Multimedia Systems (2023) 29:73–103
https://doi.org/10.1007/s00530-022-00984-w

REGULAR PAPER

A comprehensive review of facial expression recognition techniques

R. Rashmi Adyapady[1] · B. Annappa[1]

Received: 8 December 2021 / Accepted: 9 July 2022 / Published online: 30 July 2022

Abstract

Emotion recognition has opened up many challenges, which lead to various advances in computer vision and artificial intelligence. The rapid development in this field has encouraged the development of an automatic system that could accurately analyze and measure the emotions of human beings via facial expressions. This study mainly focuses on facial expression recognition from visual cues, as visual information is the most prominent channel for social communication. The paper provides a comprehensive review of recent advancements in algorithm development, presents the overall findings performed over the past decades, discusses their advantages and constraints. It explores the transition from the laboratory-controlled environment to challenging real-world (in-the-wild) conditions, focusing on essential issues that require further exploration. Finally, relevant opportunities in this field, challenges, and future directions mentioned in this paper assist the researchers and academicians in designing efficient and robust facial expression recognition systems.

Keywords Facial expression recognition · Emotion recognition · Machine and deep learning · Constrained environment · Unconstrained environment

1 Introduction

Human–computer interaction (HCI) system aims at providing systematic interaction between humans and machines. Charles Darwin [15] has firmly placed facial expressions in an evolutionary context, and has marked the origin of a study on facial expressions. In 1872, he first suggested that, facial expressions revealing basic emotions are universal and his ideas have been a centerpiece for the theory of evolution [34]. Ekman and Friesen in 1972 have proposed display rules as one of the essential aspects for the production of emotional facial expression, and interpretation of these expressions vary across cultures [14, 89]. Face being the most complex signal system is considered to be a highly differentiated part of the human body. The face of identical twins also differs in some aspects [91]. This uniqueness of face is one of the reasons for the widespread application of Facial Expression Recognition (FER). Psychologists have concluded that every part of the face conveys some affective information. According to the literature, verbal channels convey $1/3^{rd}$ of the information, and non-verbal channels carry $2/3^{rd}$ of the data [41].

Facial expression is an efficient way of emotion detection, which facilitates HCI. It is a reflection of the decision; in a social context, it initiates a social exchange or response to others. It is analyzed using Action Units (AU) or directly considers facial emotion inferences from facial expressions [106]. In day-to-day social life, understanding the emotional feelings of others is considered to be a fundamental component and intuition. In human communication, evaluation of facial emotions are an essential factor, which helps in providing evidence about oneself and to know the intentions of the other person [40, 45]. Facial expressions are muscle movements, whereas emotions are underlying mental states which may evoke these expressions. So, there is a difference between expressions and emotions, and they are not identical. Emotions are conscious experiences a person feels, and it involves intense mental process. It is closely related to psychological and physiological arousal signals. In neurobiological terms, emotions are complex action programs that are triggered by internal or external stimuli. Action

Communicated by R. Huang.

R. Rashmi Adyapady
rashmiadyapadyr.177co004@nitk.edu.in

B. Annappa
annappa@ieee.org

[1] Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, India

programs include elements such as facial expressions, action tendencies, bodily symptoms, cognitive evaluations.

Involuntary feeling and expression recognition are challenging, and is difficult for the machines to recognize the emotion in the same way as humans do [30, 60]. Expressions of the same person keep varying with time, intensity, and appearance [40]. Facial expressions vary with age, depend on gender; variations on the face like occlusions, rotation, illumination changes, and accessories degrade the performance of recognition and make it difficult for machines to recognize the expressions efficiently. Also, separating the facial features of one subject in two different interpretations becomes challenging, as they may share the same feature space [57]. There are issues relating to the selection of features, which is also a crucial task to distinguish each facial expression from the rest [127]. Lack of proper datasets is a vital issue to perform recognition accurately [57]. Hence, designing a generalized system for analyzing the expressions, which would overcome all these issues, is challenging. Most of the methods used in past literature are on facial images which contain frontal or nearly frontal view. Recognition of expression from non-frontal facial images becomes challenging [45], and obtaining accurate results needs further exploration.

1.1 Related concepts

This subsection discusses specific terminologies that play a significant role in the research of FER.

1.1.1 Facial action coding system (FACS) and action units (AUs)

The movement of facial components encodes the maneuver of individual facial muscles to represent expression states. It includes Facial Action Coding System (FACS), EMotional Facial Action Coding System (EMFACS), Maximally discriminative facial movement coding system (MAX), and AU space [123]. Ekman and Friesen in 1978 [26] developed the essential and most vital approach to encode each facial behavior, which has been fine-tuned in 2002 [10, 12]. The method is known as the FACS. FACS [12, 103] is mainly used to investigate psychopathology, emotion, pain, and so on. It encodes the changes in facial muscles, in terms of Action Units (AU), which reflect discrete momentary variations in facial appearance. FACS [62, 93] is a standardized system for manually coding the changes in facial muscle; it has a discriminative power to characterize the actions of muscles, in terms of human emotions, by following a set of prescribed rules. It includes a total of forty-four AU's, among which thirty AUs are related to facial muscle contraction, and the remaining fourteen are different actions that are not specified [44].

AU is a terminology used for describing all facial actions [119]. AUs are building blocks for the facial expression [12] and are considered as the smallest visually discriminable facial movements [44]. Coders use the contraction of a single or group of muscles to create AUs. Every muscle action and its representation has a specific meaning, and each activity is assigned a unique number when producing facial expressions [62]. To recognize the emotion classes, individual AU is detected, and based on the combination of AUs; the system further classifies them into a specific category [21, 103]. The combination of a few AUs effectively helps in representing subtle facial changes and a large variety of expression states. However, it is challenging and tedious to accurately detect and track every AU information in images because of minute changes in the facial muscles. It is challenging for psychologists to define every facial expression with the definite prototypical AUs and translate these emotion-related AUs into affective states. Differences in culture, the requirement of high-quality video equipment, high resource intense, arduous manual coding, expensive nature, and also the way of perceiving the facial expression hinders the progress towards fully automated or computer-based facial expression analysis. Expertise takes months together to learn and be professional in AU coding. The EMFACS focuses only on those facial actions which are likely to have emotional significance in it [119]. The MAX system codes discrete emotional states based on a set of formulas obtained via facial movements [87]. It discriminates various facial movements in each facial region and even distinguishes facial movements in three facial areas, forehead and eyebrows, midface (eyes and cheeks), and mouth [63]. AU space-based emotion representation uses continuous coordinates rather than binary values. It is flexible and reduces the misclassification as it is more tolerant towards the AU detection results near marginal areas [125].

With facial AU coding, one can get the knowledge of separating three facial expression categories: Macro, Micro, and Subtle expressions. Macro-expressions occur over individual or multiple regions of the face depending on the expressed expressions, and it can be observed easily in daily interactions [82, 97]. Such expressions occur between 2 to 3 seconds in duration [52] and involve the entire face. Micro-expressions occur when a person is trying to conceal or repress the felt emotional state consciously or unconsciously, and it occurs in a small region of the face [25, 81, 82]. These expressions are very rapid and involuntary and give a brief glimpse of undergoing feelings of a person which he/she is trying to conceal [24, 78]. It lasts between 1/5 to 1/25 second duration in precise length [52]. Subtle changes, in facial appearance, visual differences between human beings, and less number of frames, make it difficult for analysis of micro-expression. Micro-expressions are further categorized into three categories, simulated, neutralized, and masked expressions [5]. Subtle expressions are related to the intensity and depth of the underlying emotions. Such expressions have a low-intensity level and occur when a person starts feeling excitement.

Discussions on neurological aspects for recognizing emotion from facial expressions

A large number of psychological studies have focused on the recognition of emotions from facial expressions over several decades [1]. Muscle actions are the salient features of facial expression. Facial muscles are not only the ones that respond to emotion; even striated muscles in the neck, back, arms, and also smooth muscles of the blood vessels and alimentary tract are also responsible. Annotators can objectively measure the facial expressions without knowing the semantic meaning of emotion expressed [87], by analyzing the position and movement of facial skin and fascia causing wrinkles, lines, folds, and facial landmarks. To describe facial actions, muscles are not directly visible. Thus, FACS is the most intricate instrument used to translate skin movements into muscle patterns. The reduction of the facial expressions into the list of AUs has the advantage, of providing the means of describing any facial configuration even when it does not willingly fit into a preconceived category.

1.1.2 Elicitation of expressions

Expressions can be elicited and collected in multiple ways, such as posed, spontaneous, and in-the-wild. In posed appearance, subjects deliberately display the expressions by reproducing specific deformations in the facial muscle; these expressions are elicited based on the guidance of professionals or actors. Subjects elicit series of expressions based on the demand of instructors; basically, they are aware of being recorded. Free production, ordered production, and portrayal are the three ways of reproducing the posed expressions from the subjects [112]. Spontaneous (authentic) expressions occur naturally and are not controlled by the subjects [65, 94]. It occurs when the subjects try to express internal feelings. Not all individuals express facial expressions in the same way; it depends on one's culture, personal, and familial display rules. Two ways of eliciting such expressions depend on passive and active approaches. In a passive approach, specific emotional states are induced by displaying images or videos, or another way is recording during the interaction of two protagonists to obtain emotion-rich content [112]. In the active approach, capturing of real emotions is done directly, involving the participants themselves.

Data obtained from an unconstrained (real-time) environment includes complex emotions and variations like head pose, occlusions, illumination variation, rotations, and referred to as in-the-wild expressions. It is challenging to recognize facial expressions from these types of datasets. Three modes of getting in-the-wild emotions are crowdsourcing, the corpus of videos, or images obtained using web crawling (both posed and spontaneous expressions) [112]. Spontaneous expressions are distinct from posed ones in terms of spatial patterns, temporal patterns, morphological and dynamic properties [70, 111]. According to Picard et al. [80], five characteristics that need to be considered when eliciting emotions are subject-elicited versus event-elicited, constrained versus unconstrained settings, expressed expression versus feeling, open-recording versus hidden-recording, and emotion-purpose versus other-purpose. The transition from lab controlled environment to collecting naturalistic data in recent years has been a current leveraging and challenging topic. Working with such uncontrolled data recorded in the real-world is of utmost importance as it will be useful to know self intentions.

Discussions on the pros and cons of each elicitation approaches

a. Posed Expressions: This type of expression have different activations of facial muscles and dynamics. People tend to control and move their facial muscles, intentionally [64]. The data collected from this environment will be highly controlled, i.e., frontal exaggerated expressive faces with minimal illumination variations and occlusions [53]. These data cannot reflect real circumstances and it is hard to generalize well in real applications.
b. Spontaneous Expressions: This type of unconscious expression links to the emotional state of an individual. AU's play an essential role in analyzing and describing facial behavior [76]. Developing an intelligent HCI system capable of understanding humans' real expressions is crucial as it will be useful to deploy in real-world applications [64]. The dynamics of facial actions are problematic for FACS coders to simultaneously measure multiple AUs' intensity. Thus, it makes an annotation of such a database a tricky process.
c. In-the-Wild Expressions: Such realistic facial data plays an essential role in advancing research on facial expression analysis systems [18]. This type of expression includes unconstrained facial expressions, varied head poses, occlusions, and illuminations [19]. Compared to lab-controlled datasets, in-the-wild datasets impose a challenge making feature extraction a tedious task due to interference [53].

1.1.3 Facial occlusions and pose variations

Facial occlusions [66] occur due to the presence of obstacles like a scarf, sunglasses, hat, beards, mustaches, the appearance of hand on the face, facial hair, hair covering the frontal face. The presence of occlusions blocks the facial regions and increases the difficulty in extracting

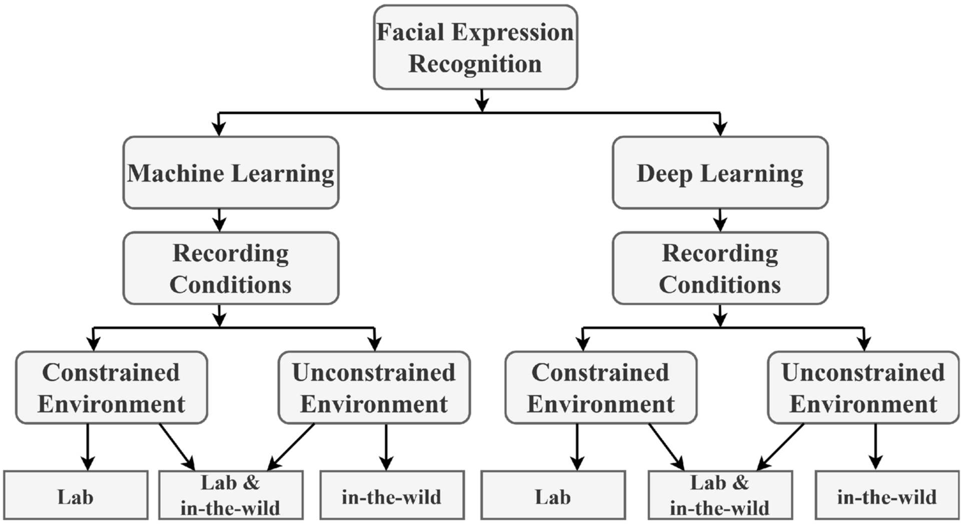


Fig. 1 The taxonomy of literature review representation

discriminative facial features, resulting in facial registration error and inaccurate face alignment [23]. These factors degrade the performance of FER systems. Facial occlusions lead to high intra-expression variations due to noise and outliers [123]. And, considering these noises and outliers can be the indicators of emotions. Head pose mainly relies on the face detection process. Yaw, roll, and pitch are the rotation angles used for estimating the orientation concerning a head-centered frame [3, 48, 67]. Pose variation is prone to errors and degrades the performance of the FER system. Robust estimation of the head pose leads to pose-invariant face recognition. Facial pose variation is one of the difficult tasks to be tackled [120], and such images require some transformations like initialization and normalization for analysis. Self-occlusion is also a significant problem that occurs due to the rigid rotation of the head and includes information loss. Hence, understanding facial occlusions and pose variations is also a key to in-the-wild FER.

