W. Li, C. Zhang, and Z. Li, "Facial emotion distribution learning by exploiting low-rank label correlations locally," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9833–9842.

[83] Z. Zhang, C. Lai, H. Liu, and Y.-F. Li, "Infrared facial expression recognition via Gaussian-based label distribution learning in the dark illumination environment for human emotion detection," *Neurocomputing*, vol. 409, pp. 341–350, Oct. 2020.

[84] A. Almowallad and V. Sanchez, "Human emotion distribution learning from face images using CNN and LBC features," in *Proc. 8th Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2020, pp. 1–6.

[85] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2001, pp. 1–9.

[86] H. Filali, J. Riffi, A. M. Mahraz, and H. Tairi, "Multiple face detection based on machine learning," in *Proc. Int. Conf. Intell. Syst. Comput. Vis. (ISCV)*, Apr. 2018, pp. 1–8.

[87] M. Böhme, M. Haker, K. Riemer, T. Martinetz, and E. Barth, *Face Detection Using a Time-of-Flight Camera* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5742. 2009, pp. 167–176.

[88] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 643–650.
[89] L. Wang and D. Rajan, "A convolutional neural network approach for face identification," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 5325–5334.
[90] D. Luo, G. Wen, D. Li, Y. Hu, and E. Huan, "Deep-learning-based face detection using iterative bounding-box regression," *Multimedia Tools Appl.*, vol. 77, no. 19, pp. 24663–24680, Oct. 2018.
[91] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2015.
[92] P. Kumar, S. L. Happy, and A. Routray, "A real-time robust facial expression recognition system using HOG features," in *Proc. Int. Conf. Comput., Analytics Secur. Trends (CAST)*, Pune, India, Dec. 2016, pp. 289–293.
[93] B. Johnston and P. D. Chazal, "A review of image-based automatic facial landmark identification techniques," *EURASIP J. Image Video Process.*, vol. 2018, no. 1, Dec. 2018.
[94] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *Image Vis. Comput.*, vol. 47, pp. 27–35, Mar. 2016, doi: 10.1016/j.imavis.2015.11.004.
[95] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3476–3483.
[96] M. Bodini, "A review of facial landmark extraction in 2D images and videos using deep learning," *Big Data Cognit. Comput.*, vol. 3, no. 1, p. 14, Feb. 2019.
[97] J. Lu, H. Sibai, and E. Fabry, "Adversarial examples that fool detectors," 2017, *arXiv:1712.02494*. [Online]. Available: http://arxiv.org/abs/1712.02494
[98] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-task learning of facial landmarks and expression," in *Proc. Can. Conf. Comput. Robot Vis.*, Montreal, QC, Canada, May 2014, pp. 98–103.
[99] B. Hasani and M. H. Mahoor, "Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Washington, DC, USA, May 2017, pp. 790–795.
[100] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," *J. Electron. Imag.*, vol. 25, no. 6, Jun. 2016, Art. no. 061407.
[101] J. Li and E. Y. Lam, "Facial expression recognition using deep neural networks," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Macau, China, Sep. 2015, pp. 1–5.
[102] M. Shin, M. Kim, and D.-S. Kwon, "Baseline CNN structure analysis for facial expression recognition," in *Proc. 25th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2016, pp. 724–729.
[103] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 433–436.
[104] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, "Enhancing CNN with preprocessing stage in automatic emotion recognition," *Proc. Comput. Sci.*, vol. 116, pp. 523–529, Jan. 2017, doi: 10.1016/j.procs.2017.10.038.
[105] C.-M. S.-H. K. Lai and M. Sarkis, "A compact deep learning model for robust facial expression recognition," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 2956–2960, 2019.
[106] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 553–560.
[107] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "HoloNet: Towards robust emotion recognition in the wild," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 472–478.
[108] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical frontalization of human and animal faces," *Int. J. Comput. Vis.*, vol. 122, no. 2, pp. 270–291, 2017.
[109] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Fully automatic face normalization and single sample face recognition in unconstrained environments," *Expert Syst. Appl.*, vol. 47, pp. 23–34, Apr. 2016, doi: 10.1016/j.eswa.2015.10.047.
