




Impact of Deep Learning Approaches on Facial Expression Recognition in Healthcare Industries

Carmen Bisogni , *Member, IEEE*, Aniello Castiglione , *Member, IEEE*, Sanoar Hossain, Fabio Narducci , *Member, IEEE*, and Saiyed Umer

Abstract—A facial expression recognition system that can provide quick assistance to the healthcare system and exceptional services to the patients is proposed in this article. The implementation of this work is divided into three components. In the first component, landmark points on the facial region are detected; a fixed-sized rectangular box is obtained by normalizing the detected face region, and then, down sampled to its varying sizes producing multi-resolution images. Different convolution neural network architectures are proposed in the second component for analyzing the textual information within the multi-resolution facial images. To extract more discriminating features and enhance the proposed system's performance, some amalgamation of transfer learning, progressive image resizing, data augmentation, and fine tuning of parameters are employed in the third component. For experimental purposes, three benchmark databases, static facial expressions in the wild, Cohn-Kanade, and Karolinska directed emotional faces, are employed with some existing methods concerning these databases. The comparison with these databases shows the superiority of the proposed system.

Index Terms—Convolutional neural networks (CNNs), deep learning, facial expression, recognition, score-level fusion.

I. INTRODUCTION

HEALTHCARE industry is an essential sector in employment, health-related facilities, revenue, and contribution for healthy civilization in smart cities. Healthcare industries include clinical trials, hospitals, medical equipment and devices, telemedicine, outside patient clinic, health insurance, doctors, nurses, and other medical professionals. A new era of technologies has enriched healthcare industries by introducing m-Health and e-Health facilities. The term m-Health is mobile health that

Manuscript received June 30, 2021; revised September 17, 2021 and November 18, 2021; accepted December 24, 2021. Date of publication January 7, 2022; date of current version May 6, 2022. Paper no. TII-21-2751. (Corresponding author: Aniello Castiglione.)

Carmen Bisogni and Fabio Narducci are with the Department of Computer Science, University of Salerno, 84084 Fisciano, Italy (e-mail: cbisogni@unisa.it; fnarducci@unisa.it).

Aniello Castiglione is with the Department of Science and Technology, University of Naples Parthenope, 80143 Naples, Italy (e-mail: castiglione@ieeee.org).

Sanoar Hossain and Saiyed Umer are with the Department of Computer Science and Engineering, Aliah University, Kolkata 700064, India (e-mail: snr.hossain12@gmail.com; saiyedumer@gmail.com).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TII.2022.3141400>.

Digital Object Identifier 10.1109/TII.2022.3141400

provides mobile device-based practices to patients to support their medicines and daily healthcare facilities. e-Health is an electronic health service that uses information and communication technology (ICT) for delivering digitally processed facilities to patients and doctors through computers and drugs. Both e-Health and m-Health have immense support to the healthcare industry benefitting patients, doctors, medical professionals, and businesses, establishing a healthy civilization in the smart cities. Healthcare is a real-time monitoring system with technological advancement in ICT. Health is an essential element and the most crucial dimension of Internet-of-Things (IoT) used to build a smart city.

There is a need for some smart health services by combining the facilities of m-Health and e-Health services. Smart health will provide support not only to patients and hospitals but also to the whole healthcare industry in the cities. The smart health system will handle the patient's information about his/her leaving from residence, weather, safest routes, temperature, traffic congestion, etc. Even the patient will be guided to his nearest healthcare services like an emergency, ambulance, medical practitioners, diagnostics, prompt deceases, etc. Doctors and caregivers receive regular information from the healthcare framework via a cloud server and monitor the skin conductance, heartbeat, and skin temperature. The final notification and decision come from doctors, alerting the traffic managers, hospital managers, and caregivers to take appropriate actions. Hence, the smart healthcare system will monitor all the issues either related to health or patients. Smart healthcare systems are real-time systems, and they have to face several challenges for maintaining, monitoring, and providing privacy protection to patient's information [1]. A patient may be an actor, actress, politician, industrialist, or government official, so the privacy protection and security of patient's information is a significant challenging issue. A smart healthcare system for patient's discomfort detection using deep learning-based techniques with the IoT has been proposed by Ahmed *et al.* [2].

Human-computer interaction is used to communicate with doctors, caregivers, clinicians, and patients through automated response sensor devices and monitoring systems in computer vision [3]. The real-time application of the facial expression recognition system (FERS) of computer vision incorporates the healthcare system with the ICT and IoT [4] [5], providing continuous interactions toward patient's monitoring. Facial expressions are the crucial nonverbal way of indicating mummies and the unique universal way for people to communicate. These are caused by facial muscle movements. The FERS reveals



Fig. 1. Example of seven types of facial expression for the FERS.

human emotions and attitudes in people's daily communication through the system. An automatic FERS helps to detect some feelings of people. It is an emergent research topic in computer vision research areas. The FERS has comprehensive potential applications in fields with various challenges. It is a contactless recognition system where a person's human face image can be captured from a distance without any intervention or interruption even if he/she is moving around or walking or sitting, or performing some activities. The facial expression plays a vital role in our daily communication with people and social interaction. Fig. 1 shows some examples of seven basic facial expressions, e.g., fear (AF), anger (AN), disgust (DI), happy (HA), neutral (NE), sad (SA) and surprise (SU) on the human face.

It has an unmistakable indication of the affective state, intellectual movement, personality, self-aided driving, attention, and a hint of person's mind. It is an easy task for a human being to recognize facial expressions, but it is pretty challenging to identify them by the computer. It is an immediate, powerful, and effective nonverbal way of communication to transit messages and convey the emotional information. A human brain can quickly identify the expressions by looking at the facial muscular movement. It is noticeable from the expressions that much of the informative and useful information is obtained from the mouth, nose, lips, and eyes, i.e., action units (AUs). The other parts of the facial region contribute to enhance the expressions in support of AU points. Here, the psychology of emotions is being considered as AUs, which are extracted from the facial region of interest. The facial area is further analyzed as texture where numerous techniques such as statistical and structural-based methods have been employed to extract more discriminant features. Apart from these techniques, recently, deep learning-based approaches with convolution neural networks (CNNs) have been employed to extract more discriminant and distinctive features. These methods can give a better performance. But most of these methods are database dependent, and these databases have been captured spontaneously under a controlled environment and have tightly controlled illumination, age, and pose variation conditions.

