RESEARCH ARTICLE

Real-Time Facial Expression Recognition Based on Image Processing in Virtual Reality

Qingzhen Gong[1] · Xuefang Liu[2] · Yongqiang Ma[3]

Received: 20 September 2024 / Accepted: 21 December 2024

Abstract
More virtual reality (VR) scenarios have become more prevalent in recent years. More and more people are getting into VR, meaning that objective physiological measures to assess a user's emotional state automatically are becoming more critical. Individuals' emotional states impact their behaviour, opinions, emotions, and decisions. They may be used to analyze VR experiences and make systems react to and engage with the user's emotions. VR environments require users to wear head-mounted displays (HMDs), blocking off their upper faces. That makes traditional Facial Expression Recognition (FER) approaches very limited in their usefulness. Thus, a Deep Learning (DL) solution combined with image processing is utilized to classify universal emotions: sadness, happiness, disgust, anger, fear and surprise. Hence, this paper suggests the Deep Automatic Facial Expression Recognition Model (DAFERM) for interactive virtual reality (VR) applications such as intelligent education, social networks, and virtual training. Two main parts comprise the system: one that uses deep neural networks (DNNs) for facial emotion identification and another that automatically tracks and segments faces. The system begins by following a marker on the front of the head-mounted display (HMD). With the help of the spatial data that has been retrieved, the positions and rotations of the face are estimated to segment the mouth. Finally, the system interacts with DNN using the pixels processed by the lips. It obtains the facial expression results in real time using an adaptive method for histogram-based mouth segmentation.

Keywords Facial expression recognition · Deep learning · Virtual reality · Image processing · Deep neural network

1 Introduction

Emotions influence psychological and physiological well-being and their involvement in perception, learning, and rational decision-making [1]. Therefore, research into human behaviour must prioritize studying emotions and identifying their expressions [2]. Elicitation technologies, including images, sounds, videos, and, more recently, virtual reality (VR), have made it possible to research human emotions in controlled laboratory environments [3]. In the last few years, virtual reality (VR) has grown in academic and business settings. Gaming, education, training, health, and marketing are some of its general uses [4]. This growth is because a new wave of affordable headsets has been introduced, making the buying of head-mounted displays (HMDs) more accessible on an international level [5]. One of the most common ways people convey and amplify nonverbal information in everyday interactions is through facial expressions [6]. Due to its wide range of potential uses in fields as diverse as healthcare, social marketing, interactive game design, driver fatigue monitoring,

 https://doi.org/10.1007/s44196-024-00729-9

Published online: 20 January 2025

emotionally sensitive robots, and machine learning, the study of human facial expressions has been gaining popularity in recent years [7].

Nevertheless, it remains challenging to implement automated Facial Expression Recognition (FER) in an unconstrained environment. Partial obstruction in the face is one of the main challenges to accurate FER outside of laboratory environments. More recently, innovations in technology, notably VR, have changed how individuals connect and their environment [8]. However, a head-mounted display covers much of the face in VR, so gamers cannot display their emotions. Therefore, it is critical to identify and represent these expressions for VR systems to provide deep social interaction [9]. To be more specific, environmental simulations are models of real-life circumstances that researchers may use to learn how people respond to widely used notions. When the event they portray is immaterial, they become crucial [10]. VR allows them to explore these events in a controlled lab environment.

Additionally, virtual reality makes it possible to isolate and change elements with minimal cost and effort compared to real life. This comprises three distinct characteristics: formats, display devices, and user interfaces [11]. The hardware's power limits the realism of realistic 3D environments, which takes more time than developing 360◦ computer-generated images [12]. Nonetheless, 3D environments are becoming more and more potent as the processing power of GPUs (graphics processing units) increases annually [13]. Additionally, a crucial component of VR applications is the interactivity of the 3D environment, which enables the recreation of real-world activities [14].

The fast advancement of machine learning in images in the last several years has led to the widespread use of deep learning in FER, effectively displacing the more conventional, manually crafted feature techniques [15]. Angry, contemptuous, disgusted, fearful, happy, sad, and surprised are the seven primary forms of facial emotions. Convolutional Neural Networks (CNNs) and other deep learning algorithms have lately surpassed their statistical equivalents. These techniques have enhanced FER performance [16]. However, it is not easy to alter many system environments, and a large amount of training data is needed to guarantee appropriate feature learning when training a deep architecture from scratch [17]. In addition, it requires costly computing power. In virtual reality (VR) environments, where the upper half of the face is entirely hidden, current CNN models perform poorly because of substantial systematic occlusion [18]. Virtual reality (VR) to detect face expression recognition (FER) is designed to use a deep neural network (DNN) with a multi-region model with high accuracy [19].

The paper's novelty is: This study stands out because it addresses HMD occlusion issues. This goal is achieved by incorporating deep learning-based facial expression detection into virtual reality. The proposed system classifies sentiments in real time using adaptive eye-tracking and histogram-based mouth segmentation. Even if just part of the face is visible, this situation occurs. The study also advises employing Gaussian normalization to improve picture processing quality and reduce illumination parameter fluctuation. This system's ability to modify virtual reality interactions based on user emotions enhances the creation of realistic and emotionally engaging virtual reality settings. This characteristic is advantageous in virtual reality.

The significant contribution of the article is

- Designing the Deep Automatic Facial Expression Recognition Model (DAFERM) for interactive virtual reality (VR) applications.
- The proposed technique uses adaptive eye-tracking and facial expression recognition to avoid virtual reality HMD occlusion concerns and simplify emotional recognition.
- A Gaussian normalization that manages illumination variation in facial expression photos improves immersion recognition performance.
- Developed and published a real-time system that detects a user's mood and adapts their VR experience to make them happier and more engaged.
- Evaluating the mathematical model of DNN for detecting and classifying the user's facial expressions such as happy, sad, surprised, fearful, angry, and disgusted.

- The numerical findings have been employed, and the recommended DAFERM model increases the classification accuracy, user performance in the VR environment, user satisfaction, and face landmark prediction ratio compared to existing techniques.

2 Literature Study

2.1 Feature Clustering and Attention-Based Networks for FER

Zhengyao Wen et al. [20] proposed Feature Clustering Networks (FCN), Multi-head Attention Networks (MAN), and Attention Fusion Networks (AFN) for Facial Expression Recognition. FCN maximizes class separability using a large-margin learning goal to extract robust features. Furthermore, MAN instantiates numerous attention heads to simultaneously create attention maps on several parts of the face. Shervin Minaee et al. [21] suggested the Attentional Convolutional Network (ACN) for Facial Expression Recognition. This provides NN with less than ten layers to compete with and even surpass deeper networks regarding emotion identification. Outcomes from the author's complete experimental research of the work utilizing 4 extensively utilized FER datasets were reassuring.