[110] H. Gao, H. K. Ekenel, and R. Stiefelhagen, "Combining view-based pose normalization and feature transform for cross-pose face recognition," in *Proc. Int. Conf. Biometrics (ICB)*, Phuket, Thailand, May 2015, pp. 487–492.
[111] J. Yang, C. Liu, and L. Zhang, "Color space normalization: Enhancing the discriminating power of color spaces for face recognition," *Pattern Recognit.*, vol. 43, no. 4, pp. 1454–1466, Apr. 2010, doi: 10.1016/j.patcog.2009.11.014.
[112] W. Deng, J. Hu, Z. Wu, and J. Guo, "Lighting-aware face frontalization for unconstrained face recognition," *Pattern Recognit.*, vol. 68, pp. 260–271, Aug. 2017.
[113] C. Ferrari, G. Lisanti, S. Berretti, and A. D. Bimbo, "Effective 3D based frontalization for unconstrained face recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Cancun, Mexico, Dec. 2016, pp. 1047–1052.
[114] N. Mendez, A. L. Bouza, L. Chang, and H. Mendez-Vazquez, "Efficient and effective face frontalization for face recognition in the wild," in *Prog. Pattern Recognit., Image Anal., Comput. Vis., Appl. 22nd Iberoamerican Congr. (CIARP)*, vol. 1, Valparaíso, Chile. Cham, Switzerland: Springer, 2018, pp. 534–541.
[115] Z. Wu and W. Deng, "One-shot deep neural network for pose and illumination normalization face recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Seattle, WA, USA, Jul. 2016, p. 6.
[116] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1415–1424.
[117] W. N. I. Al-obaydy and S. A. Suandi, *Automatic Pose Normalization for Open-Set Single-Sample Face Recognition in Video Surveillance*. Springer, 2019.
[118] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019, doi: 10.1186/s40537-019-0197-0.
[119] X. Zhu, Y. Liu, Z. Qin, and J. Li, *Emotion Classification With Data Augmentation Using Generative Adversarial Networks*. Cham, Switzerland: Springer, 2018, pp. 349–360.
[120] T. Tran, T. Pham, G. Carneiro, L. Palmer, and I. Reid, "A Bayesian data augmentation approach for learning deep models," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–10.
[121] A. J. Ratner, H. R. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré, "Learning to compose domain-specific transformations for data augmentation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017.
[122] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.
[123] C. Lin, M. Guo, C. Li, X. Yuan, W. Wu, J. Yan, D. Lin, and W. Ouyang, "Online hyper-parameter learning for auto-augmentation strategy," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6579–6588.
[124] B. Zoph and Q. V. Le, "Neural architecture each with reinforcement learning," *Mach. Learn.*, vol. 2017, pp. 1–16, 2017.
[125] E. D. Cubuk, B. Zoph, D. Man, V. Vasudevan, and Q. V. Le, "Learning augmentation strategies from data," in *Proc. Comput. Vis. Pattern Recognit.*, 2019.
[126] D. Ho, E. Liang, I. Stoica, P. Abbeel, and X. Chen, "Population based augmentation: Efficient learning of augmentation policy schedules," in *Proc. Mach. Learn. Res.*, 2019, pp. 2731–2741.
[127] M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, C. Fernando, and K. Kavukcuoglu, "Population based training of neural networks," 2017, *arXiv:1711.09846*. [Online]. Available: https://arxiv.org/abs/1711.09846
[128] F. Ahmed, H. Bari, and E. Hossain, "Person-independent facial expression recognition based on compound local binary pattern (CLBP)," *Int. Arab J. Inf. Technol.*, vol. 11, no. 2, pp. 195–203, 2014.
[129] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Elect. Eng.*, vol. 93, no. 26, pp. 429–441, Jul. 1946.
[130] S. M. Lajevardi and M. Lech, "Averaged Gabor filter features for facial expression recognition," in *Proc. Digit. Image Comput., Techn. Appl.*, 2008, pp. 71–76.
[131] T. Ahsan, T. Jabid, and U.-P. Chong, "Facial expression recognition using local transitional pattern on Gabor filtered facial images," *IETE Tech. Rev.*, vol. 30, no. 1, pp. 47–52, Sep. 2013.
[132] P. Sisodia, A. Verma, and S. Kansal, "Human facial expression recognition using Gabor filter bank with minimum number of feature vectors," *Int. J. Appl. Inf. Syst.*, vol. 5, no. 9, pp. 9–13, Jul. 2013.
[133] K. Verma and A. Khunteta, "Facial expression recognition using Gabor filter and multi-layer artificial neural network," in *Proc. Int. Conf. Inf., Commun., Instrum. Control*, 2017, vol. 24, no. 9, pp. 1–5.