Despite these current state-of-the-art methods for the FERS and their remarkable progress in affective computing, these methods still suffer from some limitations. First, the employed datasets are either laboratory-controlled datasets or the wild datasets. These images are captured under unconstrained environments. These images suffer from several challenging issues like illumination, poor resolution, occlusion, pose, age, and expression variations. So, the extraction of the face region from the input images with optimum time is challenging. Second, due to limited domain knowledge, the local-global feature

representation generates less discriminative and distinctive features. Third, the assumption might not be valid, i.e., maybe a failure to extract the local geometric information, and the AU detection tasks itself is challenging. Hence, to solve these issues here, we have proposed a novel deep-learning-based framework for the FERS to address these problems. The contributions of the proposed work are demonstrated as follows.

- 1) A fast and efficient end-to-end deep learning-based framework is achieved by adding new level of feature representation techniques added to traditional CNNs.
- 2) Powerful high-level generic features are extracted from the input image for an effective FERS under various illumination changes, pose, and age variations artifacts.
- 3) A tradeoff between batch versus epoch has been analyzed.
- 4) The proposed FERS compares in term of performances similar approaches in the literature. Nonfrontal poses and 2-D images have been also considered in the experiments.
- 5) To reduce the training loss, we have used the ReLU activation function and a combination of random weights such that the model can reduce the overfitting problem that arises because of inadequate training data and bias caused in dataset due to the variation in the expressions.

The rest of this article is organized as follows. Section II describes the related work for the proposed system. The proposed FERS is discussed in Section III, which describes the face preprocessing techniques and the proposed CNN architectures for the feature computation from both frontal and profile facial images. The database description, experimental results, and discussions are given in Section IV. Finally, Section V concludes this article.

II. RELATED WORK

Depending on the existing state-of-art methods for face representation, facial expression recognition (FER) could be broadly classified and analyzed into two categories: appearance-based methods and facial AUs-based methods. In the appearance-based methods, the entire face region has been divided into some blocks or patches, and the features are extracted from these patches using image processing techniques.

Many state-of-the-art methods and deep learning frameworks used hand-labeled points and CNN architecture for feature extraction and built an FERS. Gutta *et al.* [6] proposed a model with an ensemble radial basis function, grayscale image, and inductive decision trees for four classes (i.e., Asian, Caucasian, African, and Oriental) ethnicity recognition problem. Zhang and Wang [7] proposed a method for two-class racial classification using multiscale local binary pattern (LBP) texture features combining 2-D and 3-D texture features. Zhang *et al.* [8] described two types of features, the geometry-based features and Gabor-wavelets-based features for the FERS. Bartlett *et al.* [9] applied the Gabor filters coupled with feature selection and machine learning techniques for recognizing the facial expressions on a human face. Rose [10] applied Gabor and log-Gabor filters on low-resolution images for FER.

Gu *et al.* [11], proposed a method for FER based on the radial encoding of local Gabor features with classifiers synthesis, while Otterdout *et al.* [12] proposed a generative adversarial network based on Hilbert hypersphere with conditional

Wasserstein. Siqueira *et al.* [13] proposed ensembles with shared representation-based CNN models for predicting expressions on human face. Vo *et al.* [14] proposed pyramidic with superresolution network architecture-based image representation for detecting expressions in wild. Zhou *et al.* [15] extracted emotion features from audio and video traits individually, and then, explored intramodal and cross-modal feature fusion-based techniques for identifying emotions in wild. Wang *et al.* [16] proposed a region attention network with the backbone CNN feature extraction for predicting expressions in the wild. The major problems that occur during developing the FERS are shallow features and bias caused by various cultural and collection conditions. Current datasets have a strong build-in preference, and the corresponding proposed methods show that the conditional probability distribution between training and testing datasets is different. To address these issues, we should look deeper into this biasness and propose some novel deep CNN models. Since in our proposed methodology, we have considered that face recognition is an image classification problem, this face recognition has been extended to our work for the classification of facial expressions on human faces. The proposed FERS is based on two backbones: face preprocessing and design and analysis of features from the proposed CNN architectures. The goal of preprocessing is to enhance the expression of the region of interest and suppresses the unwanted, redundant, and inconsistent noises in the image, which need image processing techniques. Then, pixel brightness, geometric transformations, certain preprocessing using local neighborhoods of pixels, and image restoration techniques of image processing are finally required for detecting the face region from the input image. Additionally, the image augmentation techniques increase the amount of training data on the existing data by applying some image processing affine and transformation techniques such as bilateral filtering, unsharp filtering, horizontal flip, vertical flip, Gaussian blur, additive Gaussian noise, image scale, image cropping, translation, image rotation, shear mapping, image zooming, image filling, and contrast normalization methods.