2.2 Traditional and Hybrid Feature-Based Techniques

For the analysis of 2D-human face recognition and face feature regions, Surbhi Gupta et al. [22] suggested using scale-invariant feature transform (SIFT) and speeded-up robust features (SURF). This study presented a reliable and efficient face identification system that integrates the SURF and SIFT techniques for feature extraction. The suggested research mainly focuses on face detection and its accuracy with a low false-positive rate.

2.3 Cross-Connected Convolutional Networks for FER

CUIPING SHI et al. [23] discussed the Multi-branch Cross-Connection Convolutional Neural Networks (MBCC-CNN) for FERs. An MBCC-CNN model differs from single-structure convolutional neural networks, combining the residual connections, Network in Networks, and tree structure techniques in its construction. Results from experiments using the CK + , Fer2013, RAF, and FER + datasets demonstrate that the suggested MBCC-CNN technique achieves recognition rates of 71.52%, 98.48%, 88.10%, and 87.34%, respectively.

2.4 Feature Selection and Classifier Performance

Maiwan B. Abdulrazaq et al. [24] deliberated the Relief-F Feature Selection for Facial Expression Recognition using Supervised Classifiers' Performance. Research using the CK + dataset for facial expression recognition shows that KNN is the most precise classifier, with a 94.93% accuracy ratio. With a 93.95% accuracy ratio, RF is the classifier most similar to KNN. Of all the classifiers, J48 has an average accuracy ratio of 92.27%. The final three classifiers on the list are Support Vector Machine (SVM), 89.43%, 89.65%, and Multi-layer Perceptron (MLP). Zhao et al. [25] introduced the classification methods' performance assessment for FER. This research aims to assess the efficacy of supervised classifiers for face expression identification using chi-square-based minimal feature selection. With the maximum number of features used, the six classifiers SVM, multi-layer perceptron, K-Nearest neighbour, random forest, decision tree, and radial-based function are evaluated using top-ranking features. The dataset used for the study is CK + . The most accurate classifier is shown to be a random forest, which had a total accuracy ratio of 94.23%.

2.5 Adaptive Techniques for FER

Durga Ganga Rao Kola and Srinivas Kumar Samayamantula [26] presented the Local Binary Pattern with an Adaptive Window (LBP-AW) for Facial Expression Recognition. This method makes the feature vector shorter and more noise-resistant. Support Vector Machine (SVM) is being explored in terms of categorization. The suggested algorithm's performance is evaluated using the confusion matrix and recognition rate.

Chang Liu et al. [27] offered the Patch attention convolutional vision transformer (PACVT) for facial expression recognition with occlusion. Facial feature maps are extracted using a backbone convolutional neural network. To detect expressions, the Patch Attention Unit (PAU) adaptively determines the attention weights of local cues at the patch level to observe obstructed areas. After the face patches are transformed into a series of visual tokens, the Vision Transformer (ViT) records the global correlations and interactions between these tokens.

2.6 EEG and VR-Based Approaches for Emotion Recognition

The Differential Entropy for Automated Emotion Recognition in a Virtual Reality Environment with EEG Signals was proposed by Hakan Uyanık et al. [28]. This paper presents a new method for emotion detection and EEG-based facial expression recognition using the open-source dataset VREED. Two emotional states were classified using positive and negative differential entropy (DE) components, which were collected in four separate wavebands: alpha (8–13 Hz), theta (4–8 Hz), gamma (30–49 Hz), and beta (13–30 Hz). Emotion recognition using the VR-Personalized Exergames Platform (VR-PEER) was suggested by Yousra Izountar et al. [29]. A VR-PEER adaptive exergame system was created based on emotion recognition, as described in this article. Fifteen volunteers participated in the controlled trial, and all found the suggested approach helpful for their motor recovery.

3 Deep Automatic Facial Expression Recognition Model (DAFERM)

The use of facial expressions as a means of interaction between users and virtual reality devices or applications has become more common in recent years. Emotions critically impact our daily lives, so understanding and recognizing emotional responses is vital for understanding human behaviour. Research on emotion detection has relied chiefly on non-immersive 2D videos or images to induce different emotional states. The premise underlying VR is that technology may be a great tool for reproducing complex environments in real life, giving researchers new ways to study human behaviour in meticulously controlled environments. On the other hand, immersive virtual reality is rapidly gaining interest in emotion research due to its ability to replicate actual environments in a controlled laboratory environment while increasing the user's perception of their presence and activity. Its combination with implicit face expression measures and machine-learning approaches might have far-reaching effects across several fields of study, providing researchers with exciting new avenues to explore VR environments. As a result of its potential therapeutic uses, surveillance video applications, and other uses in healthcare and virtual reality systems, facial expression detection has garnered a lot of attention from users. An important method of interaction for mobile VR systems that rely on head-mounted displays is the user's expression. Thus, a rapid and precise face expression detection system is required for mobile VR systems. Consequently, to improve VR's functionality and user experience on mobility, this study creates a suite of algorithms that concentrate on wearable VR headsets and identify facial expressions.

3.1 Problem Statement

Due to factors such as dynamic lighting, head movements, and captivating settings, traditional face emotion recognition algorithms designed for still images may not always work in VR environments. Engaging users in virtual

reality requires access to real-time processing. Current facial emotion recognition algorithms could be inadequate regarding low-latency scenarios and realistic virtual reality experiences. Facial expression detection is challenging to implement with high accuracy and resilience because of user demographic diversity, occlusions, and individual facial variances. The difficulty of facial expression recognition is increased when additional modalities, such as voice or gesture detection, are included. It is currently challenging to combine and understand multimodal data in a way that offers a thorough user experience. While interacting with others in virtual reality settings, conveying emotions and other signs of engagement may be challenging because most VR helmets obscure a large portion of the face. Hence, this study recommends the Deep Automatic Facial Expression Recognition Model (DAFERM) for interactive virtual reality (VR) applications.

Figure 1 shows the VR environment with an eye-tracking system. This head-mounted virtual reality headset records the user's eye movements while engaging with virtual reality content. The procedure for eye calibration was provided before the experiment began. Following eye calibration, the researcher instructed the participant about the experiment before streaming the 360-degree VR videos and activating the eye-tracking devices. The user's raw eye data is synced with the real stimuli when the experiments are delivered simply by choosing the experiments' beginning and ending time frames. While wearing the HMD, the integrated eye tracking sensor identifies the user's stare and captures their gaze data. The information about the user's gaze is sent to the HMD via the eye-tracking sensor, and then the application receives it from the HMD. The application detects what the user is gazing at by comparing the data sent by the head-mounted display (HMD) with the image that is being created at the time. The application learns the object's properties and sets up its menus accordingly. The software then sends the combined menu and VR image to the head-mounted display. As soon as an image is received, the HMD shows it. From the available controllers, the user selects the one he chooses. Incorporating eye-tracking technology into virtual reality requires both software and hardware components. In real-time, researchers may capture the user's gaze location, duration, and pupil diameter using eye-tracking sensors. Using biometrics, one may study emotional states. It is possible to study internal processes by monitoring changes in pupil size, as dilated pupils represent psychological emotion. As the size of a person's pupils varies in proportion to their feelings and the stimuli they perceive, it may be used to gauge their psychological state. A positive connection exists between the degree of pupil dilation and the level of stimulus and emotions.

