JEI Journal of Electronic Imaging

REVIEW

Comprehensive review and analysis on facial emotion recognition methods

Rishabh Vats, Manoj Kumar,* Ritu Rai, and Gunjan Verma
GLA University, Department of Computer Engineering and Applications, Mathura, Uttar Pradesh, India

ABSTRACT. Facial emotion recognition (FER) is a significant subject in computer vision and artificial intelligence because of its tremendous academic and commercial potential, such as in cognitive science, health care, virtual reality, and video conferencing in various domains. While FER could be carried out using multiple sensors, this review includes research that exclusively uses face images, since facial expressions are among the key pieces of knowledge in human interactions. We give a brief review of FER research carried out in recent years. We divided the techniques of FER mainly into two parts, i.e., based on the type of approach and based on the type of frame, which is further subdivided into two sub-parts of each classification for more detailed division. First, it explains traditional FER approaches together with a description of the representative classes of FER systems. Deep-learning FER strategies are then addressed using deep networks that allow "end-to-end" learning. This review is also directed toward an up-to-date deep learning strategy, which is a trending topic nowadays. A brief overview of publicly accessible assessment metrics is provided in the later part of this paper, as well as a comparison with the baseline results is presented, which is a norm for quantitative analysis of FER research. The whole analysis also could act as a concise field guide for beginners in the FER sector, providing general details and a common understanding of both the recent state-of-the-art research and established researchers searching for fruitful areas for further research.

[DOI: 10.1117/1.JEI.32.4.040901]

Keywords: facial emotion recognition; conventional facial emotion recognition; deep learning facial emotion recognition; convolutional neural networks; facial action coding system

Paper 220903V received Sep. 2, 2022; revised Jun. 27, 2023; accepted Aug. 7, 2023; published Aug. 24, 2023.

1 Introduction

Facial expression of emotion is a significant factor in human communication that enables us to better understand the motives of other people. People frequently deduce other people's emotional states from their facial expressions and verbal tone, such as sorrow, anger, and happiness.[1] Non-verbal factors account for two-thirds of human communication, whereas language modules account for one-third. By carrying emotional meaning, facial expressions, like other nonverbal components, are one of the most important sources of knowledge in interpersonal communication.[2,3]

Owing to its functional importance in socially adept robotics, medical treatment, fatigue tracking for drivers, and many other computer–human interaction programs, countless studies on automated facial expression analysis have been carried out. To encode expression information

*Address all correspondence to Manoj Kumar, manoj.kumar@gla.ac.in from face representations, a variety of facial expression recognition (FER) systems were investigated in the fields of computer vision and machine learning.[4] In the early 20th century, six primary emotions were defined by Ekman and Friesen,[5] based on a cross-cultural study. This implied that regardless of background, the human species interprets those essential emotions in the same way.[6] These quintessential facial expressions are disgust, anger, fear, happiness, sorrow, and surprise.

The model of impact has always been on the basis of basic emotions, the capacity to reflect the essence, and subtlety of our regular adaptive displays,[7–9] and some other models of emotional portrayal, such as the facial action coding system (FACS)[10] and the steady model using effective dimensions, reflect a broader spectrum of emotion.[11]

2 Face Emotion Detection Techniques

The face emotion detection technique is divided into two categories: based on the type of approach and based on the type of frame. Figure 1 shows the categorization of different types of face recognition approaches.

2.1 Based on the Type of Approach

Based on the type of approach, the face emotion detection technique is categorized as

(a) Conventional method

(b) Deep learning method

2.1.1 *Conventional method*

The FER consists of three main steps in conventional FER approaches: (1) face acquisition, (2) extraction of feature, and (3) classification of expression.

Initially, a facial image is identified from the given source image and face marking (e.g., nose and both eye). On the other hand, mostly temporal and the spatial feature have been obtained from sub-parts of face. At last, the classifiers that have been pre-trained generate the recognition result using the extracted features.[1]

Happy et al.[12] envisaged a face emotion classification algorithm that uses a Haar classifier for detection of the face area, a local binary patterns (LBP) histogram having a significant difference in face image block size as a feature vector, and categorizing of different face emotions utilizing principal component analysis (PCA). The complete process is conducted in real-time for emotion categorization due to the system's minimal computational complexity. For the study of facial emotion, a flexible approach is recommended, as the different emotions and frequency of emotions deviate from one to another person. A person's frontal grayscale images are used by the system that classifies the six basic essential emotions: happiness, sadness, disgust, fear, surprise, and anger. The test results show that the accuracy of recognition of facial expressions is greater than 97% by using LBP features. The LBP histogram block features extract both local and global face image features resulting in higher precision, including an image of high resolution,

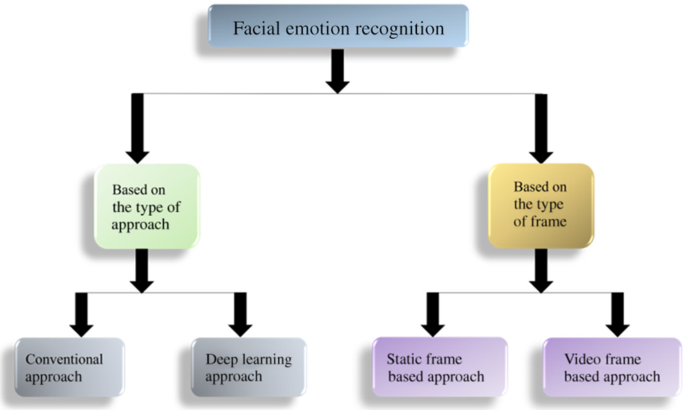
Facial emotion recognition
Based on the type of approach
Based on the type of frame
Conventional approach
Deep learning approach
Static frame based approach
Video frame based approach

Fig. 1 Classification of various FER techniques.

with a greater number of cascaded LBP block features improving the classification efficiency, although, due to the period's associated ambiguity, it is technically challenging to implement in real time. The methodology suggested is restricted to categorizing just the frontal image. But face rotation or occlusions undermine throughput of the system.

Fabian et al.[13] proposed an algorithm for computer vision to transcribe database that consists of around 1 million frames of emotional facial expressions specifically in the wild (i.e., facial images found on the internet). First, the author showed how this newly proposed algorithm would reliably identify units of action (AUs) through databases and their intensities. The program also processes photos and video frames in real-time (>30 frames per second), enabling it to handle massive amounts of data. Second, Word Net is used to download from the internet 1 million pictures of facial expressions associated with emotion keywords. This algorithm then automatically annotates these images with clusters of AUs, AU intensities, and emotions.

Gunawan et al.[14] presented an approach that detected facial expressions by using fuzzy logic to observe the shift in key features inactive appearance model. Fuzzy logic is utilized for the evaluation of emotion based on the previous experience of the FACS. The experimental outcome collection of data produced on users reveals that (i) FACS does not provide data regarding the degree of muscle activation, so the degree of core feature must be derived from some specific dataset, and (ii) the accurateness of emotion detection relies on the emotion themselves.