Since the input we examined is an image, we do not consider a temporal model. Even if temporal model could represent an advantage in terms of available amount of data for each prediction, in the presented case use (healthcare) the response time are vitally important. For this reason, a lighter architecture using only an image as input is preferable. The proposed CNN architecture is built using several convolutional layers, max-pooling, batch normalization, and drop-out layers with optimizer followed by the soft-max classifier for the final classification tasks. Our extensive random experimental results show that our proposed deep-CNN method achieves superior results for FER problems for both lab-controlled and real-world databases. The principle issues involved in the FERS designs are: face representation and classifier selection. The face representation consists of the extraction of feature descriptors from the input face image that should minimize the intraclass similarities and maximize the interclass dissimilarities. In the case of classifier selection, it does not make sense that the high-performance

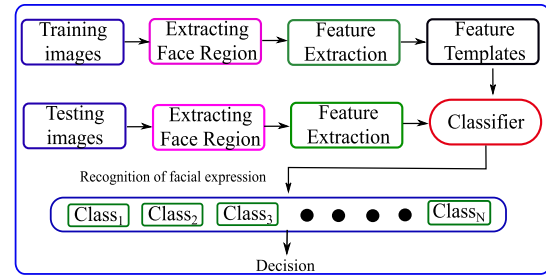


Fig. 2. Block diagram of the proposed FERS.

classifiers may always find a better separation between different classes even if there are some significant similarities between each other. Sometimes the most sophisticated classifier may fail to execute the FER and classification tasks because of inadequate face representations. If we have employed good face representation but do not select a good classifier, we cannot achieve high-performance recognition accuracy. Hence, the following sections have described the proposed FERS.

III. PROPOSED METHODOLOGY

An FERS generally consists of face representation, feature extraction, and classifier components. Looking at the importance of face recognition in computer vision systems, we have proposed a robust, efficient, and accurate model, i.e., deep CNN model for FER. An image $\mathcal{I}_{m \times n}$ with a valid face region is used as an input to the system. Our objective is to predict the types of emotion like fear, anger, disgust, surprise, sadness, happy, and neutral from the input face region $\mathcal{I}_{m \times n}$. The proposed FERS has been implemented in the following four steps.

- 1) Preprocessing: Bounded box face region is detected from the input image using the tree structure part model.
- 2) Feature extraction: From detected bounded box face region, the global generic CNN features have been extracted and prepared for the next level through a deep learning model.
- 3) The representations are further modified by using multistage progressive image resizing and transfer learning methods followed by image augmentation and fine-tuning parameter setting techniques.
- 4) Classification: Predict the type of emotion classes on the face region.

Each of these steps is described by the block diagram of the proposed system demonstrated in Fig. 2.

Recently deep CNN techniques have been successfully developed to learn discriminative features in various fields. It has widely been used in deep FER representation. Deep FER suffers from the overfitting problem. The lack of sufficient training samples and bias cause expression variations such as age variation, head pose, identity bias, and illumination variation. The proposed methods focus on these issues and overcome the computational complexity of the CNN model.

A. Face Preprocessing

In this article, we have implemented a deep learning framework applied to recognize discrete human facial emotional expressions into their corresponding categories. The input face image has been resized to the same size and is normalized to the grayscale image of each input face image $\mathcal{I}_{m \times n}$. Here, these input images are mapped to the exact locations, i.e., eye, nose's tip, etc., known as a feature map. At the lowest level of abstraction, it is assumed that preprocessing is a standard term for operation computed over intensity images. These input and output intensity images are of the same kind as the sensor's original data. A matrix of image function values usually represents an intensity image. Preprocessing aims to enhance the expression of the region of interest and suppress the unwanted, redundant, and inconsistent part of image features for further processing. Geometric transformations of images (e.g., scaling, rotation, and translation) are classified among preprocessing methods that have been used here [17]. Image preprocessing techniques are classified into four categories according to the size of the pixel neighborhood that is used for the calculation of a new pixel brightness: pixel brightness transformations, geometric transformations, specific preprocessing methods that use a local neighborhood of the processed pixel, and image restoration that requires knowledge about the entire image. Image preprocessing techniques use considerable redundancy in images. Neighboring pixels in local feature extraction corresponding to one object in real images have essentially the same or similar brightness value. If a distorted pixel can be picked out from the image, it can usually be restored as an average value of neighboring pixels [18]. Here, the required face region is detected from the input image using OpenCV library for face detection and implementation. Detected face has been resized to $\mathcal{F}_{n \times n}$ and normalized to fixed-size gray-scale image. These face images are used as input to the proposed CNN model. We extract the face region from each input image $\mathcal{I}_{m \times n}$ during preprocessing. Since the facial expressions contain very minute details, it is important to be conscious about analyzing both the facial region's expressive and nonexpressive characteristics. During face preprocessing, we have applied a tree-structured part model, which works better for both frontal and profile face region compared to Haar-like features. It has outstanding performance than the other face detection algorithm in computer vision. The tree-structured part model works on a mixture of trees with a global mixture of topological viewpoints changing. For an unconstrained image having an unknown face, the tree structure part model locates all the facial landmarks in $\mathcal{I}_{m \times n}$. Hence, the tree structure part model computes 39 landmarks for profile faces, while 68 landmark points for the frontal face. These landmark points undergo to compute four corner points of the face region $\mathcal{F}_{n \times n}$. The face preprocessing steps are shown in Fig. 3.

B. Feature Representation for Expression Classification

Feature extraction is a crucial task to extract discriminating features from the input face image $\mathcal{F}_{n \times n}$ such that the extracted part must contain more distinctive patterns. The input image may be grayscale or RGB color image. In the field of computer vision and image processing research areas, the feature extraction starts

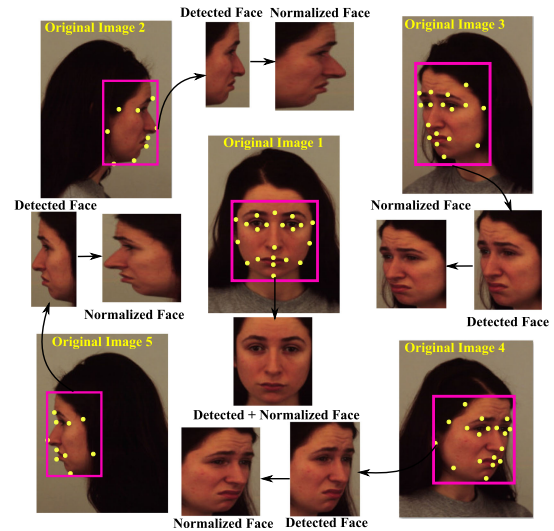


Fig. 3. Face preprocessing technique for the proposed system.

from an initial set of measured data and builds the features that are supposed to be informative and nonredundant, facilitating the subsequent learning and generalization steps. In many cases, the texture feature extraction techniques lead to better human interpretations. Moreover, it is related to the dimensionality reduction, i.e., when the input image size is too large to be processed as the representation for that image. It is transformed into a reduced set of features, also called a feature vector.