[134] J. Ou, X.-B. Bai, Y. Pei, L. Ma, and W. Liu, "Automatic facial expression recognition using Gabor filter and expression analysis," in *Proc. 2nd Int. Conf. Comput. Modeling Simulation*, Jan. 2010, pp. 215–218. [Online]. Available: http://ieeexplore.ieee.org/document/5421091/

[135] A. Harit, J. C. Joshi, and K. K. Gupta, "Facial emotions recognition using Gabor transform and facial animation parameters with neural networks," in *Proc. IOP Conf., Mater. Sci. Eng.*, 2018, vol. 331, no. 1, Art. no. 012013.

[136] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[137] X. Feng, M. Pietikäinen, and A. Hadid, "Facial expression recognition based on local binary patterns," *Pattern Recognit. Image Anal.*, vol. 17, no. 4, pp. 592–598, 2007.

[138] B. Tejinkar and S. D. Patil, "Local binary pattern based facial expression recognition using support vector machine," *Int. J. Eng. Sci.*, vol. 7, no. 8, pp. 43–49, 2018.

[139] N. Chitra and G. Nijhawan, "Facial expression recognition using local binary pattern and support vector machine," *Int. J. Innovatice Res. Adv. Eng.*, vol. 3, no. 6, pp. 103–108, 2016. [Online]. Available: http://www.ijirae.com/volumes/Vol3/iss6/17.JNAE10099.pdf

[140] G. Panchal and K. N Pushpalatha, "A local binary pattern based facial expression recognition using K-nearest neighbor (KNN) search," *Int. J. Eng. Res.*, vol. V6, no. 5, pp. 525–530, May 2017.

[141] K. S. Reddy, "A new approach for facial expression recognition using non uniform local binary patterns," vol. 7, no. 3, pp. 20–29, 2018.

[142] A. Elmadhoun, U. Kebangsaan Malaysia, M. J. Nordin, and U. K. Malaysia, "Facial expression recognition using uniform local binary pattern with improved firefly feature selection," *ARO Sci. J. Koya Univ.*, vol. 6, no. 1, pp. 23–32, Apr. 2018.

[143] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, USA, Jun. 2005, pp. 886–893.

[144] M. Dahmane and J. Meunier, "Emotion recognition using dynamic grid-based HoG features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Santa Barbara, CA, USA, Mar. 2011, pp. 884–888.

[145] J. K. J. Julina and T. S. Sharmila, "Facial recognition using histogram of gradients and support vector machines," in *Proc. Int. Conf. Comput., Commun. Signal Process. (ICCCSP)*, Chennai, India, Jan. 2017, pp. 3–7.

[146] X.-Y. Li and Z.-X. Lin, "Face recognition based on HOG and fast PCA algorithm," in *Proc. 4th Euro-China Conf. Intell. Data Anal. Appl.* Cham, Switzerland: Springer, 2018, pp. 10–22. [Online]. Available: http://link.springer.com/10.1007/978-3-319-68527-4

[147] N. Rekha and M. Kurian, "Face detection in real time based on HOG," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 3, no. 4, pp. 1345–1352, 2014.