Contributions through this work:

- To present a few relevant concepts related to FER and help researchers understand the basic and essential components of trends in this field.
- To provide a comprehensive review of traditional and advanced FER approaches, with special emphasis on the essential methodologies, used to solve FER in both constraint and unconstrained environments. Indispensable limitations and possible future directions presented in this paper help researchers find the necessary gaps in this area.

The structure of this review paper is as follows. Section 2 presents an overview of the works done in the past literature, and Sect. 3 delves into challenges and potential future directions. Applications of FER are explored in Sect. 4. Finally, concluding remarks are presented in Sect. 5.

2 Literature review

This section presents the literature on leading traditional and deep learning techniques for FER. As shown in Fig. 1, the literature review is further subdivided based on recording conditions, i.e., whether the dataset is recorded in controlled or/and uncontrolled environment situations. If the dataset is recorded in a constrained environment, it is classified under the lab recording condition. And, if the dataset is recorded in an unconstrained environment, the recording condition is In-the-wild, termed as IW in the following subsections. If the authors have utilized the datasets recorded in both constrained and unconstrained environments, the recording conditions are filled with the terms lab and IW.

2.1 Traditional machine learning (ML) approaches used in constrained and unconstrained environments

Handcrafted features are used to classify the expressions into respective emotion classes, and Table 1 highlights few works of literature based on traditional machine learning techniques. To recognize the expressions collected in a constrained environment, Ghimire et al. [31] has concatenated geometric feature descriptor Normalized Central Moments (NCM) with LBP and fed it to SVM for classification. The results show that feature descriptors extracted from domain-specific local regions outperform holistic representations. Likewise, Zhong et al. [127] has utilized LBP and its variant uniform LBP to extract features and has explored general and specific information about different facial expressions

Table 1 Summary of literature based on traditional machine learning techniques

RC	References	Dataset	Type of data	Approaches	Remarks
Lab					
	Ghimire et al. [31]	CK+	S	LBP, Normalized Central Moments (NCM), SVM	Incremental search approach has been used to determine important local regions, which has reduced the dimensions of feature and improve recognition accuracy. FER from local region selection has reduced the computation complexity of the algorithm. The performance of the proposed system has decreased when neutral emotion class has been included for evaluation as it is confused with anger and sadness emotion classes. Performance could be enhanced to discriminate facial expressions by searching and selecting the best features within the framework.
	Zhong et al. [127]	CK, MMI & GEMEP-FERA	D	Uniform LBP, SVM	The proposed framework improved recognition accuracy by combining common and specific facial patches at different scales. It provides more accurate appearance locations. The recognition rate of anger emotion class has not been as good as when compared to other emotion classes. Head pose variation and facial occlusions have not been considered in this work.
	Liong et al. [52]	CAS(ME)^2, CASME II, SMIC-HS, SMIC-NIR and SMIC-VIS	S	LBP-TOP, Bi-Weighted Oriented Optical Flow (Bi-WOOF), SVM	Reduces computational complexity as well as cost by selecting only apex frames instead of the entire video sequence. Justification is needed to understand the extent, these apex frames influence the performance of recognition.
	Niu et al. [75]	CK+, JAFFE, MMI	S & D	LBP, ORB, SVM	Overcame excessive hardware specification requirement issues of deep learning models. Improved ORB solved the problem of redundancy and feature point overlap in extraction process. On the CK+ database, the proposed solution had a low recognition rate. Requires further research to improve accuracy and computation speed.
	Boughida et al. [6]	JAFFE, CK, CK+	S	Gabor filter, PCA, Genetic Algorithm	For datasets with multiple features, the proposed technique slows down genetic algorithm convergence. Gabor filter parameters are manually selected to fetch the best features, but this does not guarantee the optimal values, necessitating further investigation.

Table 1 (continued)

RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab					
	Guo et al. [32]	SMIC (HS), CASME II, SAMM	D	Extended Local Binary Patterns on Three Orthogonal Planes (ELBPTOP)	Introduction of Whitened Principal Component Analysis (WPCA) to micro-expression recognition obtained more compact and discriminative feature representations, thus achieving computational savings. The proposed approach preserves gray scale invariance.
	Rashmi and Annappa [84]	CASME II, SAMM, and SMIC (HS)	S	Delaunay Triangulation (DT), Voronoi Diagram (VD); Stacking classifier, majority voting	DT and VD technique is utilized to segment facial ROI according to AUs. The combination of geometric and texture features extracted from the DT and VD complemented each other to recognize micro-expressions efficiently. Using advanced algorithms to detect better facial landmark points and extract more significant data make the suggested model realistic for real-world applications.

Table 1 (continued)

RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab & IW					
	Cruz et al. [13]	CK, MMI, AVEC 2011 & 2012	D	LBP, SVM	The proposed method reduces memory cost by downsampling the number of frames in video samples. The frameshave been segmented in an evenly-sized manner and may cause a boundary effect if the unlabeled apex is spotted near to the segmentation boundary.
	Chen et al. [10]	CK+, GEMEP-FERA2011, AFEW 4.0	D	Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP), geometric warp feature, multiple kernel SVM	HOG-TOP is compact and efficient in characterizing dynamic changes. The geometric warp feature has been useful in capturing facial configuration changes. The performance of in-the-wild datasets is lower compared to a lab-controlled environment, and requires further exploration.
	Pham et al. [79]	FER 2013	S	Multi-layer Perceptron (MLP), GoogLeNet, DenseNet, VGG-Face	The focus of this work is to check whether the FER result is reliable or not. The authors evaluate whether further information is needed to make a reclassification to improve FER performance based on the reliability measure. Considering uncertainty with MLP and emotion-preserving image retrieval with AUs increased the network's performance. The proposed technique can be combined with any Deep Learning architecture to boost the system's performance even more.

[*] *RC* Recording Conditions, *IW* In-the-Wild, *S* Static, *D* Dynamic

using a two-stage Multi-task Sparse Learning (MTSL). Similarly, Niu et al. [75] used LBP and improvised oriented FAST and rotated BRIEF (ORB) to extract discriminative features and SVM for classification. The improvised ORB solved the problem of conventional ORB by using region-wise division for feature point extraction. [52] has utilized LBP from Three Orthogonal Planes (LBP-TOP) features and Bi-Weighted Oriented Optical Flow (Bi-WOOF) feature extractor to encode crucial expression present at the apex frame of video sequences. Liong et al. [6] utilized Gabor filters to extract features from Region of Interest (ROI); Principal Component Analysis (PCA) was used to choose the best feature, and a genetic algorithm was employed to optimize SVM hyperparameters for FER.

To detect minute changes in the facial region, Guo et al. [32] proposed Extended Local Binary Patterns on Three Orthogonal Planes (ELBPTOP). The authors explored the second-order discriminative information in two directions of a local patch, and one is the radial differences (RDLBPTOP), and the other one is the angular differences (ADLBPTOP), as a complement to the differences between a pixel and its neighbors (LBPTOP). Similarly, Rashmi and Annappa [84] developed an ensemble technique that employs the Delaunay Triangulation (DT) and Voronoi Diagram (VD) approach to take a combination of geometric and texture features. The ensembled features are fed to the stacking classifier and majority voting approach, which takes a variety of machine learning classifiers for further classification of micro-expressions. The discriminative features extracted from the AU regions helped in identifying micro-expressions efficiently.

Similarly, in an unconstrained environment, Cruz et al. [13] has utilized LBP and its variant uniform LBP to extract features, and Chen et al. [10] has used both visual and audio modalities, and obtained features using a Histogram of Oriented Gradients from TOP (HOG-TOP). Pham et al. [79] utilized Multi-layer Perceptron (MLP) as a classifier to determine whether the current classification results are reliable or not. In case of unreliability, the facial image is used to search for similar images. Images with identical facial expressions are processed using AUs from a vast set of unlabeled face datasets.

2.2 Deep learning (DL) approaches used in constrained and unconstrained environments

Deep learning features have proved to be efficient in extracting crucial patterns from images and have better discriminative power to classify into respective emotion classes as compared to handcrafted features. Table 2 highlights a few notable works of literature performed using deep learning techniques. In a constraint environment, Zhang et al. [122] has extracted Scale-Invariant Feature Transform (SIFT) and has fed these features to Deep Neural Network (DNN) to learn discriminative patterns. A combination of CNN and image processing techniques have been utilized by Lopes et al. [57] to extract expression specific features that have proved to be efficient for FER. Whereas, Barros et al. [4] has used convolution units of CNN to identify the location of expression in a cluttered scene rather than using it for classification purposes. Likewise, Kim et al. [40] has utilized CNN to extract spatial features and trained LSTM, thereby generating discriminative spatio-temporal representation to improve FER with varying expression intensities. Similarly, Pan et al. [77] has aggregated spatial and temporal features using the aggregation layer, thus filling the gap between visual features and emotions. Zhang et al. [121] has utilized Part-based Hierarchical Bidirectional Recurrent Neural Network (PHRNN) and Multi-signal CNN (MSCNN) to extract dynamic and morphological variations of facial expressions in video sequences.

To efficiently recognize expressions from facial movements and body gestures, Sun et al. [98] has utilized CNN, Bilateral Long Short-Term Memory Recurrent Neural Networks (BLSTM-RNN) and Principal Component Analysis (PCA). Whereas, Liang et al. [50] proposed a BiLSTM architecture and fused both spatial and temporal dynamics jointly for FER. Also, Sun et al. [99] extracted spatial features from the gray-level image and optical flow features extracted from X and Y components of the emotional and neutral face image. The spatial-temporal features extracted from three-channel elements are fused using Multi-channel Deep Spatial-Temporal feature Fusion neural Network (MDSTFN) for FER. Xie and Hu [115] proposed Deep Comprehensive Multi-patches Aggregation CNN (DCMA-CNN) with Expressional Transformation-Invariant (ETI) pooling to distinguish expression sensitive elements and reduce the negative impact. Ma et al. [60] has conducted cross-model noise modeling, to eliminate data pollution in audio data and data redundancy in visual data using 2D CNN and 3D CNN, respectively. The Fusion, of uniform LBP and geometric features using autoencoders, has helped Majumder et al. [61] in representing the non-linear data in lower dimensions. Tang et al. [102] has proved that automatically extracted features are superior when compared to handcrafted features. On the other hand, Zhi et al. [126] used an evolutionary DL approach to evaluate AUs, and it proved to be efficient compared to other AU detection algorithms.

In an unconstrained environment, few DL works have proved to be efficient in solving variations of in-the-wild databases, cross-cultural problems, vast computational complexity, overfitting, and small data sample issues. Georgescu et al. [29] has used architectures of CNN and Bag-of-Visual-Words (BOVW) handcrafted features to recognize facial expressions and has employed a local and global learning approach using SVM. Amongst them, local learning SVM proved to be efficient for predicting

Table 2 Summary of literature based on deep learning techniques

RC	Reference	Dataset	Type of data	Approaches	Remarks
Lab					
	Zhang et al. [122]	BU-3DFE, Multi-PIE	S	SIFT, DNN	Deals with multi-view FER. The use of landmark points alleviates the misalignment problem. Overcomes the problem of overfitting and reduces the model complexity. Accuracy of proposed work at a 90 *deg* pose angle is less when compared to [37] on the Multi-PIE database. Accuracy is low with squint expression class on Multi-PIE and fear expression class on BU-3DFE datasets.
	Majumder et al. [61]	MMI, CK+	D	Facial keypoints (Geometric features), Uniform-LBP, autoencoders, Kohonen Self-Organizing Map (SOM), SVM	The performance is computationally efficient and accurate. The fusion of geometric and appearance features using autoencoders has provided the best representation of facial attributes to recognize facial expression. The soft thresholding technique used at the SOM classifier's output nodes reduces the problem of misleading class prediction. LBP feature extraction applied to four facial key regions reduces the redundant information. Focus is on high-level semantic concepts of expression, and the proposed method has ignored fine-grained information at local facial regions. The performance is not been investigated in the real-time environment settings. It requires the first frame of sequence to be of neutral expression, which is not always possible to get in real-life databases.
	Lopes et al. [57]	CK+, JAFFE, BU-3DFE	S	CNN	Alleviates the data shortage problem. The model has utilized less time for training the system. The proposed work is suitable to operate in real-time environment settings. The accuracy is poor for sad expression class as when compared to other emotion classes. The robustness of the model with various head poses has not been verified.