Figure 2 shows the proposed DAFERM model. The data are taken from the Facial Expression Recognition Kaggle Dataset [30]. With a fixed external camera and an HMD in a VR environment, this research thoroughly examines a more difficult situation. It is common practice for FERs to have three stages: feature extraction, face detection, and expression classification. Data pre-processing is carried out to improve or extract more useful image features and decrease unnecessary information for recognition. Image cropping and sequence sampling include processing both actual and virtual databases. Dark spots for cropping appear in both real and virtual images due to a lack of light in the areas around the eyes. The background, garments, hair, and head-mounted display (HMD) comprise the outside region. When cropping images that had nothing to do with

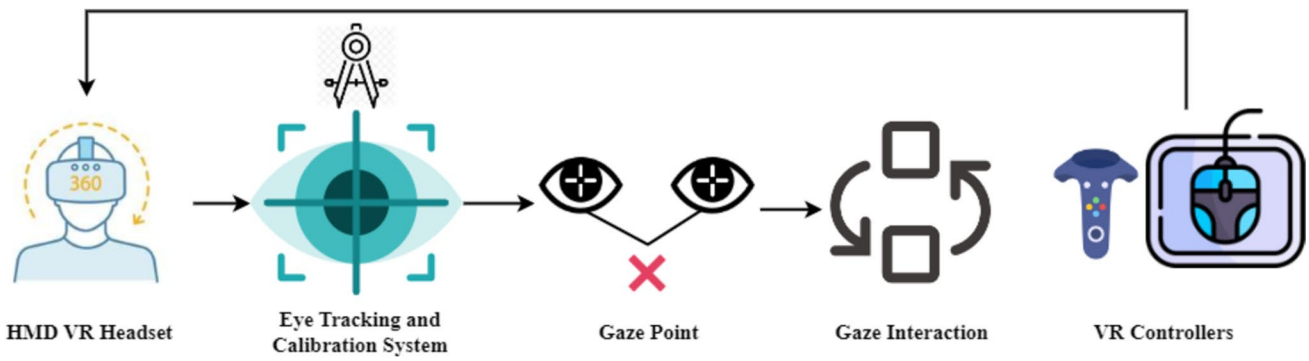


Fig. 1 VR Environment with Eye Tracking System

 https://doi.org/10.1007/s44196-024-00729-9

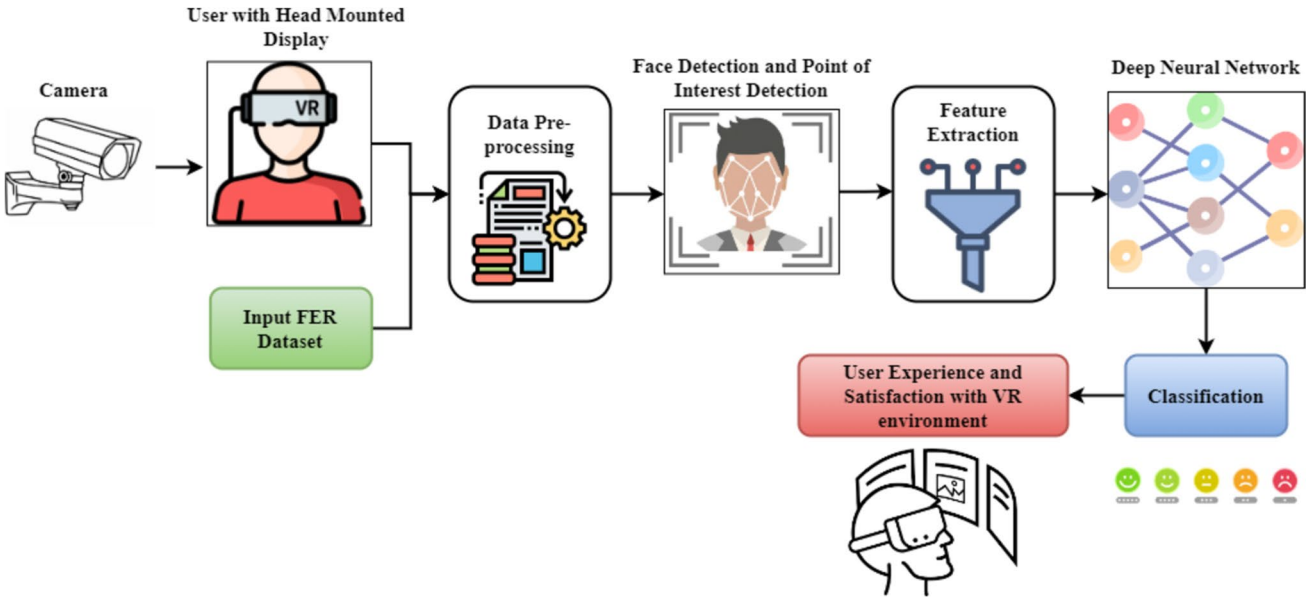


Fig. 2 Proposed DAFERM model

the facial expression movement, landmark locations have been employed as an indication. To get the virtual image landmarks, the scene capture component retrieved them. Next, it needs to find some landmarks in the facial image. Some regions that serve as landmarks are known as points of interest (POIs), while others are termed Region of Interest (ROIs).

A Haar-like feature extraction approach is suitable for face detection when working with a single target object. This study provides a DNN architecture that can learn dynamic and static features from face images and extract high-level dynamic features from optical flow series that detect the movements of facial muscles. To develop a reliable system for identification and classification using DNN, the idea of expression poses or light-invariant face recognition has been presented. To begin the learning process, this study uploads images of faces into a DNN and then classifies each expression as easy, neutral, or challenging. Raising the number of epochs, or learning iterations in a DNN model, improves classification performance (both in learning and validation accuracy). The seven emotions shown by the face were analyzed in this study: fear, anger, disgust, neutral, sadness, happiness, and surprise. Virtual reality places a premium on consumer satisfaction and experience. Emotions and six-degrees-of-freedom (6DoF) data may reveal how a user experiences things, which can reveal their body language.

In the image's interclass feature, the illumination fluctuation is reflected. Consequently, this work employs an image normalization approach to reduce the difference in the interclass variable. Impact offsets describe the difference between the image's interclass features. It has been decided to adopt Gaussian normalization since the intensity offset in the immediate area is uniform. It is possible to calculate the normalized image by using an appropriate formula.

$$\Psi(\eta,\omega) = \frac{\Omega(\eta,\omega) - \rho(\eta,\omega)}{\mu(\eta,\omega)} \tag{1}$$

As shown in Eq. (1), where $\Psi(\eta,\omega)$ denotes the input images, $\Omega(\eta,\omega)$ indicates the normalized output images, where η and ω are the number of rows and columns in the processed images. ρ denotes the local mean, and μ indicates the standard deviation calculated for face images. When the same person shows an emotion differently, the normalized image may help identify it more easily by overcoming the effects of illumination variance. In this research, standardizing the input to the suggested model is achieved via pre-processing

methods. The spatial dimensionality of the images is reduced in the suggested model by using average pooling. Batch normalization, which is performed during pre-processing, speeds up the training process, producing regularization and decreasing the number of errors related to generalization.