[148] C. Shu, X. Ding, and C. Fang, "Histogram of the oriented gradient for face recognition," *Tsinghua Sci. Technol.*, vol. 16, no. 2, pp. 216–224, Apr. 2011, doi: 10.1016/S1007-0214(11)70032-3.

[149] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition based on facial components detection and HOG features," in *Proc. Int. Workshops Elect. Comput. Eng. Subfields*, İstanbul, Turkey, 2014, pp. 64–69.

[150] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu, "A principal component analysis of facial expressions," *Vis. Res.*, vol. 41, no. 9, pp. 1179–1208, Apr. 2001.

[151] A. Garg and V. Choudhary, "Facial expression recognition using principal component analysis," *Int. J. Sci. Res. Eng. Technol.*, vol. 1, no. 4, pp. 39–42, 2012.

[152] A. P. Gosavi and S. R. Khot, "Emotion recognition using principal component analysis with singular value decomposition," in *Proc. Int. Conf. Electron. Commun. Syst. (ICECS)*, Coimbatore, India, Feb. 2014.

[153] Taqdir and J. Kaur, "Facial expression recognition with PCA and LDA," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 6996–6998, 2014.

[154] Y. Liu, Y. Cao, Y. Li, M. Liu, R. Song, Y. Wang, Z. Xu, and X. Ma, "Facial expression recognition with PCA and LBP features extracting from active facial patches," in *Proc. IEEE Int. Conf. Real-time Comput. Robot. (RCAR)*, Angkor Wat, Cambodia, Jun. 2016, pp. 368–373.

[155] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Kerkyra, Greece, Sep. 1999.

[156] S. Berretti, A. D. Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A set of selected SIFT features for 3D facial expression recognition," in *Proc. 20th Int. Conf. Pattern Recognit.*, İstanbul, Turkey, Aug. 2010, pp. 1–4.

[157] H. Tang, M. Hasegawa-Johnson, and T. Huang, "Non-frontal view facial expression recognition based on ergodic hidden Markov model supervectors," in *Proc. IEEE Int. Conf. Multimedia Expo*, Singapore, Jul. 2010, pp. 1202–1207.

[158] H. Soyel and H. Demirel, "Facial expression recognition based on discriminative scale invariant feature transform," *IET Digit. Library*, vol. 46, no. 5, pp. 4–5, 2010.

[159] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang, "Multi-view facial expression recognition," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Amsterdam, The Netherlands, Sep. 2008, pp. 3–8.

[160] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Aug. 2016.

[161] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, "Facial expression recognition using a hybrid CNN–SIFT aggregator," in *Proc. Int. Workshop Multi-Disciplinary Trends Artif. Intell.*, vol. 10607. Springer, 2017, pp. 139–149.

[162] T. F. Cootes, G. J. Edwards, and C. J. Taylor, *Active Appearance Models* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 1407. 1998, pp. 484–498.

[163] K.-E. Ko and K.-B. Sim, "Development of a facial emotion recognition method based on combining AAM with DBN," in *Proc. Int. Conf. Cyberworlds*, Singapore, Oct. 2010.

[164] B. Abboud, F. Davoine, and M. Dang, "Facial expression recognition and synthesis based on an appearance model," *Signal Process.-Image Commun.*, vol. 19, no. 8, pp. 723–740, Sep. 2004.

[165] A. S. Dhavalikar and R. K. Kulkarni, "Face detection and facial expression recognition system," in *Proc. Int. Conf. Electron. Commun. Syst. (ICECS)*, Coimbatore, India, Feb. 2014, pp. 1–7.

[166] K.-S. Cho, Y.-G. Kim, and Y.-B. Lee, "Real-time expression recognition system using active appearance model and EFM," in *Proc. Int. Conf. Comput. Intell. Secur.*, Guangzhou, China, Nov. 2006, pp. 747–750.