Table 2 (continued)

RC	Reference	Dataset	Type of data	Approaches	Remarks
	Zhang et al. [121]	CK+, Oulu-CASIA, MMI	D	Part-based Hierarchical bidirectional Recurrent Neural Network (PHRNN) & Multi-signal Convolutional Neural Network (MSCNN)	The proposed method increases variations across distinct expressions and reduces the differentiation of within-class expressions. It lacks in capturing information of expressions when there is a small motion around the critical areas of the face.
	Kim et al. [40]	MMI, CASME II	D	CNN and LSTM	The proposed approach overcomes the problem of variations in expression intensity and duration of expression. Utilization of expression-state information (onset, apex, offset) improves FER performance and the recognition rate of microexpressions. Layers of LSTM need fine-tuning using real-world datasets to improve the recognition rate further.
RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab					
	Barros et al. [4]	FABO	D	Attention Cross-channel Convolution Neural Networks (CCCNN)	The proposed approach with shunting neurons, filters the noise, and tend to learn the most relevant features. The model accuracy drops down when more than two faces are present in a sequence of images, and fed as input to recognize the location of expression. Accuracy is low for expressions like happiness, fear, and boredom.
	Sun et al. [98]	FABO	D	CNN, Bilateral Long Short-Term Memory Recurrent Neural Networks (BLSTM-RNN), PCA, SVM	The onset-apex-offset video skeleton strategy for frame sequence extraction gives the best recognition rate. The model is not much robust against noisy and stable illumination conditions. The accuracy obtained from video-words (images) are less, when compared to video-skeleton (key clips).

Table 2 (continued)

RC	Reference	Dataset	Type of Data	Approaches	Remarks
	Xie and Hu [115]	CK+, JAFFE	S & D	DCMA-CNN	The focus provided on high-level semantic information from a holistic region and fine-grained information from the local areas has helped in getting an improved performance. Expressional Transformation Invariant (ETI) pooling handles variations like noises, illumination, image rotations, and reduced negative impact. ETI pooling has enhanced the discriminative ability of a model and has helped distinguish sensitive expression elements by fusing different features. Misclassification is highest for sadness emotion class of CK+ and fear emotion class of JAFFE datasets.
	Tang et al. [102]	CK+, Oulu-CASIA, MMI	D	Geometric features, ANN, CNN	The Learning Propagation (LP) method used to fuse geometric and automatically extracted features gives the best results. Tanh activation function applied at the last second layer of ANN architecture helps overcome the gradient explosion and gradient disappearance problem of ANN. Low accuracy has been attained with hand-crafted features.
	Pan et al. [77]	RML, eNTERFACE05	D	CNN (VGG-19) & LSTM	The proposed framework can extract comprehensive features efficiently using aggregation of spatial and temporal approaches, thus fix the gap between visual features and emotions. Expensive, as it uses a massive number of network parameters and consumes vast computation time. Classification of fear emotion class is poor on both datasets.
	Ma et al. [60]	RML, eNTERFACE05, BAUM-1s	D	Audio network: 2D CNN; Video network: 3D CNN, DBN, SVM	Solves the issue of data redundancy and data pollution (denoising) by considering cross-modal feature fusion. The data preparation process takes a lot of time, making the real-time performance of the system more miserable.
RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab					

Table 2 (continued)

RC	Reference	Dataset	Type of Data	Approaches	Remarks
	Sun et al. [99]	CK+, MMI, RaFD	S	MDSTFN	Optical flow information proved to be an effective supplement to spatial features in improving FER's performance from static images. The difference between one neutral-face image and the emotional-face image is employed to compute optical flow extraction instead of evaluating the consecutive sequence of images.
	Liang et al. [50]	CK+, Oulu-CASIA, MMI	D	DSN + DTN + BiLSTM (Inception-w is utilized as a basic network)	Discriminative spatial features are crucial for FER. Average accuracy on Oulu-CASIA and MMI is decreased when convolutional layers were exceeded by three.
	Zhi et al. [126]	DISFA, BP4D	D	3D convolutional neural network (3DLeNet), Boundary Equilibrium Generative Adversarial Net- works (BEGAN), Genetic Algorithm	Among the numerous AU detection algorithms, EvoNet produced the best results. With BEGAN approach, training was easier, and it converges stably. The model is robust and provides better generalization. Training of Deep Neural networks is expensive; hence, identifying proper generation parameters and initial population numbers to avoid hardware limitation is necessary.
Lab & IW					
	Khorrami et al. [39]	CK+, Toronto Face Dataset (TFD)	S	Zero-bias CNN with Data Augmentation and Dropout (AD)	The network was trained quickly, and the number of parameters used to learn was lowered simultaneously, bypassing the bias. When visualizing discriminative spatial patterns, the authors found that most of the filters are excited by the face regions that correspond to Facial AUS.
	Kaya et al. [38]	EmotiW 2015 & 2016 challenge datasets, ChaLearn-LAP first impressions challenge dataset, RECOLA, CK+, MMI	D	SIFT, HOG, LPQ, LBP, LGBP-TOP, deep CNN, Extreme Learning Machine (ELM), Partial Least squares (PLS) regression.	A combination of strategies like weight decay and dropouts with regularization gives the best results. Usage of pre-trained CNN models complement systems with various modalities and helps in efficient feature extraction. The performance of multimodal systems is poor compared to unimodal for arousal predictions on the RECOLA dataset. Emotion classes like disgust and fear give low results on the EmotiW challenge dataset.

Table 2 (continued)

RC	Reference	Dataset	Type of Data	Approaches	Remarks
	Li and Deng [46]	RAF-DB (self-collected Database), CK+, MMI, SFEW 2.0	S	Deep Locality-Preserving Convolutional Neural Network (DLP-CNN)	DLP-CNN with Locality Preserving (LP) loss obtains more discriminative features, which improves the recognition and enhances the system's classification performance. LP loss forms a good and compact intra-class local cluster for each category. The execution time of the proposed method using LP takes more time when compared to the center loss.
Lab & IW					
	Liu et al. [54]	CK+, JAFFE, multi-view BU-3DEF, LFW	S	Conditional CoNERF	The influence of distortion induced in an unconstrained context is avoided by considering salient features retrieved from saliency-guided face patches. The deep salient features contribute to a more accurate description of multi-view facial expression images. The proposed learning strategy utilizes a global deep salient representation and requires few image data compared to existing CNN models.
	Li et al. [48]	FED-RO (self-collected Database), RAF-DB, AffectNet, CK+, MMI, Oulu-CASIA, SFEW	S	CNN with attention mechanism (ACNN)	The issues like partial occlusions are solved using an attention mechanism. The performance has a negative impact due to the misalignment of facial landmark points. Global-local-based ACNN (gACNN) suffers from extremely severe facial occlusions and novel occluders.
	Xie et al. [116]	CK+, JAFFE, TFEID, SFEW, FER2013, BAUM-2i	S	VGG-Face network, Salient Expressional Region Descriptor (SERD) and Multi-Path Variation-Suppressing Network (MPVS-Net)	DAM-CNN jointly uses SERD and MPVS-Net; it can learn discriminative features and perform well in constraint and unconstrained environment. The model overcomes severe overfitting issues. The dropout layer added to the model helps in partly improving the generalization ability of the model. The SERD approach may fail to focus on salient facial regions, due to the vast variations in unconstrained datasets. MPVS-Net (autoencoders and decoders) uses a massive amount of parameters and requires further exploration.

Table 2 (continued)

RC	Reference	Dataset	Type of Data	Approaches	Remarks
	Shao and Qian [96]	CK+, BU-3DFE, FER2013	S	Multi-Task Convolutional Neural Network (MTCNN), LBP, CNN	The model overcomes the issues of deep CNN like overfitting, high computational complexity, and insufficient data sample. A shallow network with few parameters has proved to be efficient on all three datasets. The performance of FER2013 is lower with the combination of LBP and deep CNN architecture. Expression classes like sadness and surprise gave poor accuracy on the FER2013 dataset.
	Georgescu et al. [29]	FER2013, FER+, AffectNet	S	SIFT descriptors and k-means clustering (computed by BoVW model); CNN architectures (VGG-face, VGG-f and VGG-13), k-NN, SVM	Overcomes the overfitting problem of CNN using Dense-Sparse-Dense (DSD) approach during training. Local learning with the SVM approach proved to be efficient when compared to global learning. The proposed approach cannot distinguish between voluntary and involuntary facial expressions.
	Hung et al. [35]	JAFFE, CK+, FER2013, Learning emotion database	S & D	Dense FaceLiveNet	Transfer learning approach solved the problem of a small number of data samples present in a learning-centered emotion dataset. The model does not overcome exceptional real-time situations like occlusions, illumination variations, which can occur in a real-time classroom environment.
RC	Reference	Dataset	Type of Data	Approaches	Remarks
Lab & IW					

Table 2 (continued)

RC	Reference	Dataset	Type of Data	Approaches	Remarks
	Sun and Xia [100]	CK+, JAFFE, FER2013, self-collected database	S	CNN (AlexNet, GoogleNet)	The proposed approach increases speed, accuracy, and reduces computational complexity. Overcomes data overfitting problems using an artificial face augmentation strategy. The model is not robust towards the in-the-wild database, and the recognition rate is lesser compared to the controlled environment. The performance of AlexNet and GoogleNet architectures on anger expression class is lower. The choice of kernel size for choosing the mask needs to be carefully selected. Otherwise, it may lead to the wrong selection of Region of Interest (ROI) and further decrease the system's overall performance.
	Riaz et al. [86]	CK+, FER2013, RAF-DB	S	CNN (Expression Net)	A lightweight network with a small number of parameters reduces overhead. The size of the disk is reduced, and accuracy is improved. The model overcomes the overfitting issue. The model's efficiency needs to be improved for recognition of facial expression in an unconstrained environment.
	Ruan et al. [88]	CK+, MMI, Oulu-CASIA, RAF-DB, SFEW	S & D	FDRL (Backbone network utilized isResnet18)	The proposed FDRL model accurately identifies the expression similarities, expression-specific variations and extraction of fine-grained expression features. There exits a redundancy and noise in latent features, which requires further exploration.
	Cai et al. [8]	BU-3DFE, CK+, MMI, RAF-DB	S & D	IF-GAN consisting of U-Net and PatchGAN; ResNet-101 as expression classifier.	The proposed IF-GAN alleviates identity-related information and produces a facial image for FER that is identity-free. IF-GAN can overcome pose, occlusions and illumination variations.
	Liu et al. [55]	KDEF, BU-3DFE, Multi-PIE, SFEW 2.0	S	DML-Net	The DML-Net reduces deep multiple metric loss, FER loss, and pose-estimation loss by employing dynamically learned loss weights, reducing overfitting, and enhancing recognition significantly. To achieve robust FER performance, the DML-Net reduces the effects of pose and identity.

Table 2 (continued)

RC	Reference	Dataset	Type of Data	Approaches	Remarks
	Liu et al. [56]	BU-3DFE, MMI, AFEW 8.0, DFEW	D	CEFLNet	By determining the emotional intensities of individual video clips, CEFLNet concentrates on the most informative frames for FER. No clip-wise or frame-wise annotations are required for training the model, and training can be done end-to-end.

* *RC* Recording Conditions, *IW* In-the-Wild, *S* Static, *D* Dynamic

the class label. For recognizing emotions from video sequences Kaya et al. [38] has proposed a multimodal approach. Whereas, Ruan et al. [88] proposed Feature Decomposition and Reconstruction Learning (FDRL) method to model expression similarities, characterize the expression-specific variations and reconstruct expression features. Cai et al. [8] explicitly reduced inter-subject variations created by identity-related face attributes by proposing Identity-Free conditional Generative Adversarial Network (IF-GAN). Additionally, Liu et al. [56] proposed a robust video-based FER using a clip-aware emotion-rich feature learning network (CEFLNet) to identify the emotional clips and dynamic facial expressions. CEFLNet tries to avoid the presence of irrelevant frames that render learned features unsuitable for FER and overcome the high computation complexity needed to model video-based facial expression movements.

Xie et al. [116] has proposed a Deep Attentive Multi-path Convolutional Neural Network (DAM-CNN) to locate expression-sensitive regions and has generated high-level representations that are robust against variations like gender, races, etc. Likewise, Li et al. [48], has helped to overcome the problem of facial occlusions and has improved recognition rate on both occluded and non-occluded faces using CNN with Attention (ACNN) mechanism. Li and Deng [46] has utilized Deep Locality-Preserving CNN (DLP-CNN) along with Locality Preserving (LP) loss to form a compact intra-class local clusters for faces belonging to the same emotion classes. This approach has been powerful in handling cross-cultural problems. Sun and Xia [100] has proposed AlexNet, GoogleNet architectures, to improve the accuracy of CNN architecture and has introduced a new augmentation strategy, "artificial face," to overcome the overfitting problem caused by CNN architecture. Also, Shao and Qian [96] has proposed three novel CNN architectures to overcome overfitting, high computational complexity, and shortage of training samples. Khorrami et al. [39] demonstrated that CNNs trained for emotion recognition are able to predict high-level features that firmly match facial AUs both qualitatively and quantitatively. Whereas, Liu et al. [54] proposed a novel conditional convolutional neural network enhanced random forest (CoNERF) for recognizing FER in an unconstrained environment, and the proposed model proved efficient in multi-view FER.

Hung et al. [35] utilized a learning emotion database collected by students of a National University, Taiwan. The authors performed two phases of transfer learning using Dense_FaceLiveNet to solve the problem of small data for classifying learning-centered emotions. To overcome the issue of vast parameters used for model training, Riaz et al. [86] has proposed a shallow net known as eXnet. Liu et al. [55] proposed a dynamically multi-channel metric network (DML-Net) for handling pose-aware and identity-invariant

FER, thus overcoming overfitting and vanishing gradient issues and improving the overall performance.