The modern state-of-the-art for generic CNN feature representation and emotion recognition problems could compete with the statistical and structural-based methods in computer vision systems. These generic CNN features can cope with the articulation and occlusion face images captured in an unconstrained environment and obtain a better performance. The proposed method describes a complex CNN baseline model with two components, i.e., feature extraction parts or hidden layers and classification parts. The proposed CNN model uses a CNN with five to seven deep and image perturbation layers. The model performs convolution operations using the ReLU activation function followed by max-pooling and batch normalization operations for feature extraction. Finally, two flattened layers are fully connected and used for classification tasks on the extracted features on top of the layers. The proposed CNN's performance has been increased by adding new levels to prove the image augmentation and progressive image resizing methods. These also help the model to prevent the overfitting and imbalanced data problem. Progressive image resizing methods support the model to avoid the use of much computational power. Because, here, only the pretrained weights of the last few layers are used, these weights have to be learned properly. We take advantage of image augmentation, batch normalization, activation function, and regularization method, including mix-up optimizer and label smoothing techniques.

During the feature representation of images using deep learning approaches, it has been observed that the CNN models obtain a better representation when patterns are analyzed from multiresolution of images. Additionally, increasing some layers in the architecture with increasing the resolution of the images,

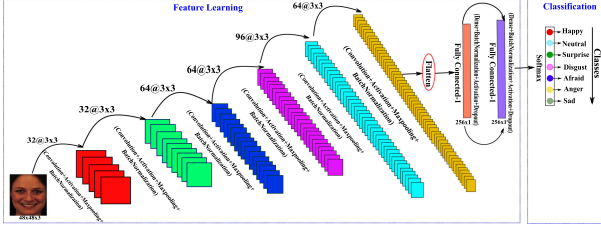


Fig. 4. CNN_1 architecture for the proposed FERS where $\mathcal{F}_{n_1 \times n_1 \times 3}$ is input to the system.

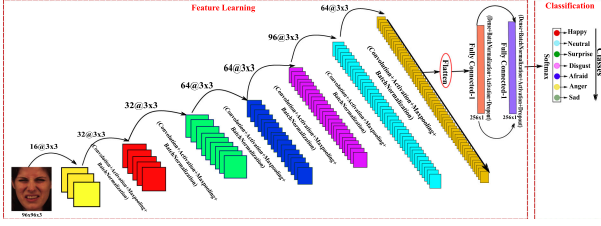


Fig. 5. CNN_2 architecture for the proposed FERS where $\mathcal{F}_{n_2 \times n_2 \times 3}$ is input to the system.

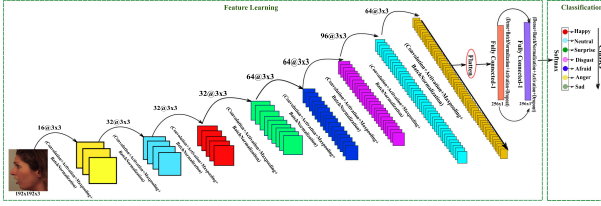


Fig. 6. CNN_3 architecture for the proposed FERS, where $\mathcal{F}_{n_3 \times n_3 \times 3}$ is input to the system.

goes deeper for analyzing some hidden patterns in the feature map. Inspired by these observations, we have applied some multiresolution of facial images with varying layers in different CNN architectures. During feature representations, we have considered facial image with three different resolutions such that the original facial image $\mathcal{F}_{n \times n}$ is down sampled to $\mathcal{F}_{n_1 \times n_1}$, $\mathcal{F}_{n_2 \times n_2}$, and $\mathcal{F}_{n_3 \times n_3}$, $n_3 = 2 \times n_2 = 4 \times n_1$. Here, for facial images, $\mathcal{F}_{n_1 \times n_1}$, $\mathcal{F}_{n_2 \times n_2}$, and $\mathcal{F}_{n_3 \times n_3}$, three different CNN architectures CNN_1 , CNN_2 , and CNN_3 are proposed. These architectures are shown in Fig. 4–6. The main differences between the networks are represented by the chosen dimension of the input layer and the amount of convolutional blocks. CNN_1 has 5 convolutional block starting with a size of (48,48,32), (24,24,64), (12,12,96), (6,6,96) and (3,3,64), the layer that flatten the data return such a size of 64. CNN_2 has 6 convolutional block starting with a size of (96,96,16), (48,48,32), (24,24,64), (12,12,96), (6,6,96) and (3,3,64), the layer that flatten the data return such a size of 64. CNN_3 has 7 convolutional block starting with a size of (192,192,16), (96,96,32), (48,48,32), (24,24,64), (12,12,96), (6,6,96), and (3,3,64), the layer that flatten the data return such a size of 64. We can thus notice that CNN_1 has the smaller architecture than CNN_2 and CNN_3 . CNN_2 has the smaller architecture than CNN_3 . However CNN_3 extracts bigger features than CNN_2 .

C. Factors Affecting Performance of the Proposed CNN

1) **Data Augmentation:** The image augmentation technique is used to expand the training samples to improve the recognition

and the ability to generalize the models. In machine learning, image augmentation techniques artificially increase training data by applying some transformation methods on the existing data. The classical augmentation techniques have employed bilateral filtering, unsharp filter, horizontal flip, vertical flip, Gaussian blur, additive Gaussian noise, image scale, image cropping and padding, translate, image rotate, shear mapping, image zooming, image filling, and contrast normalization methods for image augmentation purpose. The whole training images were flipped horizontally by applying simple image data augmentation techniques. In this article, we have involved these techniques for each resolution of the images.