[167] H.-C. Choi and S.-Y. Oh, "Realtime facial expression recognition using active appearance model and multilayer perceptron," in *Proc. SICE-ICASE Int. Joint Conf.*, Busan, South Korea, 2006, pp. 5924–5927.

[168] C. Martin, U. Werner, and H.-M. Gross, "A real-time facial expression recognition system based on active appearance models using gray images and edge images," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Amsterdam, The Netherlands, Sep. 2008, pp. 1–6.

[169] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face–pain expression recognition using active appearance models," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1788–1796, Nov. 2009, doi: 10.1016/j.imavis.2009.05.007.

[170] Sujono and A. A. S. Gunawan, "Face expression detection on kinect using active appearance model and fuzzy logic," *Proc. Comput. Sci.*, vol. 59, pp. 268–274, Jan. 2015, doi: 10.1016/j.procs.2015.07.558.

[171] F. A. M. da Silva and H. Pedrini, "Geometrical features and active appearance model applied to facial expression recognition," *Int. J. Image Graph.*, vol. 16, no. 4, pp. 1–17, 2016.

[172] D. Y. Liliana, M. R. Widyanto, and T. Basaruddin, "Human emotion recognition based on active appearance model and semi-supervised fuzzy C-means," in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS)*, Malang, Indonesia, Oct. 2016, pp. 439–445.

[173] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2015, pp. 19–27.

[174] X. Zhao, X. Liang, L. Liu, T. Li, and Y. Han, "Peak-piloted deep network for facial expression recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 9906, 2016, pp. 425–442.

[175] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines," *Sensors*, vol. 13, no. 6, pp. 7714–7734, 2013.

[176] X. Wang, C. Jin, W. Liu, M. Hu, L. Xu, and F. Ren, "Feature fusion of HOG and WLD for facial expression recognition," in *Proc. IEEE/SICE Int. Symp. Syst. Integr.*, Kobe, Japan, Dec. 2013, pp. 227–232.

[177] L. Zhang, D. Tjondronegoro, and V. Chandran, "Discovering the best feature extraction and selection algorithms for spontaneous facial expression recognition," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2012, pp. 1027–1032.