Szwoch and Pieniążek[15] presented an approach for facial emotion and its recognition solely based on the Microsoft Kinect sensor depth channel. There are nine emotions in the emotional user model, including the neutral one. The suggested approach recognizes local motion within the face region to analyze actual facial emotion. This technique completely ignores lighting circumstances and identifies a closer distance between the sensor and consumer, despite the average recognition accuracy rate being slightly greater than 50%. In order to assist various optical channel-based algorithms and make use of skeletal or face-tracking information, the proposed approach can be used.

Ghimire et al.[16] proposed a method for defining facial emotions out of a picture frame using the mixture of personality and geometrical features with that of a support vector machine (SVM). In general, the face characteristics are determined by splitting the face area onto a generic grid (holistic representation) for FER. Separating the full face area into local zones related to different fields also allows this paper to recover region-specific appearance features. Specific regions of the related domain are also used to derive geometric features. However, significant local regions are calculated using an incremental search method, and this results in a reduction of the function dimension and an increase in the precision. In addition, the findings of FER employing features across different domain territories are contrasted with those achieved utilizing holistic representation. According to empirical evidence, local region-based representations perform better than holistic representations.

Shen et al.[17] proposed a sudden process for the identification of facial emotions utilizing thermal infrared clips. First, the series features are derived from the samples of different face sub-regions in horizontal and vertical thermal infrared temperature differences. Second, a subset of features is chosen based on their F-values. And third, with a weak classifier of k-nearest neighbor, the Adaboost algorithm has been in use for classifying face emotions in dimensions of arouse. Lastly, tests on the Natural Visible and Infrared Facial Expression (USTC-NVIE) show the method's correctness. At 75.3% and 76.7%, the top most classification accuracies are achieved between maximum and minimum arousal as well as valence, which means that our approach is successful in classifying words. Further analysis depicts that the mouth area characteristics give the most important character in both valance as well as arousal category relative to the characteristics of other face regions.

Khan et al.[18] presented a model that can identify facial expressions effectively and with high accuracy, including for facial images with minimal resolution. The proposed framework is efficient in time and memory as it extracts only texture features from the perceptional prominent sectors of the face region in a pyramidal fashion. The framework is tested on several databases, including Cohn–Kanade (CK+) posed facial expression database, the random emotion of MMI facial expression database, and the FG-NET facial expression and emotion database (FEED). The presented pyramidal local binary pattern (PLBP) features are incorporated for performing person-independent recognition facial expressions. Utilizing the algorithm named Viola–Jones object detection, the region of interest was automatically obtained and initiated in the process to get PLBP feature vector.

Ghimire and Lee[19] presented an approach for the complete automatic detection of a facial emotion in the series of face images. In the consecutive video frames, as an evolution of facial expression over time, the tracking of landmark automatically uses displacements based on the elastic bunch graph matching displacement estimates. With regard to the first image in the image series, it collected and standardized feature vectors from independent markers and also pair of landmark tracking data. The prototypical emotion series for every group in face emotion, generated at the median position for landmarks monitoring outcomes by those in the facial emotion training sequences. Multi-class AdaBoost including dynamic time-warping similarities gap among prototypical facial emotion and facial input emotion feature vector is usually a weaker classifier for selecting a subset for discriminatory feature vectors. Lastly, two techniques of FER techniques are introduced, one is utilizing dynamic time-warping multi-class AdaBoost and another using the enhanced feature vector support system.

Siddiqi et al.[20] inducted a precise and robust system for FER. The system proposed to extract features of FER framework that used stepwise linear discriminant analysis (SWLDA). To reduce class variance and the low variance between different emotional classes, SWLDA focuses on selecting regional characteristics from its emotional frames that employ a portion of F-test values. The model was used to classify the hidden conditional random fields (HCRFs). Using a mixture having Gaussian density functions, HCRF can approximate a complex distribution. The system uses a strategy of hierarchical recognition to achieve optimum results. Under the given settings, the emotions are categorized into three groups on the basis of facial parts that mostly contribute to emotion. During identification, SWLDA and HCRF are used at the first level to identify the category of expression; at the next step, a separate subset of SWLDA and HCRF is used to define the class of expression within the defined category, which has been trained for that particular category only. A total of four tests were done using four publicly available data sets to validate the system. The suggested FER approach significantly outperformed previous FER approaches, with a weighted average identification rate of 96.37% across four different datasets.

Kim et al.[21] proposed a new FER spatiotemporal representation learning feature that is substantial to variations in expression severity. The proposed method uses states of descriptive emotion (e.g., apex, offset, and onset expression) that is defined in face sequences irrespective of frequency of the expression. Two parts of this paper encoded the characteristics of facial expressions. Through a convolutional neural network (CNN), the spatial image feature of representative emotion-state frames was learned in part one. To enhance the emotion class separable power of the spatial representation of features. In the second section, facial emotion with a brief-term memory of the spatial feature representation, temporal characteristics that were studied during the first portion. The authors utilized “macro- and micro-expressions facial datasets (MMI)” dataset and also a randomized micro-expression dataset (CASME II) for testing. When compared with the state-of-the-art approaches, experimental result showed that this suggested methodology achieved a high rate of recognition in both datasets.

Wei et al.[22] suggested a technique for robust face identification based on low-resolution sensor noise depth data. Depth and color knowledge were acquired by the Kinect sensor, the facial-feature points feature vector was extracted using the face tracking SDK, and the random forest (RF) algorithm recognized a facial emotion. This technique enabled deployment of real-time, intelligent interactive facial expression recognition. Stratification sampling increased sample set representation and ensured overall data structural consistency. Even if there are not enough training data, the RF method can prevent overfitting.

Suk et al.[23] presented a real time FER mobile application that operates only on camera-based smartphone. A set of SVMs was taken in use by the proposed system for classifying a set of six basic and one neutral emotion with mouth status checking. Active shape model stripped away the features of facial expression for the identification of emotions, then produced dynamic features by displacing neural and expression features. In video samples (309 samples) from the extended Cohn–Kanade (CK+) dataset, the test results showed 10-fold cross-validation with 86% accuracy. The same SVM models were utilized in the smartphone app, which ran at 2.4 frames per second on a Samsung Galaxy S3. Across seven respondents, real-time mobile emotion efficiency was shown to be around 72% for six key plus one neutral emotion.

Lajevardi et al.[24] investigated a new method for identifying facial expressions focused on a facial hybrid region. The completely automated expression recognition system includes face detection, facial identification, feature extraction, appropriate feature aggregation, and classification. Using log Gabor filters, the features are retrieved from the entire face image and the facial regions (eyes and mouth). The most appropriate traits are then selected using mutual knowledge-based criteria. The system instantaneously recognizes six facial expressions: disgust, fear, happy, sad, and surprise. A naive Bayesian classifier had been utilized to define the selected features. The proposed solution was tested extensively by using Cohn–Kanade and Japanese Female Facial Expressions (JAFFE) datasets. The efficiency of the proposed HFR method in improving classification levels has also been reported in the investigations.