2.3 Traditional machine learning versus deep learning techniques

The transition from traditional ML to DL algorithms has been ascertained quite effectively for detecting FER on both static and dynamic data. ML models have proven to be efficient when there are fewer data and are trained on fewer parameters; thus, it is unsuitable for large datasets. ML models take handcrafted features as input which is difficult to generalize [35]. Whereas DL contains multiple intermediate layers between the input and classification layers, it can automatically learn higher-level semantic characteristics from vast training data [54, 69, 99]. Using many layers of convolution and pooling layers, the CNN network can extract non-linear face features [28, 68], resulting in more excellent identification rates on the FER system. Although DL models are computationally expensive, they aid in acquiring complex features, and the CNN model's generalization ability is better than traditional ML models [35].

2.4 Summary of FER techniques based on the data

The FER systems can be divided into two categories, the work considering the static images and those that work with dynamic image sequences [40]. In static based approach, the feature vector contains the information about the current image and overlooks the temporal information [57]. On the other hand, the dynamic sequence-based method utilizes the temporal information between one or more frames to recognize the facial expressions. The FER automated systems take static or dynamic images as input and classify the expressions for controlled or uncontrolled scenario data. The dynamic sequence data can extract spatial and temporal characteristics and achieve better performance, but it leads to computational complexity and introduces noise and disruption [99]. Table 3 summarizes the FER techniques based on the type of the data.

2.5 The correlation between action units (AUs) and neural networks

Designing an accurate FER algorithm is vital for developing interactive computer systems in artificial intelligence. Most of the work in the literature has discovered that only a few regions vary as humans change their expressions, and the changes are visible around the subject's facial areas like eyes, nose, and mouth. FACS was introduced by Paul Ekman [26], who identified these regions and outlined how every facial expression is composed of several AUs, each of which corresponds to a different muscle group in the face. Facial expression categories, on the other hand, refer to global changes in facial behavior, Whereas facial AUs represent local variations on the face [126].

FACS-based FER has two stages: first, the AUs are detected, and then the emotion categories are inferred from the identified AUs. A series of AU template sequences need to be produced for each facial expression. This plays a vital role in determining facial expressions from AU analysis. The AU-based/rule-based and appearance-based techniques have been applied in the past. The AU-based method detects individual AUs and labels emotions based on the combination. Even though detecting individual AUs takes time, adding AU information into the neural network would assist in understanding facial expressions with minute changes efficiently [32, 84]. However, an AU detector must be carefully hand-engineered to assure optimal performance [39].

Distinct emotions have different combinations of AUs. By constantly updating the convolution kernel of a CNN, the neural network allows the entire network to learn more effective features. The action units and the features learned by a CNN have a significant and positive association, implying that the CNN's features are the action units [7, 35]. Different convolution kernels tend to learn distinct AUs; hence there is a strong correlation between Ekmans FACS model and neural network.

2.6 Deep learning methodologies used for the evaluation of FER on "in-the-wild" dataset

In-the-wild, FER data is challenging as it imposes various unconstraint variations making the recognition from them a difficult task. Table 4 highlights the state-of-the-art techniques used for the evaluation of in-the-wild datasets.

3 Limitations and possible future directions

This section discusses the shortcomings and prospective future paths, which may help researchers and newcomers better comprehend the opportunities in FER.

1. *Automatic analysis of FER in unconstrained real-time situations is challenging*

   Various factors like head rotation, pose variations, occlusions, illumination variations, differences in age, gender, culture, and skin tone makes it challenging for training and testing the models in an unconstrained environment. It is challenging to select efficient feature extraction and classification techniques, which work well in these varying conditions [61]. Neither conventional nor deep learning approaches are robust to overcome all the challenges in an in-the-wild dataset

Table 3 Summary of FER techniques based on the type of data

Type	Type of data	FER techniques	Performance metric (ACC)
Traditional	S		
		LBP, NCM, SVM [31]	97.25%
		LBP-TOP, Bi-WOOF, SVM [52]	58.85% (CASME II), 62.20% (SMIC-HS)
		Gabor filter, PCA, Genetic Algorithm [6]	96.30% (JAFFE), 94.20% (CK), and 94.26% (CK+)
		DT, VD, Stacking classifier, majority voting [84]	76.47% (CASME II→SAMM), 67.19% (SAMM→CASME II), 58.49% (CASME II→SMIC(HS)), 53.46% (SAMM→ SMIC(HS)), 57.86% (CASME II+SAMM→ SMIC(HS))
		MLP, GoogLeNet, DenseNet, VGG-Face [79]	69.18% (Private Test set)
	D		
		Uniform LBP, SVM [127]	91.53% (CK+), 77.39% (MMI), 80% (GEMEP-FERA)
		LBP, SVM [13]	71.8% (MMI), 56% (AVEC 2011), 76.1% (CK+)
		HOG-TOP, geometric warp feature, SVM [10]	95.7% (CK+), 45.2% (AFEW 4.0)
		ELBPTOP [32]	73.94% (CASME II), 69.06% (SMIC) [with original classes]; 63.44% (SAMM), 79.55% (CASME II) [reorganized classes]
	S & D	LBP, ORB, SVM [75]	88.5% (JAFFE), 93.2% (CK+), 79.8% (MMI)
Deep learning	S		
		SIFT, DNN [122]	85.2% (Multi-PIE), 80.1% (BU3D-FE)
		CNN [57]	96.76%
		MDSTFN [99]	98.38% (CK+), 98.75% (RaFD), 99.59% (MMI)
		Zero-bias CNN with Data Augmentation and Dropout (AD) [39]	89.8% (TFD), 96.4% (CK+)
		DLP-CNN [46]	95.78% (CK+), 51.05% (SFEW 2.0), 78.46% (MMI), 74.20% (RAF-DB (basic)), 44.55% (RAF-DB (compound))
		Conditional CoNERF [54]	94.09% (Multi-View BU-3DFE), 99.02% (CK+ and JAFFE), 60.9% (LFW)
		ACNN [48]	66.50% (FER-RO)
		MTCNN, LBP, CNN [96]	95.29% (CK+), 86.50% (BU-3DFE), 71.14% (FER2013)
		VGG-Face network, SERD, MPVS-Net [116]	95.88% (CK+), 99.32% (JAFFE), 93.36% (TFEID), 61.52% (BAUM-2i), 66.20% (FER2013), 42.30% (SFEW)
		SIFT descriptors, k-means clustering; VGG-face, VGG-f and VGG-13, k-NN, SVM [29]	75.42% (FER2013), 87.76% (FER+), 63.31% (AffectNet)
		AlexNet, GoogleNet [100]	94.67% (CK+), 53.77% (CK+ → JAFFE), 39.13% (CK+ → FER2013), 36.25% (CK+ → IW)
Type	Type of data	FER techniques	Performance metric (ACC)
Deep learning	S		
		CNN (Expression Net) [86]	73.54% (FER2013), 96.75% (CK+), 86.37% (RAF-DB)
		DML-Net [55]	88.2% (KDEF), 83.5% (BU-3DFE), 93.5% (Multi-PIE), 54.39% (SFEW)
	D		
		Facial keypoints, Uniform-LBP, Autoencoders, SOM, SVM [61]	97.55% (MMI), 98.95% (CK+)
		PHRNN, MSCNN [121]	98.50% (CK+), 86.25% (Oulu-CASIA), 81.18% (MMI)
		CNN, LSTM [40]	69.94% (MMI), 58.54% (CASME II)
		Attention CCCNN [4]	95.13%
		CNN, BLSTM-RNN, PCA, SVM [98]	99.57%
		Geometric features, ANN, CNN [102]	98.73% (CK+), 87.50% (Oulu-CASIA)
		VGG-19, LSTM [77]	65.72% (RML), 42.98% (eNTERFACE05)
		2D CNN, 3D CNN, DBN, SVM [60]	82.38% (RML), 85.69% (eNTERFACE05), 59.17% (BAUM-1s)
		DSN+DTN+BiLSTM [50]	99.6% (CK+), 91.07% (Oulu-CASIA), 80.71% (MMI)
		3DLeNet, BEGAN, Genetic Algorithm [126]	86.3% (BP4D)

Table 3 (continued)

Type	Type of data	FER techniques	Performance metric (ACC)
		SIFT, HOG, LPQ, LBP, LGBP-TOP, deep CNN, ELM, PLS regression [38]	54.55% (EmotiW 2015), 52.11% (EmotiW 2016)
		CEFLNet [56]	85.33% (BU-3DFE), 91% (MMI), 53.98% (AFEW), 65.35% (DFEW)
	S & D	DCMA-CNN [115]	93.46% (CK+), 94.75% (JAFFE)
		Dense FaceLiveNet [35]	90.97% (JAFFE), 95.89% (KDEF), 69.99% (FER2013), 79.03% (Learning emotion database), 70.02% (KDEF→FER2013), 91.93% (FER2013→ Learning emotion database)
		FDRL [88]	89.47% (RAF-DB), 62.16% (SFEW), 99.54% (CK+), 85.23% (MMI), 88.26% (Oulu-CASIA)
		IF-GAN, PatchGAN, ResNet-101 [8]	88.33% (RAF-DB),85.25% (BU-3DFE), 97.52% (CK+), 75.48% (MMI)

[*] *S* Static, *D* Dynamic, *ACC* Accuracy

[95]. Hence, it remains a challenge in the FER system. An ideal FER system needs to be developed, which can handle all these challenges in real-life situations, as these factors cause changes in visual appearance and deteriorate FER systems' performance. Emotion recognition from "in the wild" data is a difficult problem that requires several modalities to solve [74, 90]. A combination of visual and acoustic features can help achieve better FER performance in an unconstrained environment [10]. Exploring and fusing such modalities will thus be beneficial to FER and useful in HCI applications. The existing models are trained with fixed expression classes, and classifiers used for evaluations of FER cannot learn incrementally when new expression classes occur [128]. Thus, a FER system is required to detect new expression classes occurring during practical scenarios.

2. *FER from partially occluded faces is still challenging*

   Addressing the issue of facial occlusions is trivial [48], as it varies in their positions and occluders. Also, working on determining specific parameters, for the detection of facial occlusions and proper pre-processing techniques, automatically needs further exploration in the future. Thus, the robust deep learning technique with attention mechanism is of utmost importance to be developed, capable of focusing on unblocked facial patches and perceiving informative features from them to help classify the expressions into intended classes.

3. *Evaluation of real-world databases is more challenging than a lab controlled database*

   Techniques that work well on datasets collected in a controlled environment may work worse when tested in natural and unconstrained environments [4, 48] due to its distinct AU's [46, 47]. Subtle changes in facial features, co-occurrences between AU's, and the low resolution make it harder for the model to recognize these emotions efficiently. The network that learns subtle facial dynamic patterns from micro-expression can efficiently work well on recognizing posed expressions occurring in a lab environment [40]. Also, psychologists can further investigate and explore the types of characteristics (AU's differ for each side of the face, there is also variation in intensity) required to classify the real-world data [119].

4. *Recognizing emotions from videos is a long-standing problem*

   It is challenging to track subtle movements of facial muscles due to the availability of limited datasets. It needs exploiting of powerful features that can characterize facial expressions into emotion categories [77]. But, the identification of such compelling spatio-temporal features is challenging [98]. There is a need to develop a model that can identify appropriate key clips in-frame sequence to extract high-level spatio-temporal hierarchical features, which can help discriminate various facial expressions efficiently.

5. *Elicitation and annotation of micro-expression data are challenging as compared to posed datasets*

   Elicitation of micro-expression databases needs the right choice of emotional stimuli that have high ecological validity. Ground truth labeling of these expressions requires verification from trained experts or psychologists [52]. Hence, it has hindered the progress of research on micro-expression datasets. Annotating the unconstrained database is challenging due to complex and mixed emotion categories. Thus, a proper judgement based approach is required for annotating such blended facial expressions [119]. Few labelers or annotators are available that reduce the reliability and validity of emotions [46]. One option is to perform crowdsourcing rather than involving a few experts to annotate. Another is to apply transfer learning as they

Table 4 In-the-wild datasets and their state-of-the-art techniques

References	Dataset	Approaches	Performance measure	Findings
Yan et al. [117]	AFEW 6.0, CHEAVD	CNN, Bidirectional Recurrent Neural Networks (BRNN), SVM	Accuracy: 55.14% on CHEAVD and 49.22% is achieved on AFEW 6.0 datasets.	The emotions from in-the-wild datasets are recognized efficiently using three cues like facial texture, facial landmark action, and audio signal. The recognition from audio signals is deprived, as it contains noisy data. Recognition of anxious and worried emotion classes is lower, and the highest misclassification is observed in disgust emotion class due to imbalanced and fewer data samples present in these emotion categories.
Nguyen et al. [73]	FER2013, AFEW 7.0	Ensemble of Multi-Level Convolutional Neural Network (MLCNN) and temporal model with an ensemble of MLCNN and 3DCNN.	Accuracy: 74.09% on FER2013 and 49.3% on AFEW datasets.	The addition of mid-level and high-level features from few blocks played a vital role in the classification task. The filter size 3x3 proved to be performing well on most of the image classification problems.
Li et al. [49]	RAF-DB, AffectNet	Resnet-18 with separate loss and softmax loss.	Accuracy: 86.38% (basic) and 58.84% (compound) on RAF-DB and 58.89% on AffectNet datasets.	The softmax loss layer alone is not sufficient enough to discriminate facial expression recognition on an in-the-wild dataset. A separate loss function is required along with the softmax layer to recognize basic and compound expressions efficiently.
Xiaohua et al. [114]	AffectNet	Residual attention block, Bi-directional Recurrent Neural Network (Bi-RNN) with self-attention.	Accuracy: 48%	The two-level attention block yielded the best results. But, the performance of valence was poor as compared to arousal. Tukey's biweight loss function was utilized to reduce the impact of erroneous samples.
Liu et al. [53]	RAF-DB 2.0, FER2013	Point Adversarial Self Mining (PASM) (backbone network used: ResNet-34 and VGG16).	Accuracy: 88.68% on RAF-DB and 73.59% on FER2013 datasets.	The authors focused on training images rather than modifying the network. Searching for a sensitive position in each image is time-consuming. Choice of proper iteration number is expected for PASM to work efficiently.
Koujan et al. [42]	FaceVid	Deep-Exp3D, SVM.	Average accuracy: 87.98%	The network is robust against viewing angle and illumination variations, occlusions, and regresses the expression independently irrespective of the persons' identity.