3.2 Deep Neural Network

Facial expression recognition requires face identification as a pre-processing step, even though a DNN may benefit from a whole set of video frames (including backgrounds) for feature extraction. This filtering out of irrelevant background data allows the DNN to zero in on the human face area, which is recognized to contain the most important emotional expressions in the video frames. A face selection is necessary if two or more human faces are in a single video frame, making distinguishing between the main expressions difficult. Assuming that each video only has one human character's emotional expression is the basis around which our method is created. Consequently, our system for detecting faces is composed of two parts: (i) a face detector that uses the small face detector to find every face in every frame of video, and (ii) a clustering method proposed for selecting the most prominent set of faces since videos often contain other facial regions that lack expressions. This study presumed that $F_j = \left\{\left(y_{j,h}, x_{j,h}, s_{j,h}, g_{j,h}\right) \in \mathbb{N}^4\right\}$ is a set comprising face data of the *jth* frame in videos, where $h, y_{j,h}, x_{j,h}, s_{j,h}$ and $g_{j,h}$ describe the face cluster, face centre coordinate, and face size. Firstly, the set F_j should be empty, and h is equivalent to 0 as no face group is identified. If $f_i = (y_i, x_i, s_i, g_i)$ is an identified face area in the *ith* frames, from (1), this study can measure the variance between this area and the latest identified object from every face cluster. Differences in the size and position of the face define this metric such that,

$$d_i(h) = \frac{1}{2}\left\|\left(y_i - y_{i-1,h}, x_i - x_{i-1,h}\right)\right\|_2 + \frac{1}{2}\left(\left|s_i - s_{i-1,h}\right| + \left|g_i - g_{i-1,h}\right|\right) \tag{2}$$

To decrease the sensitivity during clustering, this study compared this measurement with thresholds t; consequently, this study allocated the group indices h_i for f_i based on the subsequent expression,

$$h_i = \begin{cases} h' if d_i\left(h'\right) < t \\ \max\{h\} + 1 otherwise \end{cases} \tag{3}$$

As shown in Eq. (2), where $h' = \arg\max_h d_i(h)$. In the experiment, t is equivalent to 50. To choose the most important faces, this study presumed H to be a set of face group indexes. After computing h_i for every identified face area, h_i will be auxiliary to$H$, that is, $H = G \cup \{h_i\}$. Given a set of faces in a video dataset$\left(h, y_l, x_l, s_l, g_l\right), h \in H$, this study determined the face cluster $\overline{h}$ that is detected more often than the others in this video for further processing,

$$\overline{h} = mode\{H\} \tag{4}$$

The DNN model consists of 4 convolution layers, 4 max-pooling layers, and 2 fully connected layers. Batch normalization is employed for the outputs of convolutional layers and the fully connected layers. The feature maps, F, are determined by a low-level feature and the first convolutional layer. The remaining convolution layers automatically create a feature map, which indicates high-level features like corner points, edges, and colour from the face area. Here, both high-level and low-level features are measured for categorizing face expressions. A convolution layer performs a convolutional operation on face images, J, with the aid of kernels, S_j. Further, convolved features are fed into activation functions, which are, in this case, rectified linear units (ReLUs). Statistically, a convolutional operation can be signified by Eq. (4).

$$F(j) = R\left(J * S_j\right) \tag{5}$$

 https://doi.org/10.1007/s44196-024-00729-9

As inferred from Eq. (5), where j denotes the layers in consideration, the asterisk signifies the convolutional operation and $R(.)$ represents the activation functions. If the input, x, is positive, ReLU creates x as an output; otherwise, it produces 0. ReLU is typically utilized in hidden layers because it is better than all the accessible activation functions, like tanh, sigmoid, etc. ReLU is recognized as a ramp function. Equation 6 is the depiction of ReLUs.

$$\mathrm{ReLU}(x) = \begin{cases} 0 \; if(x < 0) \\ x \; otherwise \end{cases} \tag{6}$$

Because batch normalization allows for a high learning rate without producing a vanishing gradient issue, it is used to normalize the output of the input and hidden layers by altering the mean and scale of the activation functions. Following the activation function provides improved performance. Yet before it goes into the next layer, the output of the input layer is normalized. To reduce the problem of overfitting, the pooling operation is implemented on the convolved feature maps that have been generated using Eq. 4. Subsampling describes the pooling process. Because it uses fewer DNN parameters, it can reduce the image's spatial representation. The three most common pooling processes are maximum pooling, minimum pooling, and average pooling. Lastly, the softmax layer receives output from the second completely linked layer. The deep neural net architecture uses the softmax activation function in dense layers. The predicted classes' probabilities are computed using Softmax. An output is the class that has the greatest probability. As demonstrated in Eq. 7, the softmax function is mathematically represented.

$$W_y = \frac{e^{zy}}{\sum_{x=1}^{n} e^{zx}}. \tag{7}$$

As discussed in Eq. (6), where e^{zy} and e^{zx} signify the likelihood of belonging to the classes of y and x, correspondingly, whereas n represents the number of classes. In this study, the value of m equals seven because 7 FER are considered.

The cross-entropy functions are articulated by:

$$P_{cross} = -\sum_{i=1}^{z} t_j \log(x_j) \tag{8}$$

As found in Eq. (8), where P_{cross} denotes the cross-entropy and t_j and x_j symbolize the true and estimated value correspondingly. The abovementioned processes are iterated to optimize the variables included in network models.