2.1.2 *Deep learning methods*

As a universal feed into machine learning, deep learning has grown, culminating in the state-of-the-art result with the availability of big data in several computer vision studies. Deep-learning FER approaches greatly lower reliance on face-to-face models as well as other pre-processing strategies facilitating “end-to-end” learning to take place directly from input images in the pipeline. Among various deep learning models known, the CNN is the most widely used network model.[1]

Hassouneh et al.[25] developed an algorithm for real-time emotion recognition using virtual markers through an optical flow algorithm that performs well in uneven lighting, subject head rotation, different backgrounds, and different skin tones in order to categorize the emotional expressions of physically disabled people (deaf, mute, and bedridden) and autistic children based on facial landmarks and electroencephalograph signals. Six face emotions are captured using 10 virtual markers (joy, sadness, anger, fear, disgust, and surprise). Before sending the features to the LSTM and CNN classifiers, the features are five times cross-verified.

Pranav et al.[26] proposed a two-layer model of a convolution network for recognizing facial emotions. This model uses the self-created image collection to classify five different facial expressions. The model has equivalent training and validation accuracy, indicating that it has a good fit to the data and can be generalized.

Minaee et al.[27] suggested a deep learning technique centered on an attentional convolutional network that can concentrate on important portions of the face and surpasses earlier models. The researchers also used a visualization method based on the output of the classifier to identify significant facial landmarks areas and identify emotional states. Using the findings of the experiment, the authors also illustrated how various emotions are sensitive to specific regions of the face.

Riyantoko et al.[28] utilized the Haar-cascade classifier and CNN to classify face emotion and categorized seven facial expressions. Centered on epoch varieties, the CNN model gained MSE and the value of accuracy also increases. The findings demonstrate that by increasing the epoch value, the MSE value decreased, and the accuracy value increased.

Hasani and Mahoor[29] introduced a three-dimensional (3D) CNN FER system in video format. The network design comprised of 3D Inception-ResNet layers accompanied via an LSTM system, which collectively removed the spatial relationships within face images and the temporal relationships throughout the video among different frames. Facial landmarks were used as network inputs that highlight the importance of facial parts as opposed to facial localities which may not contribute greatly to facial emotions. The proposed system was evaluated in subject-independent and cross-database activities. The system was tested utilizing four excellently known databases: CK+, MMI, FERA, and Denver Intensity of Spontaneous Facial Action Database (DISFA).

Mehendale,[30] proposed facial emotion recognition using convolutional neural networks (FERC) method. This method was used to detect new facial expression recognition system based on CNN. The foundation of the FERC is a two-part CNN: the first component focuses on background removal from the image, and the second part is concerned with extracting facial feature vectors. The five basic types of regular facial expression are identified by the FERC model's expressional vector.

Jung et al.[31] presented a method that uses deep learning, and also it was considered as a tool for automatic extraction of several features which were useful from raw data. Two different models were based on this deep network. The first deep network called deep temporal appearance network (DTAN) has extracted temporal outlook characteristics from object sequence, whereas other deep network called deep temporal geometry network extracts temporal geometry characteristics from temporal facial landmarks. The authors shown that the filters that were in learning the initial layer with the help of DTAN were having the capability of obtaining a great variation between input frames. In addition, the prominent landmarks identified by DTGN also have been shown. Using the unified deep network on databases namely CK+ and Oulu-CASIA, the highest recognition levels have been achieved. In addition, it was also shown that the combined fine-tuning approach was superior in comparison to other approaches of integration. One of them being weighted summation and other was concatenation system of features.

Ebrahimi et al.[32] suggested that one of the strongest signs of recognition of emotions is the spatio–temporal Darwinism of facial characteristics. In the proposed method, author presented an RNN application for modeling this spatio–temporal Darwinism through the accumulation of facial characteristics to execute video emotion recognition. The experiments show that this method surpasses all other modalities, and the average of classification is based on vision per frame. In addition, the author discussed two methods of fusion that work on the feature and on the level of the decision. The fusion network at the feature level integrates features from various procedures and obtains a higher validation accuracy in comparison to any single modality classifier. The experiments depict that feature level and fusion at the decision stage are additional and therefore achieve a higher classification accuracy when they are combined.

2.2 Based on the Types of Frame

(a) Static image frame

(b) Video-sequence image frame

2.2.1 *Static image-based approach*

For static-based approaches, only spatial information from the current single image encodes the representation of the function.[4]

Huang et al.[33] presented a unique approach toward FER focused on a minimal description of LBP characteristics. The JAFFE database was used in performing extensive experiments. Findings of the test indicate that this approach works better than using only facial recognition classification based on sparse representation, and it is better than standard algorithms, such as linear discriminant analysis (LDA) and PCA.

Wang et al.[34] proposed a new algorithm local step quantization (LPQ) based and fragmented FER representation. First, instead of using the sparse representation-based classification (SRC) approach, this represents the image of the testing emotion using a linear combination of the training expression images, and it uses the LPQ descriptor to extract features. The incomplete representation residual study recognizes facial emotions. The JAFFE dataset was tested on the proposed algorithm. The result indicates that the algorithm is far better in comparison to previous conventional approaches, such as LBP + SVM, 2DPCA + SVM, LDA + SVM. Compared with SRC algorithm, the efficiency is also obviously improved. Furthermore, the proposed algorithm's recognition level obtains the highest FER recognition rate when images are under occlusion.

Based on the compressive (CS) theory, Zhang et al.[35] suggested a system for successful FER. The creation of a sparse classifier representation was based on CS theory (SRC). Exploratory results on the Cohn–Kanade database of widely known facial expressions demonstrate that the SRC method achieved higher durability and efficiency in corruption, occlusion, and FER practices when compared with the nearest neighbor (NN), SVM, and the nearest subspace (NS).

Sun et al.[36] employed a multiple-channel deep neural network for recognizing face emotions in static images that adapt and combine spatial–temporal information. The primary idea behind this method is to use the emotional-gray-level faces picture as the spatial information, and to extract optical flow as the time information of a facial expression from changes between the peak expression face image and the neutral face image. A multi-channel deep spatial–temporal feature fusion neural network is described in order to do deep spatial–temporal feature extraction and fusion from static images.

2.2.2 *Video sequence-based approach*

In Ref. 37, angry, happy, sad, fear, disgust, and neutral are the six basic emotions that have been proposed as CNN-LSTM based neural networks. These networks have been trained on the CREMA-D dataset and assessed on the RAVDEES dataset. The faces in the movies have been altered using the Open Face software, which isolates the face from its environment and masks it and then placed into the CNN. The research concentrated on the application of LSTM networks, which can make use of a series of data to help with the final prediction of emotions in a clip.