Table 4 (continued)

Reference	Dataset	Approaches	Performance measure	Findings
Agrawal and Mittal [2]	FER-2013	CNN	Accuracy: Model 1: 65.77% and Model 2: 65.23%	Metric accuracy is unstable for very low and very high kernel sizes.
Reddy et al. [85]	AffectNet	Faster RCNN was used to extract face regions; Facial landmark points and XceptionNet features were extracted; SVM with Radial Basis Function (RBF) was used for classification.	Accuracy: 59%	Deep learning methods fail in an unconstrained situation, as image background dominates over the facial features. Thus, a combination of deep learning and machine learning features is essential in solving such issues.
Yan et al. [119]	RAF-AU	AU detection with CNN (AU-CNN).	Average performance of AU's with AUC-ROC: 88.73; F1-score: 65.95.	It is challenging to categorize expressions in an in-the-wild dataset based on the AU patterns. The annotation of facial expression should include both subjective (judgement-based) and objective (sign-based) elements. Facial action units differ from one side of the face to another, they are not symmetrical, and they vary in the intensity value.
Wang et al. [108]	RAF-DB, FERPlus, AffectNet, WebEmotion	Self-Cure Network (SCN) utilized ResNet 18 as a backbone network.	Accuracy: 88.14% on RAF-DB, 60.23% on AffectNet, and 89.35% on FERPlus datasets.	The proposed SCN reduces the uncertainty caused by ambiguous facial expressions, low quality images and subjectiveness of annotators for large scale FER.
Wang et al. [110]	EmotioNet 2020	Multi-view co-regularization framework with checkpoints and threshold for AU recognition.	Validation accuracy: 74.60%; Testing accuracy: 73.06%	By choosing the optimal checkpoint for each AU, recognition can be improved as AUs converge at varying speeds. The multi-view and the co-regularization loss benefit the supervised training; also, results are better than semi-supervised training.
Wang et al. [109]	FERPlus, RAF-DB, AffectNet, and SFEW	Region Attention Network (RAN) to obtain important facial regions, Region Biased Loss (RB-Loss) was used to assign a high weight to the most important region.	Accuracy: 89.16% on FERPlus, 86.9% on RAF-DB, 59.5% on AffectNet and 56.4% on SFEW datasets.	RAN can effectively capture the action units related to expressions surprise, happiness, and sadness (Cheek raiser, lip corner puller, lip corner depressor). RAN improves the performance of recognition in varying conditions like occlusions and pose variations.

Table 4 (continued)

Reference	Dataset	Approaches	Performance measure	Findings
Liang et al. [51]	FG-Emotions	Multi-Scale Action Unit (AU)-based Network (MSAU-Net) for recognition of images and TMSAU-Net with attention mechanism and temporal stream to jointly learn spatial and temporal features.	Average accuracy: 73.73% on image data; 65.86% on video sequences.	The mouth region needs to be selected when emotion is optimistic, and areas like eyes and eyebrow regions need to be chosen when emotion is passive. The facial muscle movements of positive expressions are large than negative expressions, and the accuracy is also higher.
Zhao et al. [124]	FER2013, FERPlus and FERFIN	Lightweight Emotion Recognition (LER) utilizing DenseNet architecture.	Accuracy on validation set: 71.73% on FER2013, 85.58% on FERPlus; 85.89% on FERFIN datasets.	The proposed LER model addresses the latency in natural conditions and eliminates redundant parameters. The model showed poor recognition on few primary expression classes that need further exploration.
Zhu et al. [128]	RAF-DB, AffectNet	ResNet-18 is used as backbone for DLP-CNN network.	Accuracy: 80.60% on RAF-DB and 82.17% on AffectNet datasets.	The proposed center-expression-distilled loss improved the discriminative quality of deeply learned features and avoided catastrophic forgetting. The new dimension added at the fully connected (FC) layer assigns higher prediction scores to the new expression classes than the old one using incremental learning. Solves the problem of consuming large computation resources.
Vo et al. [107]	RAF-DB, AffectNet, FERPlus	Pyramid With Super Resolution (PSR), Backbone Network: VGG-16	Weighted Accuracy (WA) of 88.98% and Unweighted Accuracy (UA) of 80.78% on RAF-DB; Accuracy of 89.75% on FER+; 60.68% (8 classes) and 63.77% (7 classes) accuracy on AffectNet datasets.	PSR deals with varying image size problems. The SR methods applied to upscale the low-resolution input images improved the network performance. By utilizing prior knowledge of the confusion about each expression, the Prior Distribution Label Smoothing (PDLS) loss function enhanced the FER problem.
Farzaneh and Qi [27]	AffectNet, RAF-DB	Resnet-18, Deep Attentive Center Loss (DACL)	Accuracy: 65.20% on AffectNet and 87.78% on RAF-DB datasets.	To improve feature discrimination, DACL can be used in conjunction with other classification tasks. More research into reducing primary emotion misclassification in in-the-wild database is required.

Table 4 (continued)

Reference	Dataset	Approaches	Performance measure	Findings
Sepas-Moghaddam et al. [95]	Light Field Faces in the Wild (LFFW), Light Field Face Constrained (LFFC)	Resnet-50 + capsule network; VGG-16 + Capsule network	Average Accuracy: 61.59% on LFFW dataset; 87.83%, 89.10%, 86.16% and 88.25% average accuracies based on cross-environment, cross-distance, cross-Pose, cross-dataset protocols on LFFC dataset respectively.	The capsule network adds network value in learning a model that completely utilizes the angular features available in Light Field images. The proposed CapsField requires more training time to extract more discriminative features.
Chen et al. [9]	300-W, AFLW, AffectNet, RAF	VGG16 and ResNet 50 as a backbone network.	Normalized mean errors (NME): 3.49 on 300-W and 1.69 on AFLW.	Residual multi-task learning framework is proposed to carry out landmark localization and expression recognition tasks. Association learning method is further proposed to enhance the two tasks. The proposed models' speed is not fast enough to make it deploy for real-world applications.
Zhu et al. [129]	RAF-DB, SFEW, FER2013	Convolutional Relation Network (CRN)	Mean Accuracy: 56.25% on RAF-DB, 67.32% on FER2013 and 54.87% on SFEW datasets	Few-shot learning is incorporated into the proposed method for transferring discriminative information to determine new emotion classes. Overcame the issue of class imbalance, which is a major challenge in the In-the-wild field.
Saurav et al. [92]	FER2013, FERPlus, RAF-DB	Dual Integrated Convolution Neural Network (DICNN)	Accuracy: 72.77% on FER2013, 85.29% on FERPlus, 86.07% on RAF-DB datasets	DICNN models overcome the issue of computationally intensive and extensive memory storage by using two custom lightweight CNNs. The proposed model utilizes 1.08M model parameters and 5.40MB storage memory and provides the best tradeoff between recognition accuracy and computational efficiency. The proposed system is ideal for real-world applications since it runs in real-time on a resource-constrained embedded platform.

Table 4 (continued)

Reference	Dataset	Approaches	Performance measure	Findings
Nan et al. [71]	RAF-DB, SFEW 2.0	Feature level super-resolution method for robust facial expression recognition (FSR-FER)	Accuracy: 76.66% on RAF-DB dataset with downsampling factor x5, 55.14% on SFEW dataset with downsampling factor x2	The proposed FSR-FER lowers the risk of privacy leaking without recovering high-resolution facial images. The performance is better on low-resolution images with more feature loss. The proposed approach overcomes the FER problem of multi-facial images in crowd scenarios. The introduction of classification-aware loss reweighting into FSR-FER achieves faster training convergence and better performance.

are pre-trained on large datasets, and can be fine-tuned to the given problem. The advent of imaging sensors has presented a new challenge for facial image analysis systems [95]. This field requires further exploitation of inter and intra view relationships to improve the face and expression recognition on an in-the-wild dataset.

6. *Building an automated framework for micro-expression recognition is difficult and needs further research*

   Small minute changes in the facial region are difficult to detect. Semi-automated systems are employed in the detection of micro-expressions. Consequently, there is a lack of a fully integrated framework to analyze micro-expressions [16].

7. *3D FER with deep learning techniques needs further exploration*

   3D data is computationally expensive and time-consuming and limits the widespread of 3D FER [125]. 3D FER database contains a small number of samples, making the deep learning models impotent to train the model [59]. But, it minimizes the drawbacks of 2D data like pose and posture variations, facial occlusions, illumination conditions, and image quality and helps in efficient recognition of facial expressions [20, 76].

8. *Deep learning models are computationally overhead and take a massive number of parameters to train the model, which is a burden.*

   The development of shallow networks (lightweight) is of utmost importance. It can overcome the drawbacks of deep learning models by reducing the number of layers in the model and increasing the performance of hardware-constraint systems [86].

9. *Overfitting caused by insufficient training samples is an significant issue in FER for appropriate recognition and improvement of accuracy*

   Deep learning models can learn complex representations and handle sophisticated learning space. But, it leads towards overfitting [45, 57]. One solution to handle these issues can be usage of data augmentation, dropouts, and simple regularization techniques [17, 86] that can randomly cut out contiguous sections of the input, without manipulating the feature maps during network training. Another possible solution is to build a deep sparse network [17] to overcome overfitting.

10. *Recognizing universal expressions is challenging*

    Primary emotions are universal. FER system poses a lot of challenges in recognizing emotion classes like sadness, fear, and disgust. There is a need for a robust system that efficiently recognizes these universal emotions in both constraint and unconstraint environments. Confusions arise when classifying fear and sadness, fear and surprise, disgust, and sad emotions [48]. Expressions like anger, sadness, and fear vary and depend on ethnicity, which makes recognition and

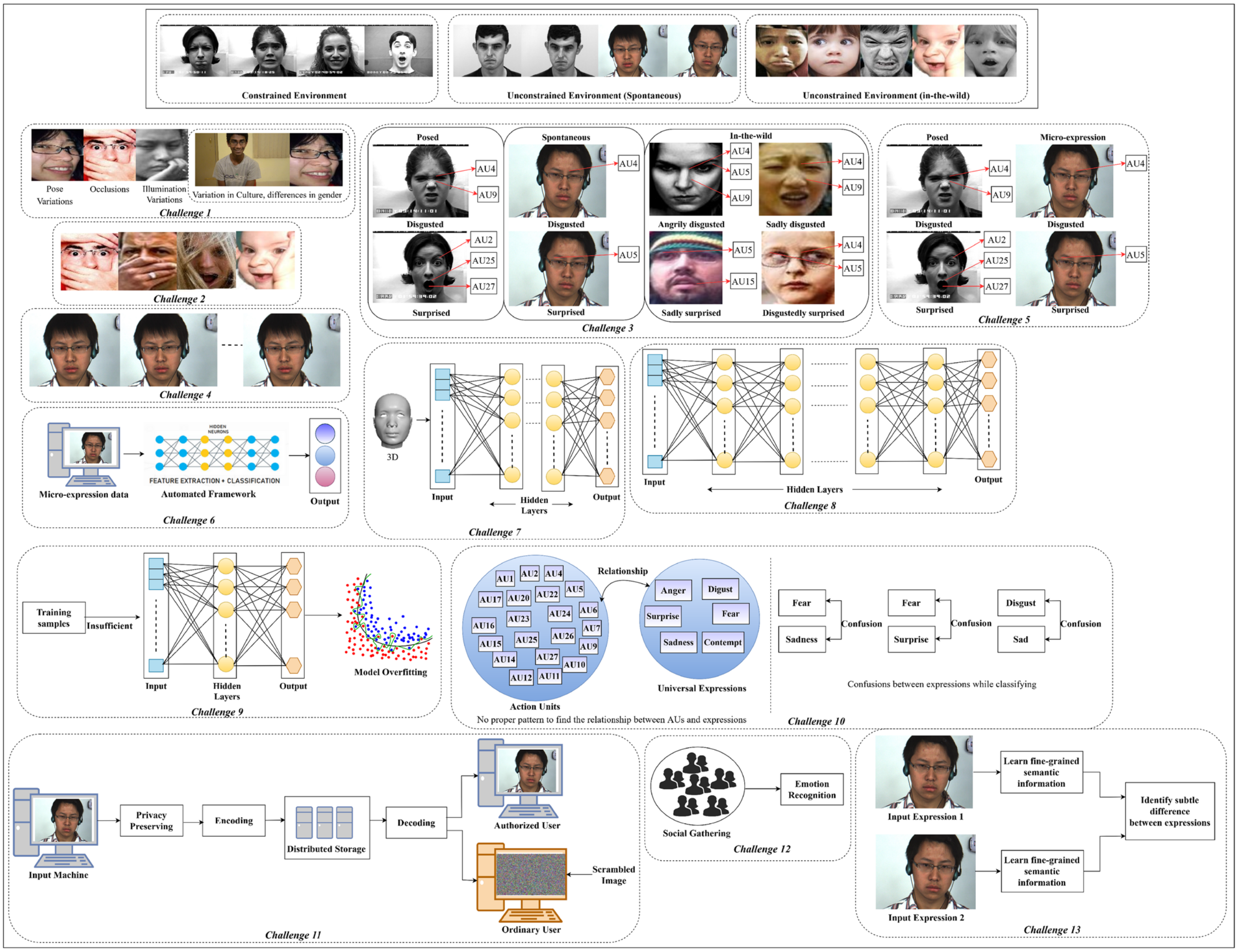


Fig. 2 Research challenges and possible future directions in the field of FER

capturing of such expressions a challenging task [105]. There is no proper consensus for universal expressions from psychologists on the relationship between patterns of AUs and emotions [119] to categorize in-the-wild facial expressions.