2) **Progressive Resizing:** Progressive image resizing is a superior technique that sequentially resizes all images while training the CNN models on smaller, i.e., tiny images to larger image sizes. The progressive resizing method is used to introduce a CNN with $n \times n$ image size, save weights, and then, retrain again for the other iterations with the images of increased size greater than n . This technique was used in superresolution [19] where low-resolution images gradually increased to the image with a higher resolution during training. The advantages of using progressive resizing are that it improves generalization and reduces overfitting problems.

3) **Fine Tuning:** Fine tune allows the higher order feature representations in the base model to make them more relevant for the face recognition tasks. For example, very deep convolutional networks (VGG) used many layers and generated a higher dimensional feature vector, and thus, inference was quite costly at run time due to huge parameters. In this case, fine-tuning techniques have been applied that freeze some layers and number of parameters and retrain the model to reduce computational overheads.

4) **Transfer Learning:** The principle concept behind the transfer learning for FER and classification problem is that a model trained on large datasets for one question is effectively used as a generic model in some way on other related issues. The model that has been trained earlier is known as the pretrained model. Our proposed deep learning CNN model uses a transfer learning technique in which the weights of the pretrained model or a set of layers from the pretrained model CNN_1 are used to a new model CNN_2 to solve such similar problems. Similarly, the weights of CNN_2 have been adopted to solve the CNN_3 model. The benefits of using transfer learning are reducing the training time and decreasing the generalization error.

5) **Optimization:** The proposed FERS problem has been solved by stochastic optimization methods to optimize our CNN model. In this article, we have used the popular first-order gradient-based Adam optimizer of the stochastic objective function. The popular optimization methods used for solving FERS problems are Adagard, SGD, RMSProp, SGD with momentum, AggMo, Demon, Demon CM, DFA, and Adadelata optimization methods. The primary advantages of using an Adam optimizer are that it works well and is suitable for problem solving for large training datasets. “Adam” can handle nonstationary objective functions as in RMSProp while overcoming the sparse gradient issues drawback that appears in RMSProp. “Adam” is favorable compared to other stochastic optimizers. The implementation of “Adam” is straightforward and computationally efficient. It takes little memory.

6) *Scores Fusion*: In the proposed system, three CNN architectures have been proposed. These architectures take images of different sizes as inputs.

$$\text{Sum-rule : } \max_{i \neq j, k=\{1, \dots, 7\}} \{c_{ik} + c_{jk}\} \quad (1)$$

$$\text{Product-rule : } \max_{i \neq j, k=\{1, \dots, 7\}} \{c_{ik} \times c_{jk}\}. \quad (2)$$

So, during recognition of facial expressions on the test sample F , three different classification score vectors are: $s_1 = (c_{11}, c_{12}, \dots, c_{17})$, $s_2 = (c_{21}, c_{22}, \dots, c_{27})$, and $s_3 = (c_{31}, c_{32}, \dots, c_{37})$, where each c_{ij} is the classification score by the CNN_i architecture and for j th expression class. Now these classification scores are fused together using score-level post-classification fusion approaches [20] for increasing the performance of the recognition system. In this article, two score-level fusion techniques: *sum-rule* [see (1)] and *product-rule* [see (2)] have been employed.

IV. EXPERIMENTATION

In this section, the experimentation of the proposed FERS has been performed. We have employed the three most challenging benchmark facial expression databases for random investigation. Each database is randomly divided into 50% of data for training set during experimentation, while the remaining 50% for testing set. These partitioning of datasets has been done ten times, and the average performance has been reported concerning each database.

A. Database Used

The FERS is a well-read field with diversified available databases. For the extensive experimental purpose, we have employed two laboratory-controlled databases such as Karolinska directed emotional faces (KDEF) [21], extended Cohn-Kanade (CK+) [22], and one uncontrolled environment database, i.e., static facial expressions in the wild (SFEW) [23]. These three datasets have been taken based on the license agreements from different universities and institutions for academic research purposes only. The ground truth of these datasets, i.e., the class labels of facial expressions has already been adjusted by their corresponding originating institutions. In the employed CK+ dataset, there are six classes: Anger, Disgust, Fear, Happy, Sad, and Surprise. Here, there are a high number of Disgust class samples, while a very low number of samples in Fear and Sad classes. Hence, some of the samples from these expression classes are also weakly expressed and are unequally distributed in the classes of the CK+ dataset. In particular, CK+ contain only six expressions, compared to KDEF and SFEW. This is because CK has no sample for the expression “neutral” and has the expression “contempt” that has not been considered since not possessed by the other datasets. All the remaining expressions are corresponding in the three datasets. Another difference of CK+ is the unbalanced expressions. In fact, SFEW and KDEF allow to do a balanced training, on the other hand, CK+ has a strong imbalance, with a difference that CK+ was collected in the laboratory, as KDEF. However, SFEW is a spontaneous dataset. It is for this reason that our work, and others in the literature, find

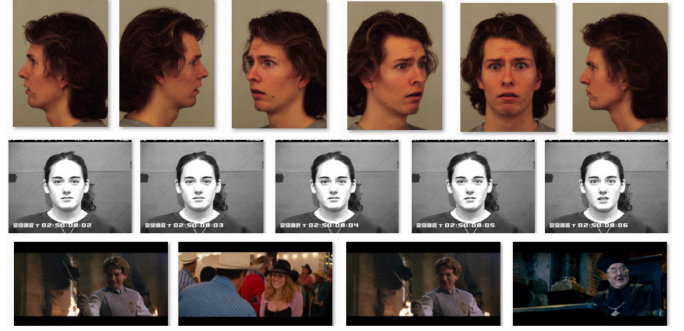


Fig. 7. Samples of images from (from top to bottom) KDEF, CK+, and SFEW databases.