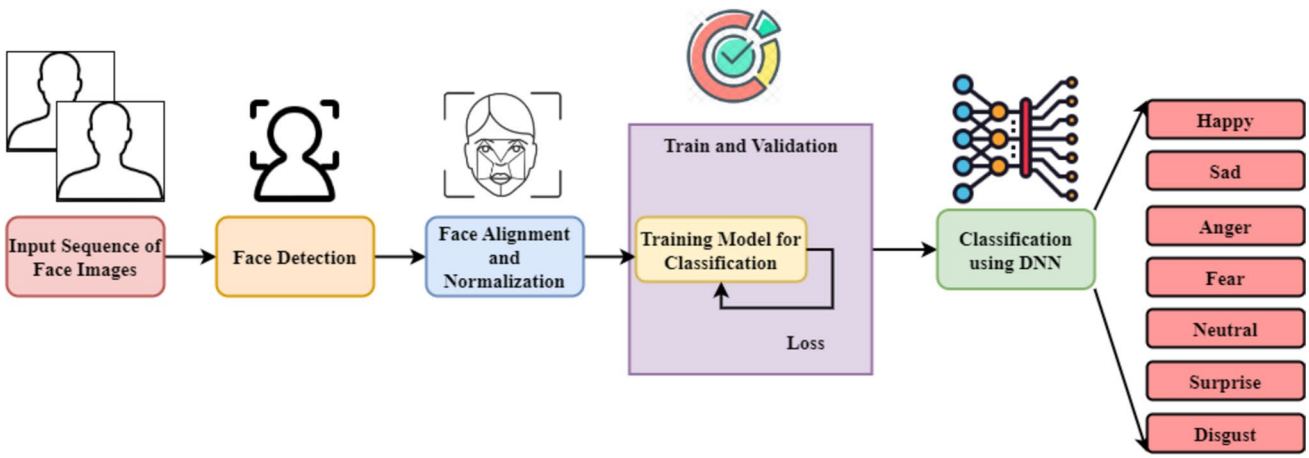


Fig. 3 Flowchart of DNN-based classification of facial expressions

Figure 3 shows the flowchart of DNN-based classification of facial expressions. Using a video camera to capture a sequence of 2D images of an individual's face allows us to accomplish our research goals in this study. To build the technological basis for expression-based emotion detection, it is necessary to observe individuals' faces continually. This is necessary as the instructional model changes from result-based evaluation to process-based assessment. Using the suggested method, which involves image-based facial expression identification, this study sorts the expressions into three groups, one for each difficulty in identifying the problems. A sequence of facial images captures the emotions in real time. The study's input images must undergo pre-processing, such as face detection, smoothing, alignment (registration), and normalization since the suggested method relies on 2D facial images. These images are selected from the recorded images and utilized as input for the recognition system. Proper comparison of facial expressions with the same or different individuals is ensured by pre-processing. Compared with existing methods, improved F1-score, classification, VR performance, face landmark prediction, and user satisfaction ratios are outcomes of the suggested DAFERM model.

4 Results and Discussion

This study discussed the Deep Automatic Facial Expression Recognition Model (DAFERM) for interactive virtual reality (VR) applications. The data on facial expression recognition are gathered from the Kaggle dataset [30]. The seven emotions in the database are anger, dissatisfaction, fear, surprise, sadness, and contempt. With each picture standing in for one of these distinct emotions, this dataset provides a great opportunity for machine learning researchers and practitioners to study and develop models for emotion recognition and analysis. The seven emotions represented in the dataset are surprise, anger, contempt, disgust, fear, joy, and sorrow. Machine learning experts and academics may use the dataset to study and build models for emotion identification and analysis, as each picture represents one of four distinct emotions. People of all ages, genders, and races are represented in the photographs. To cover all bases, the dataset collects examples of human emotions. Pictures: there are folders for each individual that include pictures that portray eight various emotions; the names of the files reflect the emotions they portray. Dataset information is stored in a CSV file. Table 1 shows the experimental setup and parameters of the suggested DAFERM model. The Kaggle dataset, which includes 2D images representing various emotions, was used to train and assess the Facial Expression Recognition (FER) model. A real-time solution that used an external camera with the VR headgear was used to bring the model to life in the virtual reality scenario. The setup enabled dynamic facial emotion recognition by synchronizing gaze and movement information with the HMD's eye-tracking data. The identification primarily targeted unobstructed facial parts, such as the mouth. A synchronization system was implemented to ensure the data from the VR HMD's eye-tracking sensors were in sync with the external camera feed. The method overcame the HMD's occlusions by retrieving face attributes from the visible portions using adaptive histogram-based mouth segmentation. The FER model effectively assessed user emotions in real-time, bridging the gap between training with 2D images and the limitations of VR applications.

Table 1 Experimental Setup and Parameters

Category	Parameters
CPU	Intel(R) Core(TM) i5-8250U
System type	64 bits
Operation system	Window 10
Simulation environment	Anaconda
Memory (RAM)	8.0 GB
Batch size	64
Learning rate	0.001

Trials were conducted in a controlled virtual reality environment to ensure the device could successfully recreate immersive experiences. https://www.kaggle.com/datasets/tapakah68/facial-emotion-recognition.

4.1 Classification Accuracy Ratio

A person's facial expressions may convey much about their mental and behavioural states. This study can effectively determine their emotional state within a few seconds of looking at a person's face. Student awareness estimation, medical care, and human–computer interaction are some of the many uses for facial expression detection. With hyperparameter adjustment, this study investigated how well the proposed DNN model performed. The idea of expression poses, or light-invariant face recognition, has been put forward to establish a reliable system for identification and classification. Our approach completely utilizes the 3D coordinates of a target face and the geometric information of a face, leading to much-improved accuracy and robust findings. Raising the number of epochs, or learning iterations in a DNN model, improves classification performance (both in learning and validation accuracy). For classification tasks, accuracy is the most important statistic. It is used to identify the percentage of the complete test set that was properly predicted, both positive and negative. Figure 4 denotes the classification accuracy ratio.

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{9}$$

As shown in Eq. (9), where TP represents the true positive, TN specifies the true negatives, FP represents the false positives, and FN denotes the false negatives.

4.2 F1-Score Ratio

RGB images' ability to capture crucial geometrical elements makes higher accuracy and condition-insensitive face information retention possible. Giving specifics on the percentage of the positively expected outcomes demonstrating precision is really about. Results showed that all labels had precisions over 50% for the models, meaning that the model accurately predicted each emotion more than half the time. The recall, or sensitivity, represents the actual rate of favourable outcomes. The f1-score metrics give the balance of precision and recall. Figure 5 demonstrates the F1-score ratio.