11. *Visual privacy-preserving is challenging*

    Most of the past literature systems rely on images that contain high resolution and ignore visual privacy (protection of user privacy). Reliable and accurate privacy-preserving methods [83, 104] are important in human-machine conversation and automatic FER system [11]. Hence, Face scrambling [36] can be a practical solution to overcome privacy issues during video streaming across the real-world data collected from the internet.

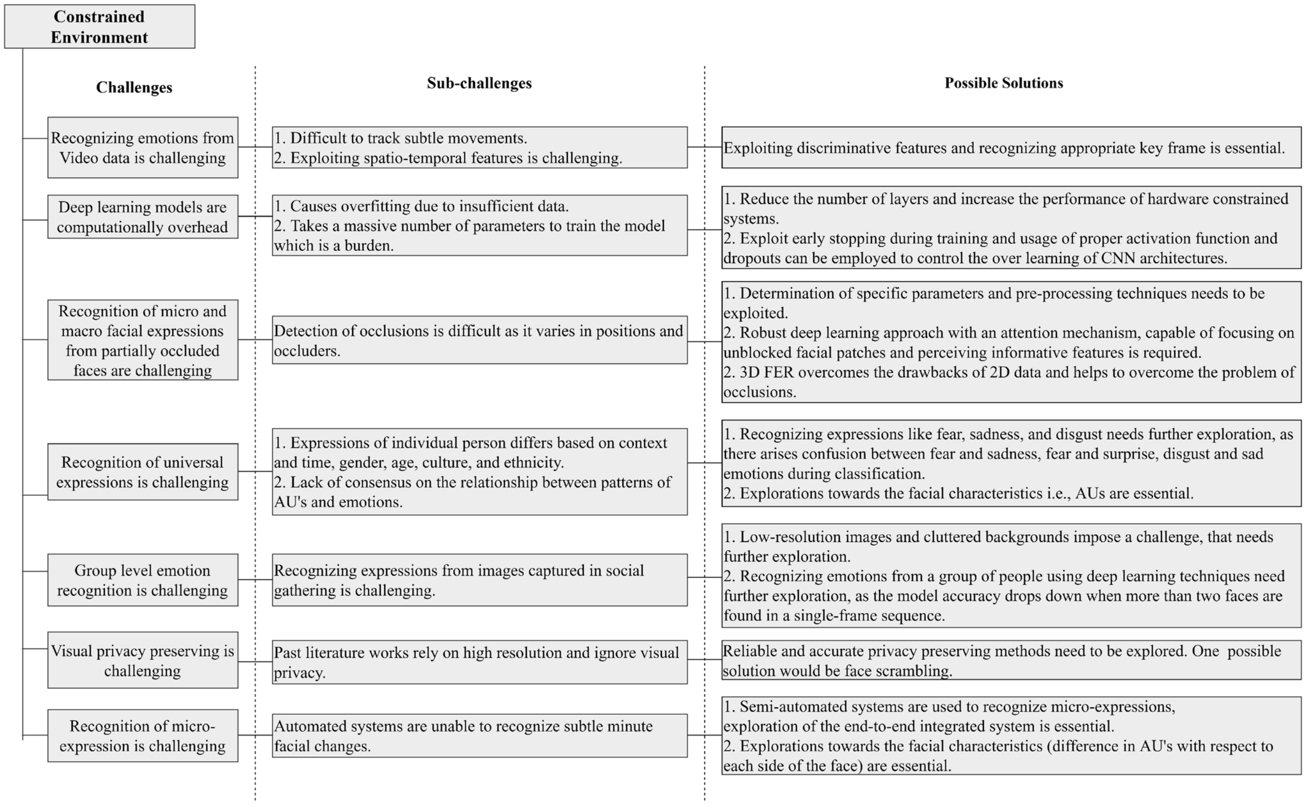
12. *Group level emotion recognition from images are challenging*

    Recognizing expressions from images captured in social gatherings and events is receiving a lot of interest. It requires further exploration as it includes cluttered background, and low image resolutions [101]. When more than two faces are detected in a single picture sequence, the accuracy diminishes. As a result, developing a deep learning model to further investigate this problem is required.

13. *Learning adequate fine-grained semantic information to identify the subtle difference between expressions is challenging*

    The shared latent features can characterize the similarities between different expressions, and expression-specific variations require consideration of corresponding important weights associated with those latent features [88].

The overall summary of the research shortcomings in the area of FER is elaborated in Fig. 2. The facial images recorded in constrained and unconstrained environments are taken from datasets like CK+ [58], CASMEII [118] (The



Constrained Environment

Challenges	Sub-challenges	Possible Solutions
Recognizing emotions from Video data is challenging	1. Difficult to track subtle movements. 2. Exploiting spatio-temporal features is challenging.	Exploiting discriminative features and recognizing appropriate key frame is essential.
Deep learning models are computationally overhead	1. Causes overfitting due to insufficient data. 2. Takes a massive number of parameters to train the model which is a burden.	1. Reduce the number of layers and increase the performance of hardware constrained systems. 2. Exploit early stopping during training and usage of proper activation function and dropouts can be employed to control the over learning of CNN architectures.
Recognition of micro and macro facial expressions from partially occluded faces are challenging	Detection of occlusions is difficult as it varies in positions and occluders.	1. Determination of specific parameters and pre-processing techniques needs to be exploited. 2. Robust deep learning approach with an attention mechanism, capable of focusing on unblocked facial patches and perceiving informative features is required. 2. 3D FER overcomes the drawbacks of 2D data and helps to overcome the problem of occlusions.
Recognition of universal expressions is challenging	1. Expressions of individual person differs based on context and time, gender, age, culture, and ethnicity. 2. Lack of consensus on the relationship between patterns of AU's and emotions.	1. Recognizing expressions like fear, sadness, and disgust needs further exploration, as there arises confusion between fear and sadness, fear and surprise, disgust and sad emotions during classification. 2. Explorations towards the facial characteristics i.e., AUs are essential.
Group level emotion recognition is challenging	Recognizing expressions from images captured in social gathering is challenging.	1. Low-resolution images and cluttered backgrounds impose a challenge, that needs further exploration. 2. Recognizing emotions from a group of people using deep learning techniques need further exploration, as the model accuracy drops down when more than two faces are found in a single-frame sequence.
Visual privacy preserving is challenging	Past literature works rely on high resolution and ignore visual privacy.	Reliable and accurate privacy preserving methods need to be explored. One possible solution would be face scrambling.
Recognition of micro-expression is challenging	Automated systems are unable to recognize subtle minute facial changes.	1. Semi-automated systems are used to recognize micro-expressions, exploration of the end-to-end integrated system is essential. 2. Explorations towards the facial characteristics (difference in AU's with respect to each side of the face) are essential.

Fig. 3 Summary of research challenge and possible future directions in the field of FER based on constrained environment settings

images of the CASMEII dataset is accessed with copyright permission©Xiaolan Fu), SAMM [16], ISED [33] and RAF-DB [47]. The Figs. 3 and 4 summarizes the shortcomings in the field of FER with potential future directions.

4 Applications

This section describes various applications of analyzing the facial expressions[1]. In the field of marketing, recording facial expressions adds quantitative data to self-reports about a product or service. Based on the study of facial expressions, market segments can be measured, and goods can be optimised. In the media and advertising industry, the audience's emotional reaction, as seen on their faces, aids in recognizing movie scenes [72] and rating them as positive or negative, thereby increasing positive emotions during the final release. Monitoring the facial expressions of patients [22] suffering from disorders can significantly promote the success of the underlying cognitive-behavioral therapy, both during the diagnosis and intervention phase in psychological research and medical applications. In website design, monitoring facial expressions of users while handling software or navigating websites helps provide insights such as satisfaction or dissatisfaction and, in turn, gain benefit.

The process of automating and understanding how instructors judge students' engagement via face is an essential application in educational research [113]. Engagement recognition helps the instructors improve the teaching strategy and improve instructional videos based on the viewers' engagement signals. Recognition of facial expression via surveillance camera helps in lie detection and suspicious crime detection. Such crucial information can benefit crime agencies to improve the safeness and take prior actions in case of emergencies. Physical fatigue recognition from facial expression [43], in turn, helps to alert the drivers for safeness. There is a need to develop a robust system to cope with these applications, as it involves recognition of facial expressions from unconstrained environments.

5 Conclusion

Facial behavior is considered an essential source of interpersonal communication. Automatic facial expression analysis is crucial to human–computer interaction and has become an

[1] https://imotions.com/blog/facial-expression-analysis/.