SFEW so competitive. Fig. 7 shows some examples of image samples from KDEF, CK+, and SFEW databases.

B. Results and Discussion

In this section, we will discuss the experimentation of the proposed system. The proposed FERS is implemented using Python 3.7.9 version, Tensorflow 2.3.1 version, Keras 2.4.3 version, CUDA Version: 11.2 and NVIDIA-SMI 460.79 Driver Version in Windows 10 Pro 64-bit, Intel(R) Core(TM)-i7-9700 CPU, 3.30GHz (8 CPU) Processor, 8 GB NVIDIA GeForce RTX 2070 SUPER XLA GPU device with 16 GB RAM. Here, we have employed both gray-scaled and RGB-colored images during experimentation as some databases have RGB images, while some have gray-scaled images. During image preprocessing, from each input image \mathcal{I} , we have detected the face region by applying the methods discussed in Section III-A. Furthermore, the detected face region \mathcal{F} is being normalized to an image \mathcal{F} of size 200×200 pixels. For recognizing the expression classes on the human face, in this article, we have employed deep learning-based approaches where the CNN architectures (see Fig. 6) has been designed that performs both feature computation from the facial region \mathcal{F} followed by classification. In this article, at first, we have started the experiment using the CNN_1 architecture where the input to this system are images of size 48×48 , $n_1 = 48$, i.e., the training samples $\mathcal{F}_{48 \times 48}$ images have been used to train the CNN_1 architecture. In contrast, the performance of the trained CNN_1 model is evaluated using testing samples. During training, learning the parameters is essential, and it depends on two factors: epochs and batch sizes. So, initially, a tradeoff between epochs and batch sizes have been performed using the CNN_1 architecture (see Fig. 4), which is shown in Fig. 8. It is observed that the performance of each database increases gradually with increasing epochs (best performance at nearly 700 epochs), while the batch size is fixed, i.e., 16. Fig. 9 demonstrates the performance of the proposed system due to the CNN_1 architecture (see Fig. 4) with image size $48 \times 48 \times 3$ with respect to KDEF, CK+, and SFEW databases.

1) *Impact of Data Augmentation*: Furthermore, it has also experimented that increasing the training samples, i.e., data augmentation techniques, well learn the parameters of CNN_1 architecture and obtains better performance. Moreover, to adapt to the diversity of training data and avoid overfitting problems, data

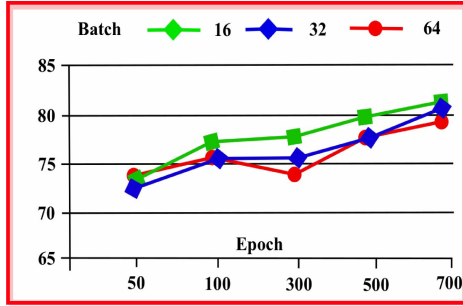


Fig. 8. Accuracy (%) of the proposed system due to the CNN architecture (see Fig. 4) for the tradeoff between epochs and batch sizes for the CK+ database.

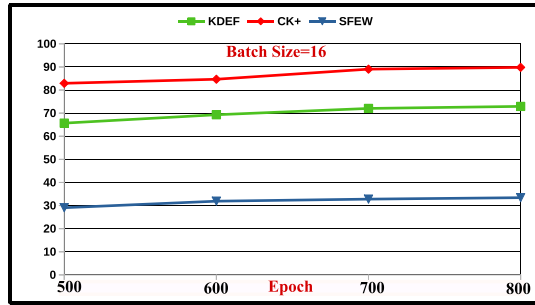


Fig. 9. Accuracy (%) of the proposed system due to the CNN architecture (see Fig. 4) with fixed batch size and number of epochs for KDEF, CK+, and SFEW database.

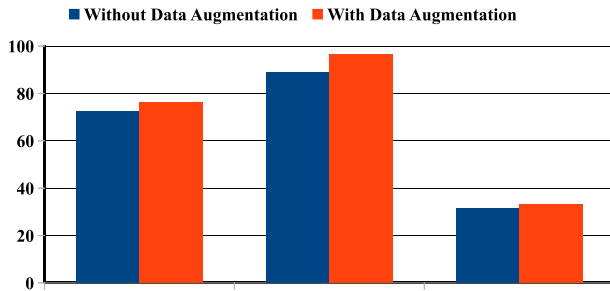


Fig. 10. Improvement of data augmentation on the performance for the proposed system with KDEF, CK+, and SFEW datasets.

augmentation plays an important role. In this article, each sample of training images is horizontally and vertically flipped. Then, affine transformations such as rotation, scaling, zooming, and shearing operations have been performed. Fig. 10 demonstrates the impact of data augmentation techniques on the performance of the CNN₁ architecture with $\mathcal{F}_{48 \times 48}$ images as inputs.

2) *Impact of Progressive Image Resizing*: The techniques of progressive image resizing have already been discussed in Section III-C2. Here, for this purpose, the original image size is $200 \times 200 \times 3$, which is further down samples into $n_1 \times n_1 \times 3$, $2n_1 \times 2n_1 \times 3$, and $4n_1 \times 4n_1 \times 3$, i.e., the CNN₁ architecture (see Fig. 4) has been trained with $\mathcal{F}_{48 \times 48 \times 3}$ images. Then, $\mathcal{F}_{96 \times 96 \times 3}$ images have been used to train the CNN₂ architecture (see Fig. 5). Finally, the CNN₃ architecture has been trained with $\mathcal{F}_{192 \times 192 \times 3}$ images. The purpose behind learning these architectures with the increasing image sizes are as follows:

- 1) the high-resolution images get trained in the network;

TABLE I
DESCRIPTION FOR FACIAL EXPRESSION DATABASE

Database	FE Class	Training	Testing
KDEF	7	1210	1213
CK+	6	663	146
SFEW	7	346	354

TABLE II
EFFECTIVENESS OF THE PROGRESSIVE IMAGE RESIZING ON THE PERFORMANCE OF THE PROPOSED FERS WITH RESPECT TO WITH AND WITHOUT DATA AUGMENTATION TECHNIQUES

Data	Image Size	With data Augmentation	Without data Augmentation
KDEF	48×48	76.28	72.40
	96×96	78.76	75.28
	192×192	81.16	77.19
CK+	48×48	97.26	89.04
	96×96	95.21	93.15
	192×192	96.58	95.20
SFEW	48×48	32.48	31.64
	96×96	34.46	33.26
	192×192	35.89	33.78

- 2) the effect of multiresolution approaches can be introduced in the network such that the texture patterns at the higher level will contribute during learning the parameters;
- 3) the system will bring deeper information that would be better for the hierarchical representations of features.