[178] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[179] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," in *Proc. Esann*, 1999, pp. 219–224.
[180] V. Vapnik, "The support vector method of function estimation," in *Nonlinear Modeling*, J. A. K. Suykens and J. Vandewalle, Eds. Boston, MA, USA: Springer, 1998, ch. 3, pp. 55–85.
[181] Z. Wang and X. Xue, "Multi-class support vector machine," in *Support Vector Machines Applications*, Y. Ma and G. Guo, Eds. Cham, Switzerland: Springer, 2014, ch. 2, pp. 23–49.
[182] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res.*, vol. 2, no. 2, pp. 265–292, Mar. 2001.
[183] S. W. Chew, S. Lucey, P. Lucey, S. Sridharan, and J. F. Conn, "Improved facial expression recognition via uni-hyperplane classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2554–2561.
[184] M. Abdulrahman and A. Eleyan, "Facial expression recognition using support vector machines," in *Proc. 23nd Signal Process. Commun. Appl. Conf. (SIU)*, Malatya, Turkey, May 2015, pp. 14–17.
[185] L. Chen, C. Zhou, and L. Shen, "Facial expression recognition based on SVM in E-learning," *IERI Proc.*, vol. 2, pp. 781–787, Jan. 2012.
[186] P. Michel and R. E. Kaliouby, "Real time facial expression recognition in video using support vector machines," in *Proc. 5th Int. Conf. Multimodal Interfaces (ICMI)*, 2003, pp. 258–264.
[187] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. ICML*, 1996, pp. 1–15.
[188] H. Fleyeh, R. Biswas, and E. Davami, "Traffic sign detection based on AdaBoost color segmentation and SVM classification," in *Proc. IEEE EuroCon*. Zagreb, Croatia, Jul. 2013, pp. 2005–2010.
[189] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *J. Mach. Learn. Res.*, vol. 1, pp. 113–141, Sep. 2001.
[190] D. Benbouzid, R. Busa-Fekete, N. Casagrande, F.-D. Collin, and B. K. Egl, "MultiBoost: A multi-purpose boosting package," *J. Mach. Learn. Res.*, vol. 13, pp. 549–553, Mar. 2012.
[191] X. Jin, X. Hou, and C.-L. Liu, "Multi-class AdaBoost with hypothesis margin," in *Proc. 20th Int. Conf. Pattern Recognit.*, İstanbul, Turkey, Aug. 2010, pp. 65–68.
[192] H. Fleyeh and E. Davami, "Multiclass AdaBoost based on an ensemble of binary AdaBoosts," *Amer. J. Intell. Syst.*, vol. 3, no. 2, pp. 57–70, 2013.
[193] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class AdaBoost," *Statist. Interface*, vol. 2, no. 3, pp. 349–360, 2009.
[194] S. Prabhakar, J. Sharma, and S. Gupta, "Facial expression recognition in video using AdaBoost and SVM," *Int. J. Comput. Appl.*, vol. 104, no. 2, pp. 1–4, Oct. 2014.
[195] C. S. Fahn, M. H. Wu, and C. Y. Kao, "Real-time facial expression recognition in image sequences using an AdaBoost-based multi-classifier," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2009, pp. 8–17.
[196] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 1–122, 2001.
[197] F. Shen, J. Liu, and P. Wu, "Double complete D-LBP with extreme learning machine auto-encoder and cascade forest for facial expression analysis," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1947–1951.
[198] A. Dapogny, K. Bailly, and S. Dubuisson, "Pairwise conditional random forests for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3783–3791.
[199] X. Pu, K. Fan, X. Chen, L. Ji, and Z. Zhou, "Facial expression recognition from image sequences using twofold random forest classifier," *Neurocomputing*, vol. 168, pp. 1173–1180, Nov. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231215006220
[200] M. Abdulrahman and A. Eleyan, "Facial expression recognition using support vector machines," in *Proc. 23nd Signal Process. Commun. Appl. Conf. (SIU)*, May 2015, pp. 276–279.
[201] N. Kauser and J. Sharma, "Facial expression recognition using LBP template of facial parts and multilayer neural network," in *Proc. Int. Conf. I-SMAC (IoT Social, Mobile, Analytics Cloud) (I-SMAC)*, Feb. 2017, pp. 445–449.
[202] N. Ben, G. Zhenxing, and G. Bingbing, Tech. Rep., Sep. 2021.
[203] Y. Wang, Y. Li, Y. Song, and X. Rong, "Facial expression recognition based on random forest and convolutional neural network," *Information*, vol. 10, no. 12, p. 375, Nov. 2019. [Online]. Available: https://www.mdpi.com/2078-2489/10/12/375
[204] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[205] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–12.
[206] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[207] M. D. Zeiler and R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks," in *Proc. 1st Int. Conf. Learn. Represent. (ICLR)*, 2013, pp. 1–9.
[208] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.
[209] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1–9.
[210] M. Liu, S. Li, S. Shan, and X. Chen, "AU-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, Jul. 2015, doi: 10.1016/j.neucom.2015.02.011.
[211] A. T. Lopes, E. de Aguiar, A. F. D. Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320316301753
[212] G. Wen, Z. Hou, H. Li, D. Li, and J. Xun, "Ensemble of deep neural networks with probability-based fusion for facial expression recognition," *Cognit. Comput.*, vol. 9, no. 5, pp. 597–610, Oct. 2017.
[213] G. Wang and J. Gong, "Facial expression recognition based on improved LeNet-5 CNN," in *Proc. Chin. Control Decis. Conf.*, Jun. 2019, pp. 5655–5660.
[214] Y. Li, X.-Z. Li, and M.-Y. Jiang, "Facial expression recognition with cross-connect LeNet-5 network," *Acta Automatica Sinica*, vol. 44, no. 1, pp. 176–182, Jan. 2018.
[215] X. Chen, X. Yang, M. Wang, and J. Zou, "Convolution neural network for automatic facial expression recognition," in *Proc. Int. Conf. Appl. Syst. Innov. (ICASI)*, May 2017, pp. 814–817.
[216] Y. Chen, J. Du, Q. Liu, and B. Zeng, "Robust expression recognition using ResNet with a biologically-plausible activation function," in *Proc. Pacific-Rim Symp. Image Video Technol.*, Jun. 2018, pp. 426–438.
[217] H. Guo and J. Chen, "Dynamic facial expression recognition based on ResNet and LSTM," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 790, Apr. 2020, Art. no. 012145.
[218] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[219] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
[220] H. Jun, L. Shuai, S. Jinming, L. Yue, W. Jingwei, and J. Peng, "Facial expression recognition based on VGGNet convolutional neural network," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 4146–4151.
[221] A. Fathallah, L. Abdi, and A. Douik, "Facial expression recognition via deep learning," in *Proc. IEEE/ACS 14th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Los Alamitos, CA, USA, Oct. 2017, pp. 745–750, doi: 10.1109/AICCSA.2017.124.
[222] A. Graves, J. Schmidhuber, C. Mayer, M. Wimmer, and B. Radig, "Facial expression recognition with recurrent neural networks," in *Proc. Int. Workshop Cognition Technocial Syst.*, 2008.
[223] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 939–947, Jan. 2019.
[224] H. Kobayashi and F. Hara, "Dynamic recognition of basic facial expressions by discrete-time recurrent neural network," *Nippon Kikai Gakkai Ronbunshu, C Hen, Trans. Jpn. Soc. Mech. Eng. C*, vol. 62, no. 594, pp. 644–651, 1996.
[225] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 467–474.
[226] Q. V. Le, N. Jaitly, and G. E. Hinton, "A simple way to initialize recurrent networks of rectified linear units," 2015, *arXiv:1504.00941*. [Online]. Available: http://arxiv.org/abs/1504.00941