Gupta et al.[38] developed a method for recognizing facial expressions that combines the spatial and temporal convolutional properties that are present throughout the video. In order to train an end-to-end FER system from video, the author used 15 spatial and 15 temporal streams or spatial correlating to the temporal flow. Through training, the representation of the video is enhanced, and the over-fitting issue brought on by a shortage of datasets is avoided. It is also found that the suggested methodology is the best for collecting spatial–temporal information from video datasets when compared with other methods that are already in use. The datasets used include RML, MMI, BAUM-1s, eNTERFACE05, and FER-2013.

Hajarolasvadi et al.[39] devised a method for spotting face expressions of emotion in video clips. The program is then assessed in scenarios that involve both people and without them. For video frames, 60 geometric features were calculated. The geometric traits are then clustered using k-means to find the k most discriminant frames for each sequence of video. Then, to determine the best representative k, various classifiers, such as linear SVM and Gaussian SVM, were employed. Finally, GoogleNet, AlexNet, ResNet-50, VGG-16, and VGG-19 were utilized to evaluate. In addition, depending on multiple facial landmark areas, the impact of geometric features on keyframe choice for person-dependent and person-independent situations is examined. The retrieved characteristics via CNNs were displayed using the t-distributed stochastic neighbor embedding method to examine the discriminative capacity under various conditions.

Kulkarni et al.[40] proposed an approach for facial emotion identification from continuing face images. For a more accurate representation, the approach isolates the discriminative component from prominent facial areas before combining it with surface and direction highlights. To gauge the degree of happiness and sadness, as well as to evaluate whether a person is indeed happy or sad or just acting that way, subcategories in main phrases such as joyful and sad were categorized. Furthermore, by selecting the profoundly discriminative highlights, the author reduces the information measurement.

3 Dataset Description

For comparative and comprehensive studies, numerous databases have been used in the FER area. Following are the brief description of popular datasets used in FER.

3.1 Extended Cohn–Kanade Dataset (CK+)

CK+ includes 593 sequences of videos that cover both posed as well as un-posed spontaneous emotions, as well as supplementary metadata as shown in Fig. 2. Ages ranges between 18 and 30 years with 123 subjects, who are mostly female. It is possible to evaluate image sequences both for action units and prototypical feelings. Offering guidelines and base tests that monitor facial expressions, AUs as well as identification of emotions. Pixel resolutions of the image 640×480 and 640×490 for gray-scale values with 8-bit precision.[41]

3.2 Denver Intensity of Spontaneous Facial Action Database

It consists of 27 adult subjects with distinct ethnicities (12 females and 15 males), 130,000 high-resolution stereo video frames (1024×768) are included. The AU intensities (0 to 5 scale) were rated manually across all video sequence utilizing two different human FACS experts. In addition to each image, the database contains 66 facial landmark points as shown in Fig. 3. Each facial image's original resolution is 1024×768 pixels.[42]

3.3 Compound Emotion

Compound emotion (CE) consists of 5060 images for its 230 human subjects, correlating to 22 simple and composite emotion groups (130 females and 100 males, average age 23).

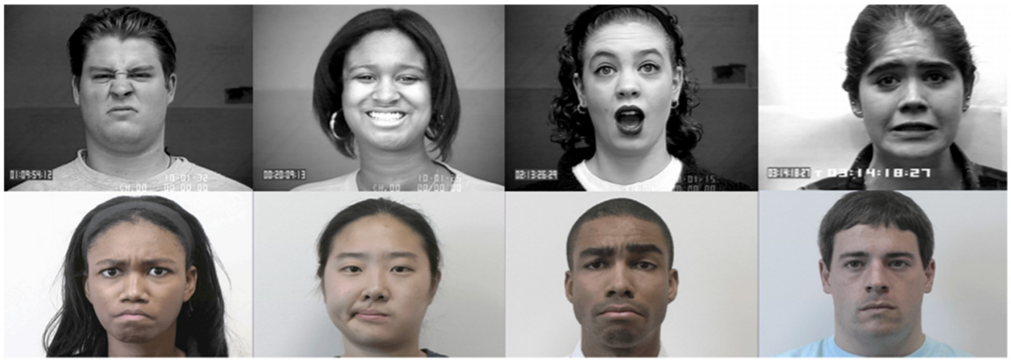

Fig. 2 Examples from CK+ database.[41] All top-leveled images are obtained from the original CK database and also the expanded data are symbolic of those under. Examples of the labels EMOTION and AU are: (i) disgust: AU 1 + 4 + 15 + 17, (ii) happy: AU 6 + 12 + 25, (iii) surprise: AU 1 + 2 + 5 + 25 + 27, (iv) fear: AU 1 + 4 + 7 + 20, (v) angry: AU 4 + 5 + 15 + 17, (vi) contempt: AU 14, (vii) sadness: AU 1 + 2 + 4 + 15 + 17, and (viii) neutral: AU0.

Fig. 3 Facial portraits of 25 of the 27 participants. Two participants did not agree to the publication use of their images.[42]

The majority of races and ethnicities are Caucasians, Asians, Africans, and Hispanics. There are reduced facial occlusions, with no glasses or facial hair. In order to fully expose their eyebrows, male subjects were asked to shave their faces as cleanly as possible and all participants also were requested to expose their foreheads. The pictures are color images as shown in Fig. 4, with a pixel resolution of 3000 × 4000 with a Canon IXUS.[43]

3.4 Binghamton University 3D Facial Expression

Since two-dimensional (2D) static images of the face were widely used throughout FER, Yin et al. created a database of annotated 3D facial expressions at Binghamton University, named Binghamton University 3D Facial Expression (BU-3DFE) 3D, for research in 3D human faces also facial expressions, as well as for the advancement of a generalized understanding of the human mind. Figures 5 and 6 show sample images of the BU-3DFE dataset. It consists of a

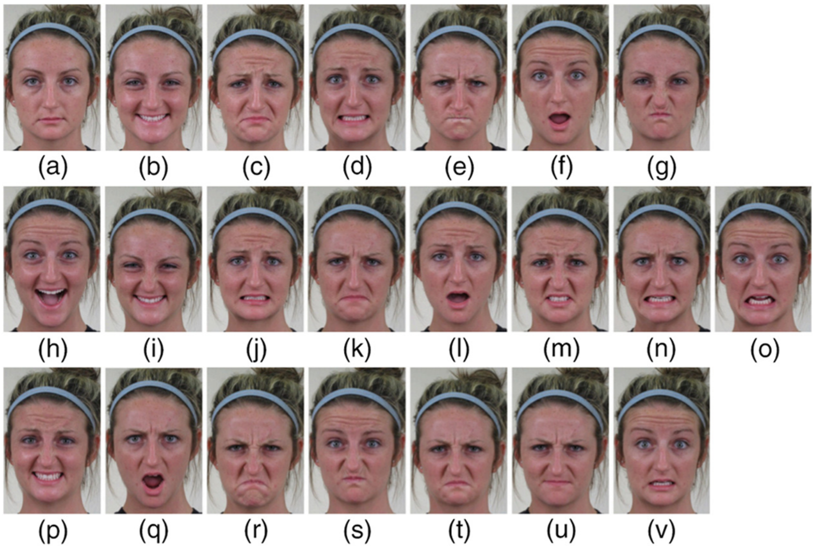

Fig. 4 Test images in database for 22 categories:[43] (a) neutral, (b) happy, (c) sad, (d) fearful, (e) angry, (f) surprised, (g) disgusted, (h) happily surprised, (i) happily disgusted, (j) sadly fearful, (k) sadly angry, (l) sadly surprised, (m) sadly disgusted, (n) fearfully angry, (o) fearfully surprised, (p) fearfully disgusted, (q) angrily surprised, (r) angrily disgusted, (s) disgustedly surprised, (t) appalled, (u) hatred, and (v) awed.