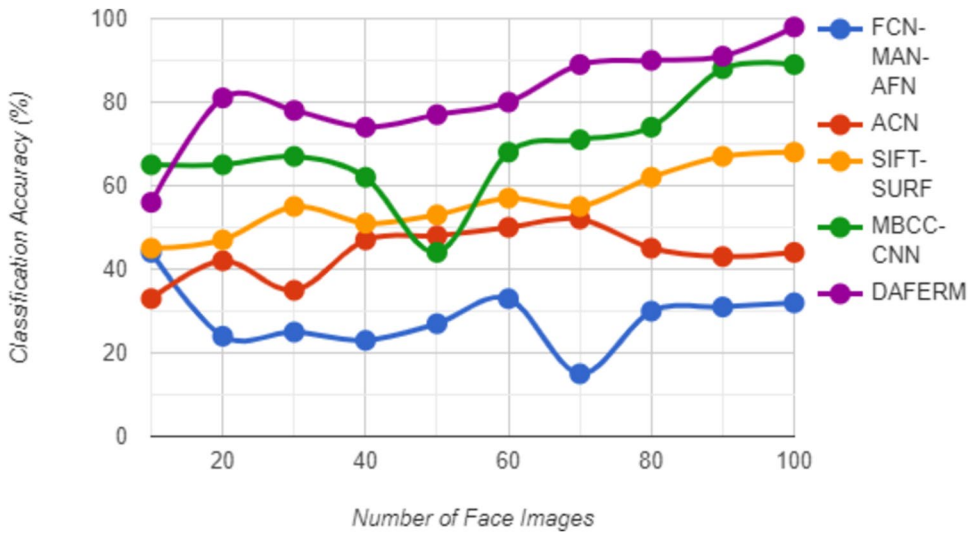
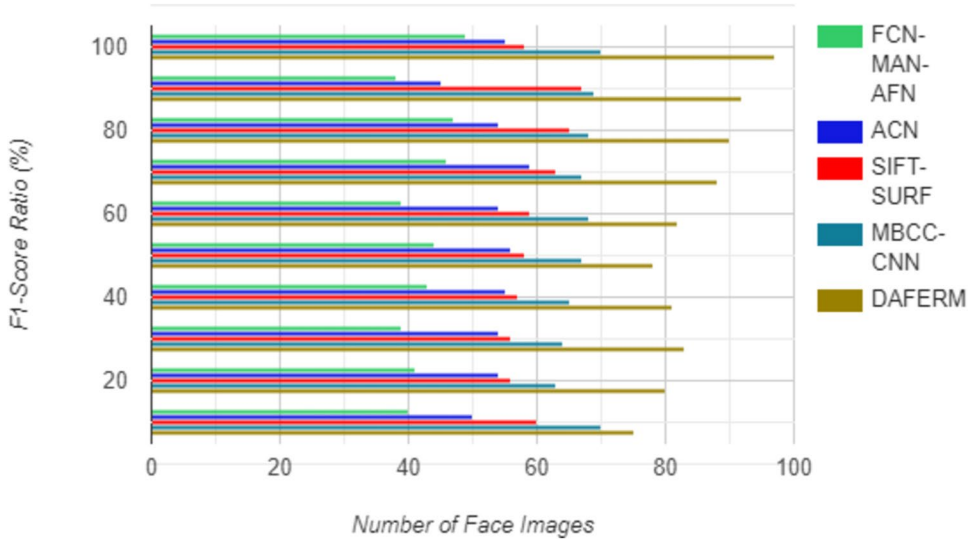


Fig. 4 Classification Accuracy Ratio

Fig. 5 F1-Score Ratio



$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$F1 - \text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

As inferred from Eqs. (10)–(12), where TP represents the true positive, TN indicates the true negatives, FP symbolizes the false positives, and FN denotes the false negatives. Table 2 express Fig. 5 values.

4.3 User Performance in VR Environment

Virtual Reality (VR) and other interactive and immersive technologies are quickly replacing the costly prototype as the primary means of visually representing a product, design, or environment. Researchers in the field of human–computer interaction should keep the intricacy of the communicative act in mind as they seek to model and improve this connection by creating smart applications that make automated systems easier to use and more engaging for the typical user. Virtual reality systems must be able to detect and portray facial emotions to provide immersive social interaction. This research uses technical expertise to create three-dimensional models, new virtual reality technologies to control virtual objects in three dimensions, the human factors method to assess how real and virtual environments interact, and modelling tools to see and assess the outcomes. The effects of the suggested systems' features on users, such as immersion, interaction

Table 2 F1-Score Ratio (%)

Number of Face Images	FCN-MAN-AFN (%)	ACN (%)	SIFT-SURF (%)	MBCC-CNN (%)	DAFERM (%)
20	60	62.8	65.4	70.9	80.8
40	64.2	67.4	69.8	75.2	85.4
60	68.4	71.6	73.2	79.5	89.7
80	72.1	74.3	76.9	82.7	92.3
100	75.3	78.5	80.7	85.4	95.6

 https://doi.org/10.1007/s44196-024-00729-9

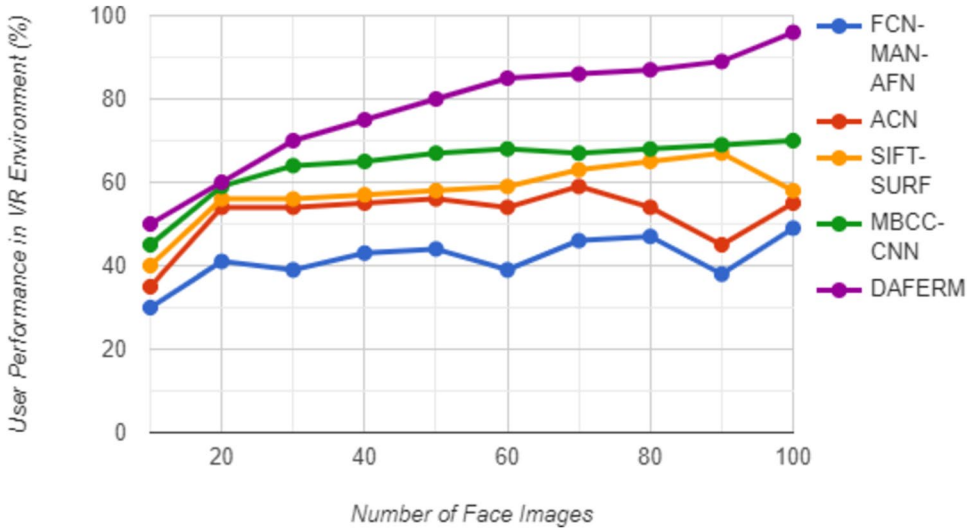


Fig. 6 User Performance in VR environment

types, and the use of eye-level view in spatial design, were investigated by studying users' spatial decisions, performance, and design outcomes in systems. Figure 6 illustrates the user performance in a VR environment.

4.4 Face Landmark Prediction Ratio

Eyes, noses, and corners of the mouth are common facial cues that may be used to rotate the depicted face. Secondly, a feature vector is constructed by concatenating features collected from the clipped face area. The next step is to train a machine learning system to recognize facial expressions using the feature vector. Successful recognition results have been achieved in face recognition employing information such as landmarks and 3D curves describing faces' geometric shapes, intensity, and eigenfaces.

To locate and remove faces from an input image, pre-processing is used to identify human faces. Step two involves using a library to identify facial landmarks for each face. The human face is first divided into the lower and upper half to facilitate feature extraction further. Both texture and geometric-based feature classes are taken into account in the suggested model. Figure 7 and Table 3 signifies the face landmark prediction ratio.