Hence, the use of progressive image resizing increases the performance of the recognition system and reduces the overfitting problems. The impact of progressive image resizing on the performance of the proposed system has been reported in Table II, and from this table, it has been observed that for KDEF, CK+, and SFEW databases, the proposed FERS has obtained a better performance for $\mathcal{F}_{192 \times 192 \times 3}$ images than $\mathcal{F}_{48 \times 48 \times 3}$ and $\mathcal{F}_{96 \times 96 \times 3}$ images, i.e., the image size $192 \times 192 \times 3$ than for 96×96 and 48×48 image sizes. Moreover, it has also been derived that both progressive images resizing and data augmentation techniques together are very much effective for the CNN models for recognizing the facial expression classes on the facial region.

3) *Impact of Transfer Learning*: In the transfer learning, we have used two different approaches: in the first approach, we freshly train CNN₁, CNN₂, and CNN₃ architectures with the corresponding image sizes, i.e., refreshed models will be used, and in the second approach, we have used the trained CNN₁ model as a pretrained model for CNN₂ such that only upper layers of the CNN₂ model will be trained with $\mathcal{F}_{96 \times 96 \times 3}$ images. Similarly, the upper layers of the CNN₃ architecture will be trained by $\mathcal{F}_{192 \times 192 \times 3}$ images, while for the remaining layers, the weights of the trained CNN₂ model will be used. Hence, for this approach, the pretrained models will be used. Fig. 11 demonstrates the impact of transfer learning approaches on the performance of the proposed FERS. Here, only the version of the CNN₃ model trained with $\mathcal{F}_{192 \times 192 \times 3}$ images has been shown using both progressive images resizing and data augmentation techniques.

TABLE III
PERFORMANCE COMPARISON (IN % OF ACCURACY) OF OUR PROPOSED CNN MODEL WITH OTHER EXISTING STATE-OF-THE-ART METHODS FOR KDEF, CK+, AND SFEW DATABASES

Method	Accuracy (%)	Method	Accuracy (%)	Method	Accuracy (%)
Vgg16 (24)	65.08	Fei (30)	87.74	Vgg16 (24)	24.78
Gabor (25)	53.18	Siqueira (13)	91.40	ResNet50 (34)	24.98
LBP (26)	49.34	Zhou (15)	92.82	Inception-v3 (35)	29.52
HoG (27)	58.29	Vo (14)	93.41	Liu (36)	26.14
GLCM (28)	41.67	Zeng (31)	93.73	Zhang (37)	32.67
Zernik (29)	52.39	Ouellet (32)	94.43	Zhou (15)	31.29
Fei (30)	69.20	Khorrami (33)	95.16	Wang (16)	34.93
Vo (14)	81.05	Proposed	97.83	Proposed	35.91
Proposed	82.63				

(a)

(b)

(c)

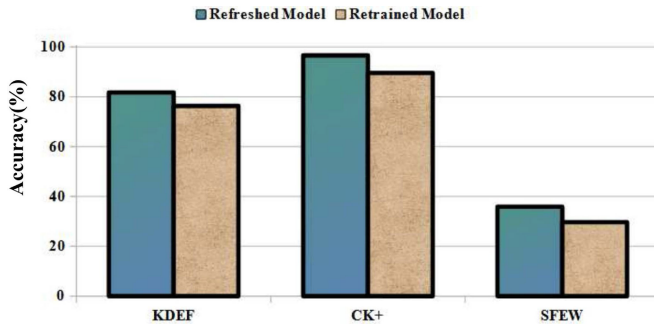


Fig. 11. Impact of transfer learning in terms of improvement in accuracy achieved for the proposed FERS on KDEF, CK+, and SFEW datasets.

TABLE IV
ACCURACY(%) OF THE PROPOSED CNN MODELS DUE TO SCORE-LEVEL FUSION APPROACHES

	CNN_1	CNN_2	CNN_3	Sum-Rule	Product-Rule
KDEF	76.28	78.81	81.16	81.91	82.63
CK+	95.21	96.58	97.26	97.67	97.83
SFEW	32.48	34.46	35.53	35.88	35.91

4) **Impact of Score Fusion:** Score fusion techniques defined in Section III-C have been applied on the classification scores obtained by the CNN architectures, and the results are reported in Table IV. Since both *product-rule* and *sum-rule* are postclassification score level fusion techniques. The *sum-rule* computes the combined class scores, while the *product-rule* derives the final score by multiplying individual scores. Both these rules are based on arithmetic rules that perform better than any other fusion technique. From Table IV, it is observed that for the SFEW and CK+ database the *product-rule* achieves slightly better performance than the *sum-rule*, whereas for the KDEF database this performance is better for the *product-rule*. Here, for the KDEF database, we have sufficient training and testing samples per class. In SFEW and CK+, there are an insufficient amount of training samples. So, this situation is always a limiting factor related to the performance.