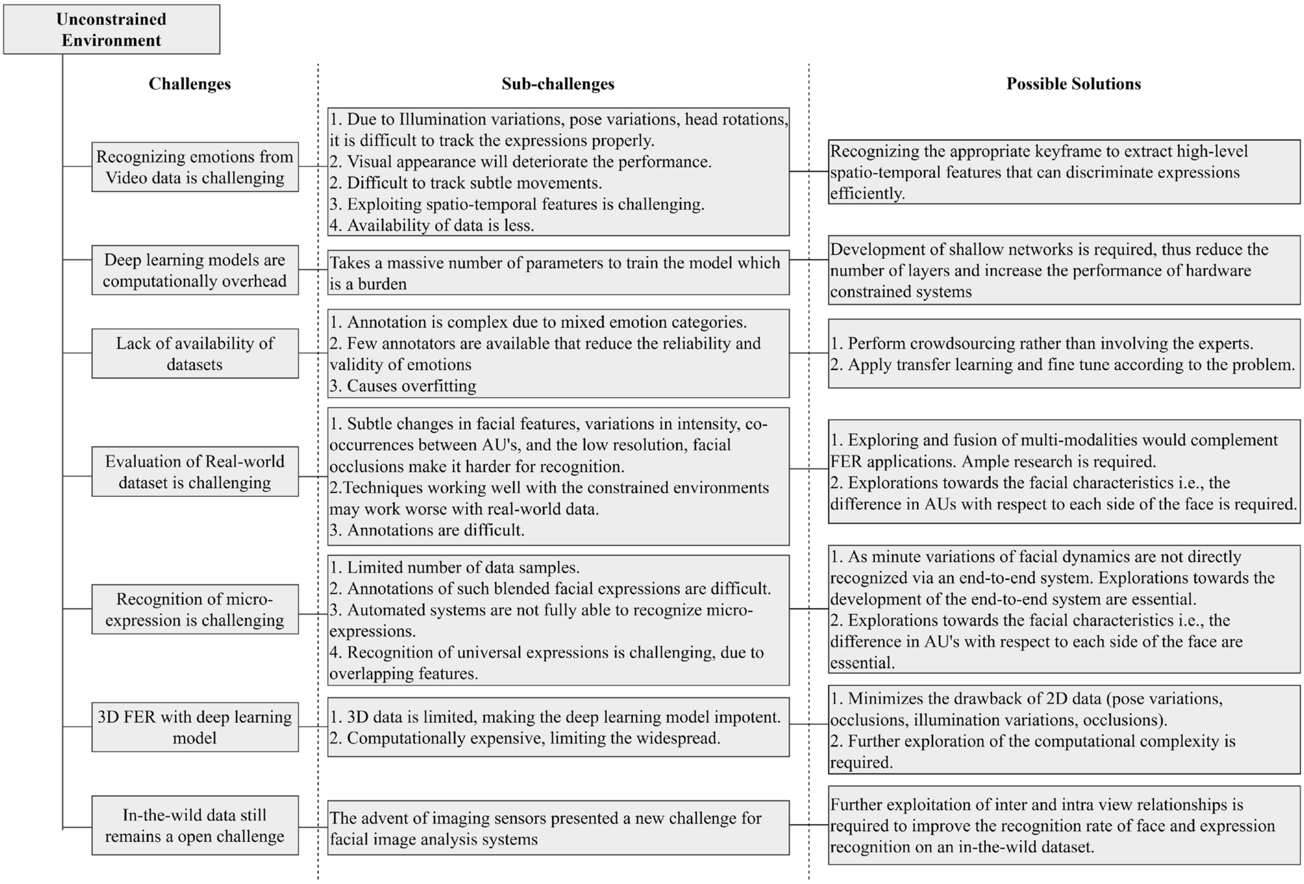

Unconstrained Environment
Challenges
Sub-challenges
Possible Solutions
Recognizing emotions from Video data is challenging
1. Due to Illumination variations, pose variations, head rotations, it is difficult to track the expressions properly.
2. Visual appearance will deteriorate the performance.
2. Difficult to track subtle movements.
3. Exploiting spatio-temporal features is challenging.
4. Availability of data is less.
Recognizing the appropriate keyframe to extract high-level spatio-temporal features that can discriminate expressions efficiently.
Deep learning models are computationally overhead
Takes a massive number of parameters to train the model which is a burden
Development of shallow networks is required, thus reduce the number of layers and increase the performance of hardware constrained systems
Lack of availability of datasets
1. Annotation is complex due to mixed emotion categories.
2. Few annotators are available that reduce the reliability and validity of emotions
3. Causes overfitting
1. Perform crowdsourcing rather than involving the experts.
2. Apply transfer learning and fine tune according to the problem.
Evaluation of Real-world dataset is challenging
1. Subtle changes in facial features, variations in intensity, co-occurrences between AU's, and the low resolution, facial occlusions make it harder for recognition.
2.Techniques working well with the constrained environments may work worse with real-world data.
3. Annotations are difficult.
1. Exploring and fusion of multi-modalities would complement FER applications. Ample research is required.
2. Explorations towards the facial characteristics i.e., the difference in AUs with respect to each side of the face is required.
Recognition of micro-expression is challenging
1. Limited number of data samples.
2. Annotations of such blended facial expressions are difficult.
3. Automated systems are not fully able to recognize micro-expressions.
4. Recognition of universal expressions is challenging, due to overlapping features.
1. As minute variations of facial dynamics are not directly recognized via an end-to-end system. Explorations towards the development of the end-to-end system are essential.
2. Explorations towards the facial characteristics i.e., the difference in AU's with respect to each side of the face are essential.
3D FER with deep learning model
1. 3D data is limited, making the deep learning model impotent.
2. Computationally expensive, limiting the widespread.
1. Minimizes the drawback of 2D data (pose variations, occlusions, illumination variations, occlusions).
2. Further exploration of the computational complexity is required.
In-the-wild data still remains a open challenge
The advent of imaging sensors presented a new challenge for facial image analysis systems
Further exploitation of inter and intra view relationships is required to improve the recognition rate of face and expression recognition on an in-the-wild dataset.

Fig. 4 Summary of research challenges and possible future directions in the field of FER based on unconstrained environment settings

interesting area over the past decades. Considerable work has been carried out in FER's field with a significant increase in the recognition rate using conventional and learning-based approaches. Conventional approaches may recognize facial expressions efficiently in a controlled environment. Significant variations like pose, facial occlusions, illumination changes, and head movements make it difficult for hand-crafted networks to recognize valuable features and classify facial expressions into intended emotion classes. These models cannot attain a reasonable recognition rate when analyzing a few universal emotions due to the confusion with other class labels. ML models may have proved efficient for frontal faces but cannot attain excellent performance on databases collected in an unconstrained environment, which is of utmost importance. Learning-based approaches have gained an increasing ability to achieve state-of-the-art performance in FER. It has the advantage of learning in-depth, discriminative, and abstract patterns from raw images, using multi-layer architecture than hand-crafted features. It tends to learn complex hierarchical feature representations from the images, to improve the classification rate.

This study comprehensively reviews and summarizes the technologies and existing problems in this area. Existing scientific issues, real-world applications, and future directions presented in this paper aid the researchers explore the field efficiently. Even though numerous works have been carried out in the field of FER, systems do suffer from some drawbacks that need some efficient solution and further exploration shortly. There is a requirement to build an adequate system that is robust in constrained and unconstrained environments and aims to achieve accurate recognition as these methods will be exceedingly helpful in real-life scenarios.

Author contributions Conceptualization, literature review, original draft preparation: RAR, supervision, guidance and review: AB.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

References

1. Adolphs, R.: Recognizing emotion from facial expressions: psychological and neurological mechanisms. Behav. Cogn. Neurosci. Rev. **1**(1), 21–62 (2002)
2. Agrawal, A., Mittal, N.: Using cnn for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. Vis. Comput. **36**(2), 405–412 (2020)
3. Azmi, R., Yegane, S.: Facial expression recognition in the presence of occlusion using local gabor binary patterns. In: 20th Iranian Conference on Electrical Engineering (ICEE2012), IEEE, pp 742–747 (2012)
4. Barros, P., Parisi, G.I., Weber, C., Wermter, S.: Emotion-modulated attention improves expression recognition: a deep learning model. Neurocomputing **253**, 104–114 (2017)
5. Bhushan, B.: Study of facial micro-expressions in psychology. In: Understanding facial expressions in communication, Springer, pp 265–286 (2015)
6. Boughida, A., Kouahla, M.N., Lafifi, Y.: A novel approach for facial expression recognition based on gabor filters and genetic algorithm. Evol. Syst. **13**(2), 331–345 (2022)
7. Breuer, R., Kimmel, R.: A deep learning perspective on the origin of facial expressions. arXiv preprint arXiv:1705.01842 (2017)
8. Cai, J., Meng, Z., Khan, A.S., O'Reilly, J., Li, Z., Han, S., Tong, Y.: Identity-free facial expression recognition using conditional generative adversarial network. In: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, pp 1344–1348 (2021)
9. Chen, B., Guan, W., Li, P., Ikeda, N., Hirasawa, K., Lu, H.: Residual multi-task learning for facial landmark localization and expression recognition. Pattern Recogn. **115**, 107893 (2021)
10. Chen, J., Chen, Z., Chi, Z., Fu, H.: Facial expression recognition in video with multiple feature fusion. IEEE Trans. Affect. Comput. **9**(1), 38–50 (2016)
11. Chen, J., Konrad, J., Ishwar, P.: Vgan-based image representation learning for privacy-preserving facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 1570–1579 (2018)
12. Cohn, J.F., Ambadar, Z., Ekman, P.: Observer-based measurement of facial expression with the facial action coding system. The handbook of emotion elicitation and assessment pp 203–221 (2007)
13. Cruz, A.C., Bhanu, B., Thakoor, N.S.: Vision and attention theory based sampling for continuous facial emotion recognition. IEEE Trans. Affect. Comput. **5**(4), 418–431 (2014)
14. Dailey, M.N., Joyce, C., Lyons, M.J., Kamachi, M., Ishi, H., Gyoba, J., Cottrell, G.W.: Evidence and a computational explanation of cultural differences in facial expression recognition. Emotion **10**(6), 874–893 (2010)
15. Darwin, C., Prodger, P.: The expression of the emotions in man and animals. Oxford University Press, USA (1998)
16. Davison, A.K., Lansley, C., Costen, N., Tan, K., Yap, M.H.: Samm: a spontaneous micro-facial movement dataset. IEEE Trans. Affect. Comput. **9**(1), 116–129 (2016)
17. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
18. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.P: Acted facial expressions in the wild database. Australian National University, Canberra, Australia, Technical Report TR-CS-11 2:1 (2011)
19. Dhall, A., Goecke, R., Joshi, J., Sikka, K., Gedeon, T.: Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In: Proceedings of the 16th international conference on multimodal interaction, pp 461–466 (2014)
20. Drira, H., Amor, B.B., Srivastava, A., Daoudi, M., Slama, R.: 3d face recognition under expressions, occlusions, and pose variations. IEEE Trans. Pattern Anal. Mach. Intell. **35**(9), 2270–2283 (2013)
21. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. Proc. Natl. Acad. Sci. **111**(15), E1454–E1462 (2014)
22. Edla, D.R., Ansari, M.F., Chaudhary, N., Dodia, S.: Classification of facial expressions from eeg signals using wpt and svm for wheelchair control operations. Proc. Comput. Sci. **132**, 1467–1476 (2018)
23. Ekenel, H.K., Stiefelhagen, R.: Why is facial occlusion a challenging problem? In: International Conference on Biometrics, Springer, pp 299–308 (2009)
24. Ekman, P.: Lie catching and microexpressions. The philosophy of deception pp 118–133 (2009)
25. Ekman, P., Friesen, W.V.: Nonverbal leakage and clues to deception. Psychiatry **32**(1), 88–106 (1969)
26. Ekman, P., Friesen, W.V.: Facial action coding system: investigator's guide. Consulting Psychologists Press, Washington, DC (1978)
27. Farzaneh, A.H., Qi, X.: Facial expression recognition in the wild via deep attentive center loss. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 2402–2411 (2021)
28. Ge, H., Zhu, Z., Dai, Y., Wang, B., Wu, X.: Facial expression recognition based on deep learning. Computer Methods and Programs in Biomedicine p 106621 (2022)
29. Georgescu, M.I., Ionescu, R.T., Popescu, M.: Local learning with deep and handcrafted features for facial expression recognition. IEEE Access **7**, 64827–64836 (2019)
30. Ghimire, D., Lee, J.: Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. Sensors **13**(6), 7714–7734 (2013)
31. Ghimire, D., Jeong, S., Lee, J., Park, S.H.: Facial expression recognition based on local region specific features and support vector machines. Multimed. Tools Appl. **76**(6), 7803–7821 (2017)
32. Guo, C., Liang, J., Zhan, G., Liu, Z., Pietikäinen, M., Liu, L.: Extended local binary patterns for efficient and robust spontaneous facial micro-expression recognition. IEEE Access **7**, 174517–174530 (2019)
33. Happy, S., Patnaik, P., Routray, A., Guha, R.: The indian spontaneous expression database for emotion recognition. IEEE Trans. Affect. Comput. **8**(1), 131–142 (2017)
34. Hess, U., Thibault, P.: Darwin and emotion expression. Am. Psychol. **64**(2), 120–128 (2009)
35. Hung, J.C., Lin, K.C., Lai, N.X.: Recognizing learning emotion based on convolutional neural networks and transfer learning. Appl. Soft Comput. **84**, 105724 (2019)
36. Jiang, R., Ho, A.T., Cheheb, I., Al-Maadeed, N., Al-Maadeed, S., Bouridane, A.: Emotion recognition from scrambled facial images via many graph embedding. Pattern Recogn. **67**, 245–251 (2017)
37. Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2983–2991 (2015)
38. Kaya, H., Gürpınar, F., Salah, A.A.: Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image Vis. Comput. **65**, 66–75 (2017)
39. Khorrami, P., Paine, T., Huang, T.: Do deep neural networks learn facial action units when doing expression recognition? In: Proceedings of the IEEE international conference on computer vision workshops, pp 19–27 (2015)
40. Kim, D.H., Baddar, W.J., Jang, J., Ro, Y.M.: Multi-objective based spatio-temporal feature representation learning robust to