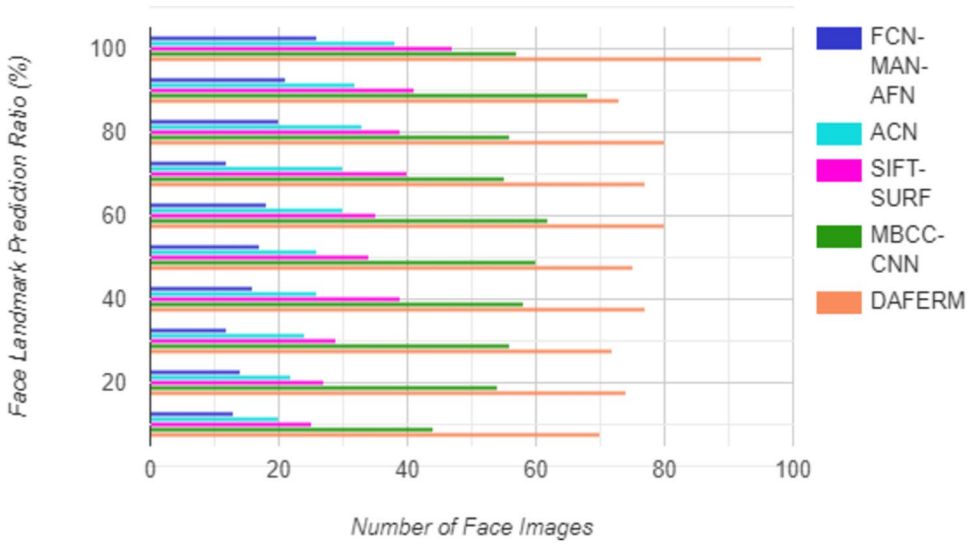


Fig. 7 Face Landmark Prediction Ratio

Table 3 Face Landmark Prediction Ratio

Number of Face Images	FCN-MAN-AFN (%)	ACN (%)	SIFT-SURF (%)	MBCC-CNN (%)	DAFERM (%)
20	62.4	65.7	68.5	73.9	81.9
40	65.1	69.2	71.9	77.3	85.2
60	68.3	72.1	75.7	80.8	88.4
80	71.8	75.4	78.9	84.2	91.6
100	74.5	79.2	82.6	87.3	94.8

4.5 User Satisfaction Ratio

The visual system prioritizes the satisfaction of the 3D scene presentation above all other requirements. This study examined the user's subjective experience, including their level of satisfaction and degree of usability, in the experiment. For example, virtual reality equipment may collect biometric data like eye movement and the user's activities and responses inside an experience, including their choices. Business brands may use this data to learn more about customer preferences and average problems. This study explains VR as an environment for gathering data on user satisfaction. Visualizing new product concepts in Virtual Reality (VR) offers distinct benefits due to the enhanced interaction it makes possible. Figure 8 illustrates the user satisfaction ratio.

5 Discussion

In a real-time virtual reality (VR) context, this model's balance between computational efficiency and feature extraction makes it well suited for face expression recognition (FER). Using four convolutional layers, the model captures surface-level information (edges and textures) and deeper semantic features (such as the underlying facial muscle movements) without introducing any needless complexity. Incorporating four max-pooling layers reduces dimensionality, decreasing computational overhead and overfitting; batch normalization stabilizes and accelerates training by normalizing layer inputs.

Improving performance by adding extra convolutional layers after training the model to comprehend more complex feature hierarchies is possible. If this improvement increases the need for computer power, it might affect virtual reality (VR) applications that depend on real-time processing. Overfitting or falling returns could occur from adding more layers without enough training data, which might weaken the system's resilience.

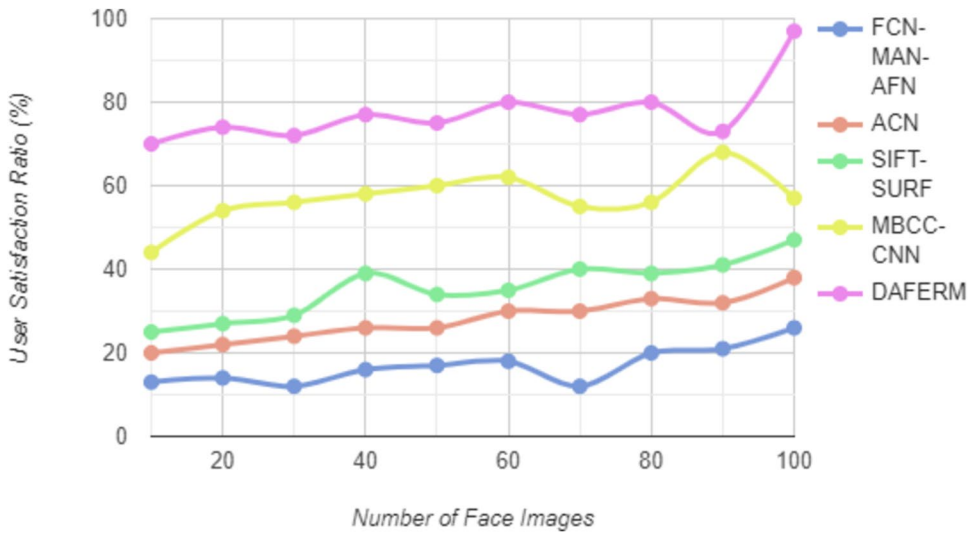


Fig. 8 User Satisfaction Ratio

Based on previous research, the architecture was designed to mimic the performance of CNNs with similar topologies on FER tasks. The importance of tailoring CNN depth to specific applications has been highlighted in previous research. This model uses a rather deep network that follows FER best practices to deal with VR-specific problems like occlusion and real-time constraints. The architecture shows a well considered and functional design choice.

6 Conclusion

This study presents the Deep Automatic Facial Expression Recognition Model (DAFERM) for interactive virtual reality (VR) applications. Based on a DNN, the suggested model learnt the user's facial expressions to classify and recognize their emotions concerning the severity of the VR scenario challenges. Training and validation were both shown to be reasonably accurate in the studies. Training and validation accuracy are both affected by the input image quality. The combination of 2D and 3D views and carefully chosen facial key points formed the basis of the identification. The system's performance tests demonstrate that it can generally identify the user's face interest regions over the whole pitch angle, consistently distinguish expressions with an accuracy of over 90%, and remain highly resilient when exposed to changes in light intensity. Through eye-tracking technology and virtual reality, this research examined the effects of spatial and design features on users' emotional arousal and visual attention in an immersive environment.

Acknowledgements This work was supported by the cultivation and construction of electronic information key disciplines in the School of Physical and Electronic Information Engineering, Jining Normal University. This work was supported by PhD Innovation Research Fund Project of Jining Normal University (Number: jsbsjj2335, jsbsjj2336, jsbsjj2413). Intelligent Recognition and Image Processing Research Center(Number: jskypt2436).

Author Contributions Qingzhen Gong: write the manuscript Xuefang Liu: editing data curation, writing—original draft preparation Yongqiang Ma: collect and analyse the data.

Funding This work was supported by the cultivation and construction of electronic information key disciplines in the School of Physical and Electronic Information Engineering, Jining Normal University. This work was supported by PhD Innovation Research Fund Project of Jining Normal University (Number: jsbsjj2335, jsbsjj2336, jsbsjj2413). Intelligent Recognition and Image Processing Research Center(Number: jskypt2436).