[227] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested LSTM for facial expression recognition," *Neurocomputing*, vol. 317, pp. 50–57, Nov. 2018, doi: 10.1016/j.neucom.2018.07.028.
[228] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 808–822.
[229] Y. Lv, Z. Feng, and C. Xu, "Facial expression recognition via deep learning," in *Proc. Int. Conf. Smart Comput.*, Hong Kong, Nov. 2014, pp. 1–5.
[230] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3359–3368.
[231] J. Chen, J. Konrad, and P. Ishwar, "VGAN-based image representation learning for privacy-preserving facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1570–1579.
[232] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 559–565.
[233] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.
[234] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 20–29.
[235] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2168–2177.
[236] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *Int. J. Comput. Vis.*, vol. 126, no. 5, pp. 550–569, May 2018.
[237] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2015, pp. 1–8.
[238] M. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.
[239] C. Wu, S. Wang, and Q. Ji, "Multi-instance hidden Markov model for facial expression recognition," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, May 2015, pp. 1–6.
[240] S. Kumawat, M. Verma, and S. Raman, "LBVCNN: Local binary volume convolutional neural network for facial expression recognition from image sequences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 207–216.
[241] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Nov. 2016, pp. 445–450.
[242] B. Islam, F. Mahmud, and A. Hossain, "High performance facial expression recognition system using facial region segmentation, fusion of HOG & LBP features and multiclass SVM," in *Proc. 10th Int. Conf. Electr. Comput. Eng. (ICECE)*, Dec. 2018, pp. 42–45.
[243] S. K. Eng, H. Ali, A. Y. Cheah, and Y. F. Chong, "Facial expression recognition in JAFFE and KDEF datasets using histogram of oriented gradients and support vector machine," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 705, Dec. 2019, Art. no. 012031, doi: 10.1088/1757-899x/705/1/012031.
[244] D. B. Vishal, S. C. Devendra, and M. D. Chaudhari, "Use of KNN classifier for emotion recognition based on distance measures," *Int. J. Eng. Adv. Technol.*, vol. 9, 2019.
[245] R. Safa, H. Rafika, and C. S. Ben, "Facial expression recognition system on SVM and HOG techniques," *Int. J. Image Process.*, vol. 15, 2021.
[246] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning social relation traits from face images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3631–3639.
[247] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," Dec. 2016, *arXiv:1612.02903*. [Online]. Available: https://arxiv.org/abs/1612.02903
[248] R. Jack, O. Garrod, H. Yu, R. Caldara, and P. Schyns, "Facial expressions of emotion are not culturally universal," *Proc. Nat. Acad. Sci. USA*, vol. 109, pp. 4–7241, Apr. 2012.
[249] Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim, "Deep generative-contrastive networks for facial expression recognition," Mar. 2017, *arXiv:1703.07140*. [Online]. Available: https://arxiv.org/abs/1703.07140
[250] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognit. Lett.*, vol. 119, pp. 49–61, Mar. 2019.
[251] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, Y. Zong, and N. Sun, "Multi-cue fusion for emotion recognition in the wild," Tech. Rep., Dec. 2016, pp. 458–463.
[252] O. Xi, K. Shigenori, G. H. G. Ester, S. Shengmei, D. Wan, M. Huaiping, and H. Dong-Yan, Tech. Rep., 2017.
[253] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," Tech. Rep., 2017.
[254] S. M. Lajevardi and M. Lech, "Facial expression recognition using neural networks and log-Gabor filters," in *Proc. Digit. Image Comput., Techn. Appl.*, Canberra, ACT, Australia, 2008, pp. 77–83.
[255] D. M. Vo and T. H. Le, "Deep generic features and SVM for facial expression recognition," in *Proc. 3rd Nat. Found. Sci. Technol. Develop. Conf. Inf. Comput. Sci. (NICS)*, Sep. 2016, pp. 80–84.
[256] A. Ravi, "Pre-trained convolutional neural network features for facial expression recognition," 2018, *arXiv:1812.06387*. [Online]. Available: http://arxiv.org/abs/1812.06387
[257] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," Tech. Rep., 2017, pp. 248–255.
[258] B. Martinez and M. F. Valstar, *Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition*. Cham, Switzerland: Springer, 2016.