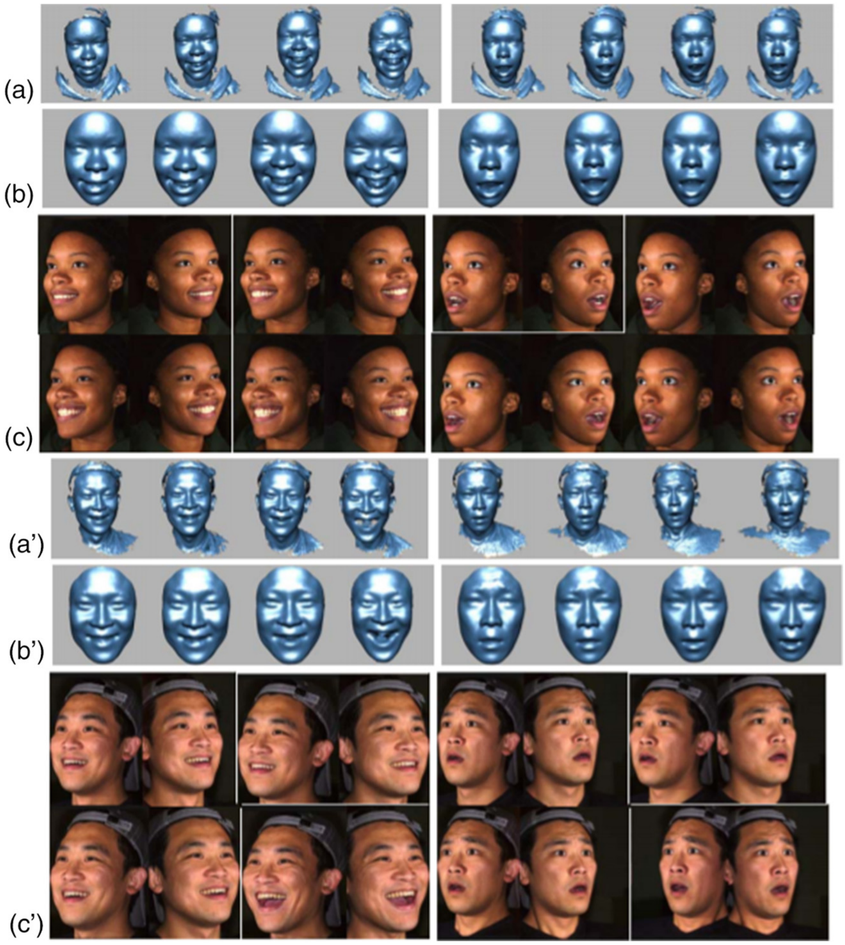

Fig. 5 Examples of expressions: the left four are gladness and the right four are astonishment, each with four degrees of intensity. (a), (a′) Raw models and (b), (b′) cropped facial region models. (c), (c′) Two textured views.[44]

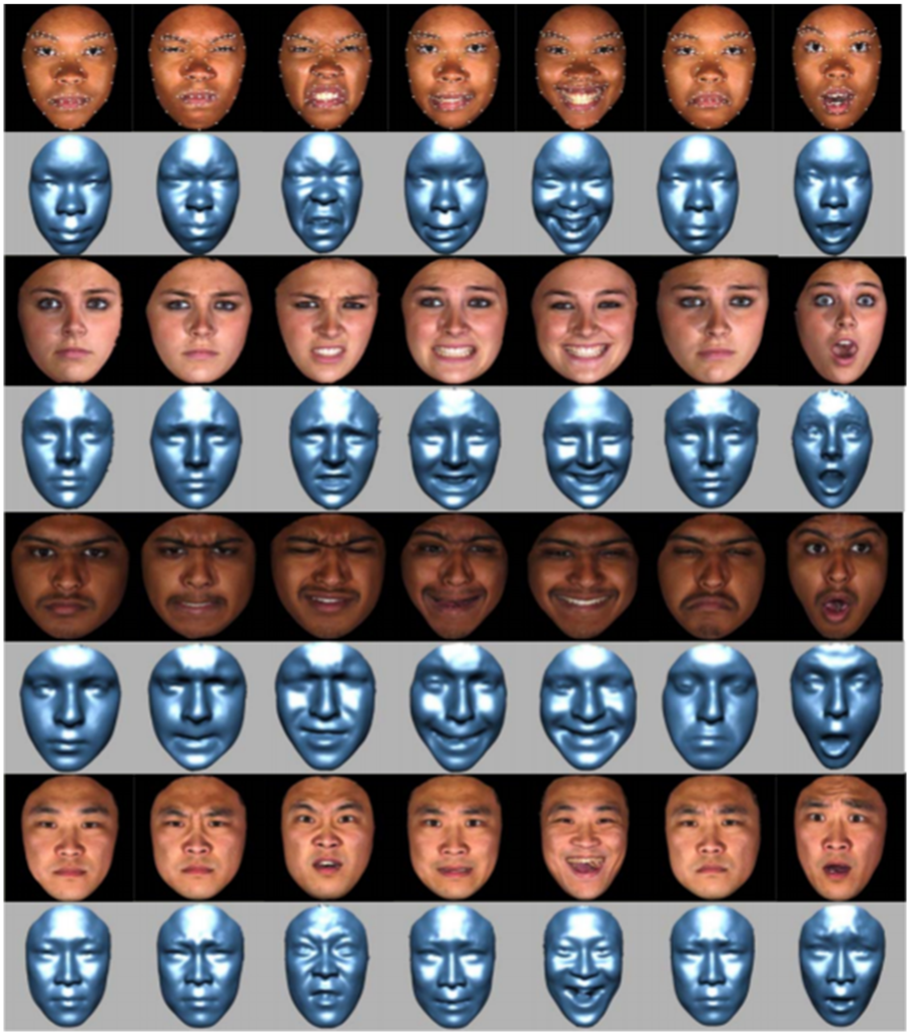

Fig. 6 In four sample individuals, seven emotions (neutral, angry, disgust, fear, happiness, sorrow, and surprise) are expressed. Face form is being designed, as well as frontal view texturing. The first row shows a selection of the selected feature points.[44]

total of 100 subjects displaying six emotions, 56 females and 44 males. The database contains 25 3D facial emotion models per subject, and a set of 83 manually annotated facial landmarks linked to each model. Each facial image has an original dimension of 1040 pixels × 1329 pixels.[44]