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Cha, H.S., Choi, S.J., Im, C.H.: Real-time recognition of facial expressions using facial electromyograms recorded around the eyes for social virtual reality applications. IEEE Access **8**, 62065–62075 (2020)
2. Hassouneh, A., Mutawa, A.M., Murugappan, M.: Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. Inform. Med. Unlocked **20**, 100372 (2020)
3. Ge, H., Zhu, Z., Dai, Y., Wang, B., Wu, X.: Facial expression recognition based on deep learning. Comput. Methods Programs Biomed. **215**, 106621 (2022)
4. Lee, H.J., Lee, D.: Study of process-focused assessment using an algorithm for facial expression recognition based on a deep neural network model. Electronics **10**(1), 54 (2021)
5. Lee, J.R., Wang, L., Wong, A.: Emotionnet nano: an efficient deep convolutional neural network design for real-time facial expression recognition. Front. Artif. Intell. **3**, 609673 (2021)
6. Kim, J., Kang, J.K., Kim, Y.: A resource efficient integer-arithmetic-only FPGA-based CNN accelerator for real-time facial emotion recognition. IEEE Access **9**, 104367–104381 (2021)
7. Ngo, Q.T., Yoon, S.: Facial expression recognition based on weighted-cluster loss and deep transfer learning using a highly imbalanced dataset. Sensors **20**(9), 2639 (2020)
8. Sungheetha, A., Sharma, R.: 3D image processing using machine learning based input processing for man-machine interaction. J. Innov. Image Process. (JIIP) **3**(01), 1–6 (2021)
9. Zhang, K., Li, Y., Wang, J., Cambria, E., Li, X.: Real-time video emotion recognition based on reinforcement learning and domain knowledge. IEEE Trans. Circuits Syst. Video Technol. **32**(3), 1034–1047 (2021)
10. David, D.S., Samraj, M.: A comprehensive survey of emotion recognition system in facial expression. Artech J. Eff. Res. Eng. Technol **1**, 76–81 (2020)
11. Garcia, A.S., Fernandez-Sotos, P., Vicente-Querol, M.A., Lahera, G., Rodriguez-Jimenez, R., Fernandez-Caballero, A.: Design of reliable virtual human facial expressions and validation by healthy people. Integr. Comput. -Aided Eng. **27**(3), 287–299 (2020)
12. Mohan, K., Seal, A., Krejcar, O., Yazidi, A.: Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks. IEEE Trans. Instrum. Meas. **70**, 1–12 (2020)
13. Jeong, D., Kim, B.G., Dong, S.Y.: Deep joint spatiotemporal network (DJSTN) for efficient facial expression recognition. Sensors **20**(7), 1936 (2020)
14. He, Z., Li, Z., Yang, F., Wang, L., Li, J., Zhou, C., Pan, J.: Advances in multimodal emotion recognition based on brain–computer interfaces. Brain Sci. **10**(10), 687 (2020)
15. Bhatti, Y.K., Jamil, A., Nida, N., Yousaf, M.H., Viriri, S., Velastin, S.A.: Facial expression recognition of instructor using deep features and extreme learning machine. Comput. Intell. Neurosci. **2021**, 1–17 (2021)
16. Crenn, A., Meyer, A., Konik, H., Khan, R.A., Bouakaz, S.: Generic body expression recognition based on synthesis of realistic neutral motion. IEEE Access **8**, 207758–207767 (2020)
17. Wedyan, M., Falah, J., Alturki, R., Giannopulu, I., Alfalah, S.F., Elshaweesh, O., Al-Jumaily, A.: Augmented reality for autistic children to enhance their understanding of facial expressions. Multimodal Technol. Interact. **5**(8), 48 (2021)
18. Sikkandar, H., Thiyagarajan, R.: Deep learning based facial expression recognition using improved cat swarm optimization. J. Ambient. Intell. Humaniz. Comput. **12**, 3037–3053 (2021)
19. Pabba, C., Kumar, P.: An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. Expert. Syst. **39**(1), e12839 (2022)
20. Wen, Z., Lin, W., Wang, T., Xu, G.: Distract your attention: Multi-head cross attention network for facial expression recognition. Biomimetics **8**(2), 199 (2023)
21. Minaee, S., Minaei, M., Abdolrashidi, A.: Deep-emotion: facial expression recognition using attentional convolutional network. Sensors **21**(9), 3046 (2021)
22. Gupta, S., Thakur, K., Kumar, M.: 2D-human face recognition using SIFT and SURF descriptors of face's feature regions. Vis. Comput. **37**, 447–456 (2021)
23. Shi, C., Tan, C., Wang, L.: A facial expression recognition method based on a multi-branch cross-connection convolutional neural network. IEEE access **9**, 39255–39274 (2021)
24. Abdulrazaq, M.B., Mahmood, M.R., Zeebaree, S.R., Abdulwahab, M.H., Zebari, R.R., Sallow, A.B.: An analytical appraisal for supervised classifiers' performance on facial expression recognition based on relief-F feature selection. J. Phys. Conf. Ser. **1804**(1), 012055 (2021)
25. Zhao, Z., Liu, Q., Wang, S.: Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE Trans. Image Process. **30**, 6544–6556 (2021)
26. Kola, D.G.R., Samayamantula, S.K.: A novel approach for facial expression recognition using local binary pattern with adaptive window. Multimed. Tools Appl. **80**, 2243–2262 (2021)

27. Liu, C., Hirota, K., Dai, Y.: Patch attention convolutional vision transformer for facial expression recognition with occlusion. Inf. Sci. **619**, 781–794 (2023)
28. Uyanık, H., Ozcelik, S.T.A., Duranay, Z.B., Sengur, A., Acharya, U.R.: Use of differential entropy for automated emotion recognition in a virtual reality environment with EEG signals. Diagnostics **12**(10), 2508 (2022)
29. Izountar, Y., Benbelkacem, S., Otmane, S., Khababa, A., Masmoudi, M., Zenati, N.: VR-PEER: a personalized exergame platform based on emotion recognition. Electronics **11**(3), 455 (2022)
30. Roman, K.: Facial emotion recognition dataset (2023)

Authors and Affiliations

Qingzhen Gong[1] · Xuefang Liu[2] · Yongqiang Ma[3]

✉ Xuefang Liu
yjxuefang90022@yeah.net

Qingzhen Gong
gqz0125@126.com

Yongqiang Ma
nsd-myq@126.com

[1] School of Physical and Electronic Information Engineering, Jining Normal University, Ulanqab 012000, Inner Mongolia, China

[2] School of Information Engineering, Jingdezhen University, JingDeZhen 333000, Jiangxi, China

[3] School of Computer and Big Data, Jining Normal University, Ulanqab 012000, Inner Mongolia, China