OLUFISAYO S. EKUNDAYO is from Ondo, Nigeria. He received the B.Sc. degree in mathematical sciences (computer science option) from the University of Agriculture, Abeokuta, Nigeria, in 2008, and the M.Sc. degree in computer science from the University of Ibadan, Nigeria, in 2011. He is currently pursuing the Ph.D. degree in computer science with the University of KwaZulu-Natal, South Africa. He worked as a Computer Science Lecturer at Achievers University, Owo, Nigeria, from 2012 to 2018. He has published more than five articles. His research interests include machine learning, pattern recognition, computer vision, and affective computing.

SERESTINA VIRIRI (Senior Member, IEEE) received the B.Sc. degree in mathematics and computer science and the M.Sc. and Ph.D. degrees in computer science, respectively. He has been in academia, since 1998. He is currently a Full Professor in computer science with the University of KwaZulu-Natal, South Africa. He has published extensively in several accredited journals and international and national conference proceedings. He has supervised to completion several Ph.D. and M.Sc. candidates. He is a rated Researcher by the National Research Foundation (NRF) of South Africa. His main research interests include artificial intelligence, computer vision, image processing, machine learning, medical image analysis, pattern recognition, and other image processing related fields, such as biometrics and nuclear medicine. He serves as a reviewer for several machine learning and computer vision-related journals. He also serves on program committees for numerous international and national conferences.