expression intensity variations for facial expression recognition. IEEE Trans. Affect. Comput. **10**(2), 223–236 (2017)
41. Ko, B.: A brief review of facial emotion recognition based on visual information. sensors **18**(2), 401 (2018)
42. Koujan, M.R., Alharbawee, L., Giannakakis, G., Pugeault, N., Roussos, A.: Real-time facial expression recognition "in the wild" by disentangling 3d expression from identity. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, pp 24–31 (2020)
43. Lee, K., Yoon, H., Song, J., Park, K.: Convolutional neural network-based classification of driver's emotion during aggressive and smooth driving using multi-modal camera sensors. Sensors **18**(4), 957 (2018)
44. Lee, S.H., Ro, Y.M.: Partial matching of facial expression sequence using over-complete transition dictionary for emotion recognition. IEEE Trans. Affect. Comput. **7**(4), 389–408 (2016)
45. Li, S., Deng, W.: Deep facial expression recognition: A survey. arXiv preprint arXiv:1804.08348 (2018a)
46. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Trans. Image Process. **28**(1), 356–370 (2018)
47. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2852–2861 (2017)
48. Li, Y., Zeng, J., Shan, S., Chen, X.: Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Trans. Image Process. **28**(5), 2439–2450 (2018)
49. Li, Y., Lu, Y., Li, J., Lu, G.: Separate loss for basic and compound facial expression recognition in the wild. In: Asian Conference on Machine Learning, pp 897–911 (2019)
50. Liang, D., Liang, H., Yu, Z., Zhang, Y.: Deep convolutional bilstm fusion network for facial expression recognition. Vis. Comput. **36**(3), 499–508 (2020)
51. Liang, L., Lang, C., Li, Y., Feng, S., Zhao, J.: Fine-grained facial expression recognition in the wild. IEEE Trans. Inf. Forensics Secur. **16**, 482–494 (2020)
52. Liong, S.T., See, J., Wong, K., Phan, R.C.W.: Less is more: micro-expression recognition from video using apex frame. Signal Process. **62**, 82–92 (2018)
53. Liu, P., Lin, Y., Meng, Z., Deng, W., Zhou, J.T., Yang, Y.: Point adversarial self mining: A simple method for facial expression recognition in the wild. arXiv preprint arXiv:2008.11401 (2020)
54. Liu, Y., Yuan, X., Gong, X., Xie, Z., Fang, F., Luo, Z.: Conditional convolution neural network enhanced random forest for facial expression recognition. Pattern Recogn. **84**, 251–261 (2018)
55. Liu, Y., Dai, W., Fang, F., Chen, Y., Huang, R., Wang, R., Wan, B.: Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition. Inf. Sci. **578**, 195–213 (2021)
56. Liu, Y., Feng, C., Yuan, X., Zhou, L., Wang, W., Qin, J., Luo, Z.: Clip-aware expressive feature learning for video-based facial expression recognition. Inf. Sci. **598**, 182–195 (2022)
57. Lopes, A.T., de Aguiar, E., De Souza, A.F., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn. **61**, 610–628 (2017)
58. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 ieee computer society conference on computer vision and pattern recognition-workshops, IEEE, pp 94–101 (2010)
59. Ly, T.S., Do, N.T., Kim, S.H., Yang, H.J., Lee, G.S.: A novel 2d and 3d multimodal approach for in-the-wild facial expression recognition. Image and Vision Computing (2019)
60. Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., Košir, A.: Audio-visual emotion fusion (avef): a deep efficient weighted approach. Inform. Fusion **46**, 184–192 (2019)
61. Majumder, A., Behera, L., Subramanian, V.K.: Automatic facial expression recognition system using deep network-based data fusion. IEEE Transact. Cybern. **48**(1), 103–114 (2016)
62. Martinez, B., Valstar, M.F., Jiang, B., Pantic, M.: Automatic analysis of facial actions: A survey. IEEE transactions on affective computing (2017)
63. Matias, R., Cohn, J.F., Ross, S.: A comparison of two systems that code infant affective expression. Dev. Psychol. **25**(4), 483 (1989)
64. Mavadati, M., Sanger, P., Mahoor, M.H.: Extended disfa dataset: Investigating posed and spontaneous facial expressions. In: proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1–8 (2016)
65. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: Disfa: a spontaneous facial action intensity database. IEEE Trans. Affect. Comput. **4**(2), 151–160 (2013)
66. Min, R., Hadid, A., Dugelay, J.L.: Improving the recognition of faces occluded by facial accessories. In: Proc. IEEE Int. Conf. Face and Gesture, pp 442–447 (2011)
67. Min, R., Hadid, A., Dugelay, J.L.: Efficient detection of occlusion prior to robust face recognition. The Scientific World Journal 2014 (2014)
68. Minaee, S., Abdolrashidi, A., Su, H., Bennamoun, M., Zhang, D.: Biometrics recognition using deep learning: A survey. arXiv preprint arXiv:1912.00271 (2019)
69. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter conference on applications of computer vision (WACV), IEEE, pp 1–10 (2016)
70. Namba, S., Makihara, S., Kabir, R.S., Miyatani, M., Nakao, T.: Spontaneous facial expressions are different from posed facial expressions: morphological properties and dynamic sequences. Curr. Psychol. **36**(3), 593–605 (2017)
71. Nan, F., Jing, W., Tian, F., Zhang, J., Chao, K.M., Hong, Z., Zheng, Q.: Feature super-resolution based facial expression recognition for multi-scale low-resolution images. Knowl.-Based Syst. **236**, 107678 (2022)
72. Navarathna, R., Carr, P., Lucey, P., Matthews, I.: Estimating audience engagement to predict movie ratings. IEEE Trans. Affect. Comput. **10**(1), 48–59 (2017)
73. Nguyen, D.H., Kim, S., Lee, G.S., Yang, H.J., Na, I.S., Kim, S.H.: Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks. IEEE Transactions on Affective Computing (2019)
74. Nigam, S., Singh, R., Misra, A.: A review of computational approaches for human behavior detection. Arch. Comput. Method. Eng. **26**(4), 831–863 (2019)
75. Niu, B., Gao, Z., Guo, B.: Facial expression recognition with lbp and orb features. Computational Intelligence and Neuroscience 2021 (2021)
76. Nonis, F., Dagnes, N., Marcolin, F., Vezzetti, E.: 3d approaches and challenges in facial expression recognition algorithms-a literature review. Appl. Sci. **9**(18), 3904 (2019)
77. Pan, X., Ying, G., Chen, G., Li, H., Li, W.: A deep spatial and temporal aggregation framework for video-based facial expression recognition. IEEE Access **7**, 48807–48815 (2019)
78. Pfister, T., Li, X., Zhao, G., Pietikäinen, M.: Recognising spontaneous facial micro-expressions. In: Proc. IEEE Int. Conf. Com-put. Vision, pp 1449–1456 (2011)
79. Pham, T.T.D., Kim, S., Lu, Y., Jung, S.W., Won, C.S.: Facial action units-based image retrieval for facial expression recognition. IEEE Access **7**, 5200–5207 (2019)

80. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Transact. Pattern Anal. Mach. Intell. **10**, 1175–1191 (2001)
81. Polikovsky, S., Kameda, Y., Ohta, Y.: Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In: Proc. IEEE Third Int. Conf. Crime Detection and Prevention (ICDP), pp 1–6 (2009)
82. Qu, F., Wang, S.J., Yan, W.J., Li, H., Wu, S., Fu, X.: CAS $(\mathrm{ME})^2$: a database for spontaneous macro-expression and micro-expression spotting and recognition. IEEE Trans. Affect. Comput. **9**(4), 424–436 (2017)
83. Rahulamathavan, Y., Phan, R.C.W., Chambers, J.A., Parish, D.J.: Facial expression recognition in the encrypted domain based on local fisher discriminant analysis. IEEE Trans. Affect. Comput. **4**(1), 83–92 (2012)
84. Rashmi, R.A., Annappa, B.: Micro expression recognition using delaunay triangulation and voronoi tessellation. IETE J. Res. **0**(0), 1–17 (2022)
85. Reddy, G.V., Savarni, C.D., Mukherjee, S.: Facial expression recognition in the wild, by fusion of deep learnt and hand-crafted features. Cogn. Syst. Res. **62**, 23–34 (2020)
86. Riaz, M.N., Shen, Y., Sohail, M., Guo, M.: Exnet: an efficient approach for emotion recognition in the wild. Sensors **20**(4), 1087 (2020)
87. Rinn, W.E.: The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. Psychol. Bull. **95**(1), 52 (1984)
88. Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., Wang, H.: Feature decomposition and reconstruction learning for effective facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7660–7669 (2021)
89. Russell, J.A.: Culture and the categorization of emotions. Psychol. Bull. **110**(3), 426–450 (1991)
90. Samadiani, N., Huang, G., Cai, B., Luo, W., Chi, C.H., Xiang, Y., He, J.: A review on automatic facial expression recognition systems assisted by multimodal sensor data. Sensors **19**(8), 1863 (2019)
91. Samal, A., Iyengar, P.A.: Automatic recognition and analysis of human faces and facial expressions: a survey. Pattern Recogn. **25**(1), 65–77 (1992)
92. Saurav, S., Gidde, P., Saini, R., Singh, S.: Dual integrated convolutional neural network for real-time facial expression recognition in the wild. Vis. Comput. **38**(3), 1083–1096 (2022)
93. Sebe, N., Cohen, I., Gevers, T., Huang, T.S.: Multimodal approaches for emotion recognition: a survey. Internet Imaging VI Int. Soc. Optics Photon. **5670**, 56–68 (2005)
94. Sebe, N., Lew, M.S., Sun, Y., Cohen, I., Gevers, T., Huang, T.S.: Authentic facial expression analysis. Image Vis. Comput. **25**(12), 1856–1863 (2007)
95. Sepas-Moghaddam, A., Etemad, A., Pereira, F., Correia, P.L.: Capsfield: Light field-based face and expression recognition in the wild using capsule routing. arXiv preprint arXiv:2101.03503 (2021)
96. Shao, J., Qian, Y.: Three convolutional neural network models for facial expression recognition in the wild. Neurocomputing **355**, 82–92 (2019)
97. Shreve, M., Godavarthy, S., Goldgof, D., Sarkar, S.: Macro-and micro-expression spotting in long videos using spatio-temporal strain. In: Face and Gesture 2011, IEEE, pp 51–56 (2011)
98. Sun, B., Cao, S., He, J., Yu, L.: Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy. Neural Netw. **105**, 36–51 (2018)
99. Sun, N., Li, Q., Huan, R., Liu, J., Han, G.: Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recogn. Lett. **119**, 49–61 (2019)
100. Sun, X., Pingping, Xia SL.: A roi-guided deep architecture for robust facial expressions recognition. Inf. Sci. **522**, 35–48 (2020)
101. Tan, L., Zhang, K., Wang, K., Zeng, X., Peng, X., Qiao, Y.: Group emotion recognition with individual facial emotion cnns and global image based cnns. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp 549–552 (2017)
102. Tang, Y., Zhang, X.M., Wang, H.: Geometric-convolutional feature fusion based on learning propagation for facial expression recognition. IEEE Access **6**, 42532–42540 (2018)
103. Tian, Y.L., Kanade, T., Cohn, J.F.: Facial expression analysis. In: Handbook of face recognition, Springer, pp 247–275 (2005)
104. Ullah A, Wang J, Anwar MS, Ahmad A, Nazir S, Khan HU, Fei Z.: Fusion of machine learning and privacy preserving for secure facial expression recognition. Secur. Commun. Netw. 2021 (2021) 1–12
105. Valstar, M., Pantic, M.: Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In: Proc, pp. 65–70. Int'l Conf. Language Resources and Evaluation, Workshop EMOTION (May 2010)
106. Valstar, M.F., Mehu, M., Jiang, B., Pantic, M., Scherer, K.: Meta-analysis of the first facial expression recognition challenge. IEEE Transact. Syst. Man Cybern Part B (Cybern) **42**(4), 966–979 (2012)
107. Vo, T.H., Lee, G.S., Yang, H.J., Kim, S.H.: Pyramid with super resolution for in-the-wild facial expression recognition. IEEE Access **8**, 131988–132001 (2020)
108. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6897–6906 (2020a)
109. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans. Image Process. **29**, 4057–4069 (2020b)
110. Wang, P., Wang, Z., Ji, Z., Liu, X., Yang, S., Wu, Z.: Tal emotionet challenge 2020 rethinking the model chosen problem in multi-task learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 412–413 (2020c)
111. Wang, S., Wu, C., He, M., Wang, J., Ji, Q.: Posed and spontaneous expression recognition through modeling their spatial patterns. Mach. Vis. Appl. **26**(2–3), 219–231 (2015)
112. Weber, R., Soladié, C., Séguier, R.: A survey on databases for facial expression analysis. In: VISIGRAPP (5: VISAPP), pp 73–84 (2018)
113. Whitehill, J., Serpell, Y.C., Foster, A., Movellan, J.R.: The faces of engagement: automatic recognition of student engagement from facial expressions. IEEE Trans. Affect. Comput. **5**(1), 86–98 (2014)
114. Xiaohua, W., Muzi, P., Lijuan, P., Min, H., Chunhua, J., Fuji, R.: Two-level attention with two-stage multi-task learning for facial emotion recognition. J. Vis. Commun. Image Represent. **62**, 217–225 (2019)
115. Xie, S., Hu, H.: Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. IEEE Trans. Multimed. **21**(1), 211–220 (2018)
116. Xie, S., Hu, H., Wu, Y.: Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. Pattern Recogn. **92**, 177–191 (2019)
117. Yan, J., Zheng, W., Cui, Z., Tang, C., Zhang, T., Zong, Y.: Multicue fusion for emotion recognition in the wild. Neurocomputing **309**, 27–35 (2018)
118. Yan, W.J., Li, X., Wang, S.J., Zhao, G., Liu, Y.J., Chen, Y.H., Fu, X.: Casme ii: an improved spontaneous micro-expression

database and the baseline evaluation. PLoS ONE **9**(1), e86041 (2014)

119. Yan, W.J., Li, S., Que, C., Pei, J., Deng, W.: Raf-au database: In-the-wild facial expressions with subjective emotion judgement and objective au annotations. In: Proceedings of the Asian Conference on Computer Vision (2020)
120. Zhang, F., Yu, Y., Mao, Q., Gou, J., Zhan, Y.: Pose-robust feature learning for facial expression recognition. Front. Comp. Sci. **10**(5), 832–844 (2016)
121. Zhang, K., Huang, Y., Du, Y., Wang, L.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans. Image Process. **26**(9), 4193–4203 (2017)
122. Zhang, T., Zheng, W., Cui, Z., Zong, Y., Yan, J., Yan, K.: A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Trans. Multimed. **18**(12), 2528–2536 (2016)
123. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: From facial expression recognition to interpersonal relation prediction. Int. J. Comput. Vision **126**(5), 550–569 (2018)
124. Zhao, G., Yang, H., Yu, M.: Expression recognition method based on a lightweight convolutional neural network. IEEE Access **8**, 38528–38537 (2020)
125. Zhao, X., Zou, J., Li, H., Dellandréa, E., Kakadiaris, I.A., Chen, L.: Automatic 2.5-d facial landmarking and emotion annotation for social interaction assistance. IEEE Transact Cybern. **46**(9), 2042–2055 (2015)
126. Zhi, R., Zhou, C., Li, T., Liu, S., Jin, Y.: Action unit analysis enhanced facial expression recognition by deep neural network evolution. Neurocomputing **425**, 135–148 (2021)
127. Zhong, L., Liu, Q., Yang, P., Huang, J., Metaxas, D.N.: Learning multiscale active facial patches for expression analysis. IEEE Transact. Cybern. **45**(8), 1499–1510 (2014)
128. Zhu, J., Luo, B., Zhao, S., Ying, S., Zhao, X., Gao, Y.: Iexpressnet: Facial expression recognition with incremental classes. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 2899–2908 (2020)
129. Zhu, Q., Mao, Q., Jia, H., Noi, O.E.N., Tu, J.: Convolutional relation network for facial expression recognition in the wild with few-shot learning. Expert Syst. Appl. **189**, 116046 (2022)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.