3.5 MMI

Over 2900 high-resolution still images and video sequences of 75 subjects make up the MMI. The existence of AUs (event coding) is fully transcribed throughout the video sequences and partially coded at the frame level, indicating where an AU would be at each frame's neutral, onset, apex, or offset point. There are 238 video sequences in total, with 28 male and female themes addressed. Figure 7 shows sample images of the MMI dataset where each image is 720 × 576 pixels in size.[45]

3.6 Japanese Female Facial Expressions

The JAFFE collection has 213 pictures of 10 different Japanese female models expressing seven various face expressions (six simple expressions and a single neutral emotion) as shown in Fig. 8. Each of the 60 images of Japanese persons is categorized using six emotive characteristics. Each facial image is 256 pixels by 256 pixels in its original size.[46]

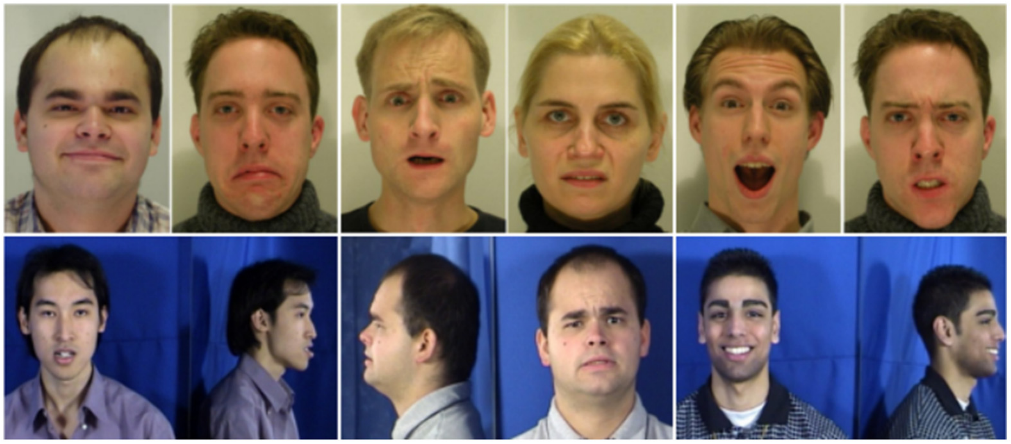

Fig. 7 Example of images from MMI database.[45]

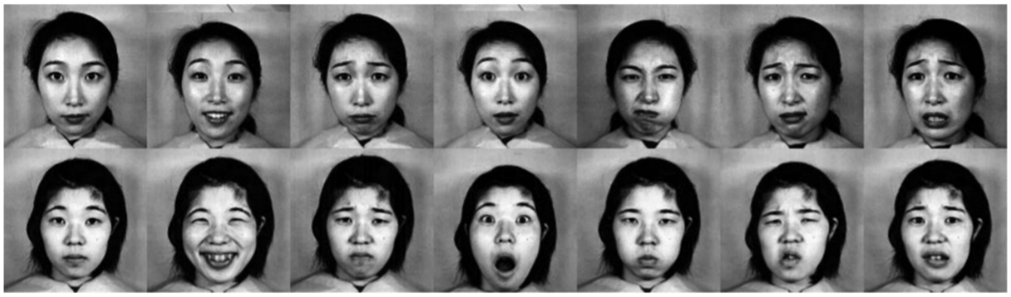

Fig. 8 Example of images from JAFFE database.[46]

3.7 Binghamton-Pittsburgh 3D Dynamic Spontaneous

Binghamton-Pittsburgh 3D dynamic spontaneous (BP4D-spontaneous) is a 3D video database that includes a large range of 41 young adults with spontaneous facial expressions (23 women and 18 men) as shown in Fig. 9. The subjects were aged 18 to 29. Six are African American, four are Hispanic, and 20 are Euro-Americans. In the 2D and 3D domains, the facial characteristics were tracked using both person-specific and generic approaches. The database encourages researchers to investigate 3D spatiotemporal aspects during minor facial gestures in order to gain a better understanding of the interaction between posture and motion dynamics in facial AUs, as well as natural face behavior. Each facial image has an original size of 1040 pixels × 1329 pixels.

3.8 RAF-DB Database

Around 30,000 facial photos of thousands of people are stored in the RAF-DB. The EM method was used to filter out inaccurate classifications after each shot was independently labeled roughly

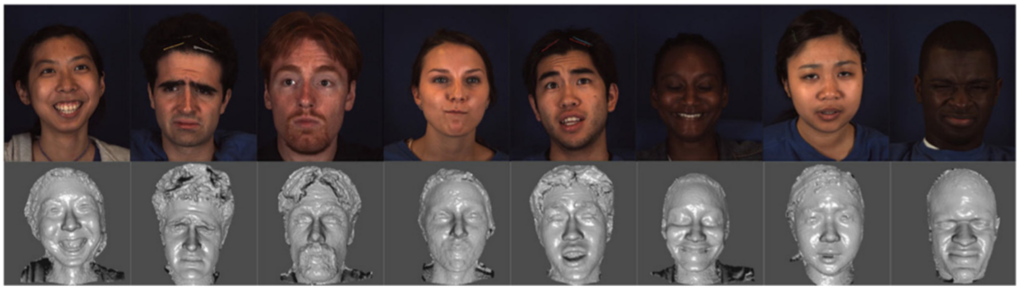

Fig. 9 2D and 3D examples of eight emotional expressions from BP4D-spontaneous database.

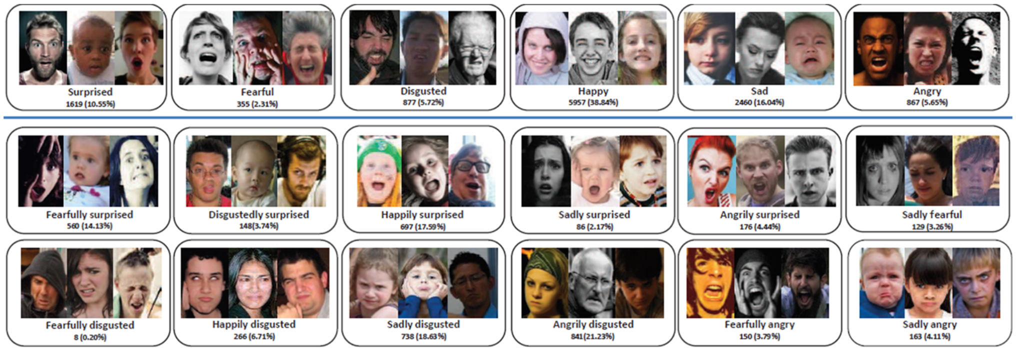

Fig. 10 Example: six-class basic emotions and all 12-class CE from RAF-DB.[47]

40 times. Crowd-sourcing demonstrates that various aspects of the real world frequently exhibit compound or mixed emotions. Figure 10 shows some sample images of RAF-DB dataset. It is the first wild database containing compound expressions. The cross-database research indicates that in the RAF-DB, the units of operation of fundamental emotions are much more vibrant than in laboratory-controlled ones or even deviate from them.[47]

4 Conclusion

This paper reviews the techniques that are used in recognizing facial emotion. The FER techniques are divided into two parts (categorization) based on the type of approach and the type of frame. The techniques based on the type of approach are further divided into conventional methods and deep learning methods. Table 1 describes the conventional methods and the types of datasets used in each paper. Table 2 describes the deep learning approaches and the types of datasets used. Table 3 compares the deep learning approaches based on different parameters on each of the datasets used. The techniques used in the type of frame are divided into parts: static image frame and video-sequence image frame. Table 4 briefly describes the different static

Table 1 Comparison of various conventional methods.

Author	Dataset	Accuracy (%)
Happy et al.[12]	Self-created	97
Fabian et al.[13]	DISFA, CFEE	—
Gunawan et al.[14]	User-generated	95.67
Szwoch et al.[15]	FEEDB	50
Ghimire et al.[16]	CK+	91.95
Shen et al.[17]	Normal facial expression visible and infrared (USTC-NVIE)	76.7
Khan et al.[18]	CK+, MMI, and FEED	96.7
Ghimire et al.[19]	CK+	95.17 (multi-class AdaBoost with DTW), 97.35 (SVM)
Siddiqi et al.[20]	CK+, JAFEE, extended Yale B face (B+), MMI	96 (CK+), 96 (B+), 96 (MMI), 96 (JAFEE)
Kim et al.[21]	MMI, CASME II	78.61 (MMI), 60.98 (CASME II)
Wei et al.[22]	Self-generated	68
Suk et al.[23]	CK+	72
Lajevardi et al.[24]	CK, JAFEE	91.8 (CK), 97.9 (JAFEE)

Table 2 Comparison of various FER methods based on deep learning.

Author	Dataset used	Training accuracy (%)	Testing accuracy (%)
Hassouneh et al.[25]	Self-created	—	87.25
Pranav et al.[26]	Self-created	—	78.04
Minaee et al.[27]	FER-2013, CK+, FERG, JAFFE	—	70.02 (FER-2013), 99.3 (FERG), 92.8 (JAFEE), 98 (CK+)
Riyantoko et al.[28]	FER-2013	—	92
Hasani et al.[29]	CK+, FERA, MMI, and DISFA	—	96 (CK+), 100 (MMI), 94 (FERA), 68 (DISFA)
Mehendale [30]	CK+, Caltech faces, CMU, NIST	—	p6 (NIST), 85 (Caltech faces), 78 (CMU)
Jung et al.[31]	CK+, Oulu-CASIA	—	97.25 (CK+), 81.46 (Oulu-CASIA), 70.24 (MMI)
Ebrahimi et al.[32]	Acted facial expressions in the wild 5.0	52.320	50.092
Suk et al.[23]	CK+	—	72
Lajevardi et al.[24]	CK, JAFEE	—	91.8 (CK), 97.9 (JAFEE)

Table 3 Static frame based method.

Reference	Dataset used	Training accuracy	Testing accuracy (%)
Huang et al.[33]	JAFFE	—	62.86
Wang et al.[34]	JAFFE	—	70
Zhang et al.[35]	CK	—	98.10
Sun et al.[36]	CK+, MMI, and RaFD	67.18 (CK+, RaFD), 86.80 (CK+, MMI), 75.13 (MMI, RaFD)	98.38 (CK+), 99.17 (RaFD), 99.59 (MMI)

Table 4 Video sequence-based approach.

Author	Dataset used	Training accuracy (%)	Testing accuracy (%)
Hans et al.[37]	CREMA-D, RAVDEES	94.3 (CREMA-D)	75.4 (CREMA-D), 63.35 (RAVDEES)
Gupta et al.[38]	RML, MMI, BAUM-1 s, FER-2013, eNTERFACE05	—	93.33 (BAUM-1 s), 95.42 (FER-2013), 96.25 (MMI), 90.41 (RML), 97.08 (eNTERFACE05)
Hajarolasvadi et al.[39]	SAVEE, RML	—	99.89 (SAVEE), 98.23 (RML)
Kulkarni et al.[40]	MMI, JAFEE	79	90

image frame techniques along with the dataset. The paper also reviews the different datasets that are widely used in the recognition of facial emotion.

Acknowledgments

The authors declare that they have no conflict of interest. Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

References

1. B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors* **18**(2), 401 (2018).
2. A. Mehrabian, "Communication without words," in *Communication Theory*, pp. 193–200, Routledge (2017).
3. K. Kaulard et al., "The MPI facial expression database a validated database of emotional and conversational facial expressions," *PLoS One* **7**(3), e32321 (2012).
4. S. Li and W. Deng, "Deep facial expression recognition: a survey," *IEEE Trans. Affect. Comput.* **13**, 1195–1215 (2020).
5. P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Person. Soc. Psychol.* **17**(2), 124 (1971).
6. P. Ekman, "Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique," *Psychol. Bull.* **115**, 268–287 (1994).
7. Z. Zeng et al., "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(1), 39–58 (2008).
8. E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: a survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(6), 1113–1133 (2014).
9. B. Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," in *Advances in Face Detection and Facial Image Analysis*, M. Kawulok, M. E. Celebi, and B. Smolka, Eds., pp. 63–100, Springer (2016).
10. P. Ekman, *Facial Action Coding System (FACS)*, A Human Face, Salt Lake City (2002).
11. H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: current trends and future directions," *Image Vision Comput.* **31**(2), 120–136 (2013).
12. S. Happy, A. George, and A. Routray, "A real time facial expression classification system using local binary patterns," in *4th Int. Conf. Intell. Hum. Comput. Interact. (IHCI)*, IEEE, pp. 1–5 (2012).
13. C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 5562–5570 (2016).
14. A. A. Gunawan et al., "Face expression detection on kinect using active appearance model and fuzzy logic," *Procedia Comput. Sci.* **59**, 268–274 (2015).
15. M. Szwoch and P. Pieniążek, "Facial emotion recognition using depth data," in *8th Int. Conf. Hum. Syst. Interact. (HSI)*, IEEE, pp. 271–277 (2015).
16. D. Ghimire et al., "Facial expression recognition based on local region specific features and support vector machines," *Multimedia Tools Appl.* **76**(6), 7803–7821 (2017).
17. P. Shen, S. Wang, and Z. Liu, "Facial expression recognition from infrared thermal videos," in *Intelligent Autonomous Systems 12*, pp. 323–333, D. K. Pratihar and L. C. Jain, Eds., Springer (2013).
18. R. A. Khan et al., "Framework for reliable, real-time facial expression recognition for low resolution images," *Pattern Recognit. Lett.* **34**(10), 1159–1168 (2013).
19. D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors* **13**(6), 7714–7734 (2013).
20. M. H. Siddiqi et al., "Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields," *IEEE Trans. Image Process.* **24**(4), 1386–1398 (2015).
21. D. H. Kim et al., "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).
22. W. Wei, Q. Jia, and G. Chen, "Real-time facial expression recognition for affective computing based on kinect," in *IEEE 11th Conf. Ind. Electron. and Appl. (ICIEA)*, IEEE, pp. 161–165 (2016).
23. M. Suk and B. Prabhakaran, "Real-time mobile facial expression recognition system-a case study," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. Workshops*, pp. 132–137 (2014).
24. S. M. Lajevardi and Z. M. Hussain, "Feature extraction for facial expression recognition based on hybrid face regions," *Adv. Electr. Comput. Eng.* **9**(3), 63–67 (2009).
25. A. Hassouneh, A. Mutawa, and M. Murugappan, "Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods," *Inf. Med. Unlocked* **20**, 100372 (2020).
26. E. Pranav et al., "Facial emotion recognition using deep convolutional neural network," in *6th Int. Conf. Adv. Comput. and Commun. Syst. (ICACCS)*, IEEE, pp. 317–320 (2020).
27. S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: facial expression recognition using attentional convolutional network," *Sensors* **21**(9), 3046 (2021).
28. P. Riyantoko et al., "Facial emotion detection using Haar-cascade classifier and convolutional neural networks," *J. Phys. Conf. Ser.* **1844**(1), 012004 (2021).
29. B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. Workshops*, pp. 30–40 (2017).

30. N. Mehendale, “Facial emotion recognition using convolutional neural networks (FERC),” *SN Appl. Sci.* **2**(3), 1–8 (2020).
31. H. Jung et al., “Joint fine-tuning in deep neural networks for facial expression recognition,” in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2983–2991 (2015).
32. S. Ebrahimi Kahou et al., “Recurrent neural networks for emotion recognition in video,” in *Proc. 2015 ACM Int. Conf. Multimodal Interact.*, pp. 467–474 (2015).
33. M.-W. Huang, Z.-W. Wang, and Z.-L. Ying, “A new method for facial expression recognition based on sparse representation plus LBP,” in *3rd Int. Congr. Image and Signal Process.*, IEEE, Vol. 4, pp. 1750–1754 (2010).
34. Z. Wang and Z. Ying, “Facial expression recognition based on local phase quantization and sparse representation,” in *8th Int. Conf. Natural Computat.*, IEEE, pp. 222–225 (2012).
35. S. Zhang, X. Zhao, and B. Lei, “Robust facial expression recognition via compressive sensing,” *Sensors* **12**(3), 3747–3761 (2012).
36. N. Sun et al., “Deep spatial-temporal feature fusion for facial expression recognition in static images,” *Pattern Recognit. Lett.* **119**, 49–61 (2019).
37. A. S. A. Hans and S. Rao, “A CNN-LSTM based deep neural networks for facial emotion detection in videos,” *Int. J. Adv. Signal Image Sci.* **7**(1), 11–20 (2021).
38. R. Gupta and L. Vishwamitra, “Facial expression recognition from videos using CNN and feature aggregation,” in *Mater. Today: Proc.* (2021).
39. N. Hajarolasvadi, E. Bashirov, and H. Demirel, “Video-based person-dependent and person-independent facial emotion recognition,” *Signal, Image Video Process.* **15**(5), 1049–1056 (2021).
40. P. Kulkarni and T. Rajesh, “Video based sub-categorized facial emotion detection using LBP and edge computing,” *Rev. d’Intell. Artif.* **35**(1), 55–61 (2021).
41. P. Lucey et al., “The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression,” in *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit.-Workshops*, IEEE, pp. 94–101 (2010).
42. S. M. Mavadati et al., “DISFA: a spontaneous facial action intensity database,” *IEEE Trans. Affect. Comput.* **4**(2), 151–160 (2013).
43. S. Du, Y. Tao, and A. M. Martinez, “Compound facial expressions of emotion,” *Proc. Natl. Acad. Sci. U. S. A.* **111**(15), E1454–E1462 (2014).
44. L. Yin et al., “A 3D facial expression database for facial behavior research,” in *7th Int. Conf. Autom. Face and Gesture Recognit. (FGR06)*, pp. 211–216 (2006).
45. M. Valstar et al., “Induced disgust, happiness and surprise: an addition to the mmi facial expression database,” in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, Paris, France, Vol. **10**, p. 65 (2010).
46. M. Lyons et al., “Coding facial expressions with Gabor wavelets,” in *Proc., Third IEEE Int. Conf. Autom. Face and Gesture Recognit.*, IEEE Computer Society, Nara, Japan, pp. 200–205 (1998). A database of facial expressions collected for the research accompanies the article and is available for use by other researchers. JAFFE facial expression image database: https://doi.org/10.5281/zenodo.3451524.
47. S. Li, W. Deng, and J. Du, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 2584–2593 (2017).

Rishabh Vats is an assistant professor in the Department of Computer Science and Engineering, GLA University, Mathura, India. He has completed his MTech from GLA University, India, in the Department of Computer Science and Engineering. His current research interests include computer vision, facial image recognition, medical image processing, and artificial intelligence. He has published many academic papers (SCIE indexed) as well as conferences.

Manoj Kumar is a professor at GLA University received a PhD in computer engineering and applications from GLA University, Mathura, India, in 2016. His total teaching experience is more than 20 years. Currently, he is professor at GLA University. His research interest areas are computer vision, image processing, medical imaging, and cloud computing. He has published more than 50 papers in various journals and conferences. Also, he has published 2 books as well as 9 patents and 3 patents have been granted. Currently, he is guiding 8 PhD students. Also, running 3 different government projects. He is also a active reviewer in various SCI indexed journals. He has served as a reviewer in various national and international conferences.

Ritu Rai is an assistant professor in the Department of Computer Science and Engineering, GLA University, Mathura, India. She has completed her MTech from GLA University, India, in the Department of Computer Science and Engineering. Her current research interests include image recognition, medical image processing, and artificial intelligence. She has published many academic papers (SCIE indexed) as well as conferences.

Gunjan Verma is an assistant professor at the School of Computer Applications, Manav Rachna International Institute of Research and Studies (MRIIRS), Faridabad, India. She is pursuing a PhD in Computer Science and Technology from GLA University, Mathura, India. Before joining MRIIRS, she was a researcher in Computer Vision and Image Processing lab, at GLA University, India from 2019–2022. Her current research interests include computer vision, object detection, multimedia information processing, underwater image enhancement and interpretation, artificial intelligence, nature-inspired algorithms, and deep learning. She has published many academic papers in international journals (SCI indexed) as well as conferences. She has served as a reviewer in various conferences, including ISCON-2019, ISCON-2020, FSAET-2021 and FSAET-2022 as well as a reviewer in renowned journals.