5) **Comparison:** To compare the performance of the proposed methodology, the proposed system has achieved a better performance in Table III(a) for the KDEF database, in Table III(b) for the CK+ database, and in Table III(c) for the SFEW database than the competing methods reported in these

tables. Here, for comparison purposes, the competing methods reported in Table III(a)–(c) have been implemented accordingly, and the performance due to these competing methods have been compared with the same training–testing protocol as done for the proposed system. Here, the proposed system has achieved a better performance for each database and it is due to the following facts:

- 1) an efficient end-to-end deep learning-based framework designed for high-level generic features have been adopted for the proposed FERS;
- 2) for reducing the training loss and to overcome overfitting problems, the multiresolution of facial images, data augmentation, progressive image resizing, transfer learning, and fine tuning of parameters techniques are employed for the solution of inadequate training data and bias caused in dataset due to the expressions variation.

V. CONCLUSION

Smart healthcare provides information to patient about traffic congestion or safe route, temperature from weather, nearest healthcare centers, etc. It guides the patient about the emergency ambulance. Smart healthcare gets information constantly from patients through hand-held intelligent devices and notifies the practitioners. It helps to improve the quality of life. Smart cameras have embedded face detection modules used to capture the patient's input face image constantly in a smart home under normal lighting conditions and send it to the cloud server to get regular notification from doctors, caregivers, etc. Here, the implementation of the proposed system is a backbone of quick assistance to the healthcare system and exceptional services to the patients. Generally, there are seven basic emotions on the human face and a mixture of these expressions are being used to predict pain intensities, depression analysis, cognitive analysis, and social ethics problems in humans beings. In this article, the employed databases were captured in real-time scenarios and the methods behind the implementation of the proposed system were robust and efficient. This proposed system was capable to detect and recognize patients' emotions in the real-time healthcare framework. The proposed FERS detected patient face and analyzed it. It recognized human facial expressions/emotions based on the texture of the input visualize information from physiological parameters and transferred them to the cloud server

for further processing. Here, a novel method for an FERS was proposed, which had three components. In the first component, an image preprocessing task was performed, where, from a body silhouette image, the face region was extracted using the facial landmark points. Then, in the second component, the multiresolution images were considered, and some CNN architectures were proposed for each resolution of the detected face region. Here, the images undergo the CNN architectures and were classified into seven different facial expression classes based on learning the CNN models' parameters. To enhance the recognition system's performance and handle the challenging issues of the FERs, some advanced techniques like image augmentation, progressive image resizing, transfer learning, and fine tuning of parameters were employed in the third component. Finally, fusion methods were applied to the best performance of the different CNN models to achieve a better performance than the existing state-of-the-art methods. Extensive experimentation was performed using three benchmark databases, such as KDEF, CK+, and SFEW. The performance of the proposed system was compared with some existing methods for these databases. Comparing the performance of the competing methods and proposed method showed the superiority of the proposed system.

REFERENCES

- [1] C. Guo, P. Tian, and K.-K. R. Choo, "Enabling privacy-assured fog-based data aggregation in E-healthcare systems," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1948–1957, Mar. 2021.
- [2] I. Ahmed, G. Jeon, and F. Piccialli, "A deep learning-based smart healthcare system for patient's discomfort detection at the edge of Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10318–10326, Jul. 2021.
- [3] H. Huang, T. Gong, N. Ye, R. Wang, and Y. Dou, "Private and secured medical data transmission and analysis for wireless sensing healthcare system," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1227–1237, Jun. 2017.
- [4] Z. Xi, Y. Niu, J. Chen, X. Kan, and H. Liu, "Facial expression recognition of industrial internet of things by parallel neural networks combining texture features," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2784–2793, Apr. 2021.
- [5] Z. Lv, Y. Han, A. K. Singh, G. Manogaran, and H. Lv, "Trustworthiness in industrial IoT systems based on artificial intelligence," *IEEE Trans. Ind. Informat.*, vol. 17, no. 2, pp. 1496–1504, Feb. 2021.
- [6] S. Gutta, H. Wechsler, and P. J. Phillips, "Gender and ethnic classification of face images," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 194–199.
- [7] G. Zhang and Y. Wang, "Multimodal 2D and 3D facial ethnicity classification," in *Proc. IEEE 5th Int. Conf. Image Graph.*, 2009, pp. 928–932.
- [8] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 454–459.
- [9] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscek, I. Fasel, and J. Movellan, "Recognizing facial expression: Machine learning and application to spontaneous behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 568–573.
- [10] N. Rose, "Facial expression classification using Gabor and log-Gabor filters," in *Proc. IEEE 7th Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 346–350.
- [11] W. Gu, C. Xiang, Y. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis," *Pattern Recognit.*, vol. 45, no. 1, pp. 80–91, 2012.
- [12] N. Othoudout, M. Daoudi, A. Kacem, L. Ballihi, and S. Berretti, "Dynamic facial expression generation on Hilbert hypersphere with conditional Wasserstein generative adversarial nets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 848–863, Feb. 2022.
- [13] H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5800–5809.
- [14] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131 988–132 001, 2020.
- [15] H. Zhou *et al.*, "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *Proc. Int. Conf. Multimodal Interact.*, 2019, pp. 562–566.
- [16] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, Jan. 2020.
- [17] S. Umer, R. K. Rout, C. Pero, and M. Nappi, "Facial expression recognition with trade-offs between data augmentation and deep learning features," *J. Ambient Intell. Human. Comput.*, vol. 13, pp. 721–735, 2022, doi: 10.1007/s12652-020-02845-8.
- [18] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*. Stamford, CT, USA: Cengage Learning, 2014.
- [19] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.
- [20] S. Umer, B. C. Dhara, and B. Chanda, "Face recognition using fusion of feature learning techniques," *Measurement*, vol. 146, pp. 43–54, 2019.
- [21] M. V. Zavarez, R. F. Berriel, and T. Oliveira-Santos, "Cross-database facial expression recognition based on fine-tuned deep convolutional network," in *Proc. 30th SIBGRAPI Conf. Graph., Patterns Images*, 2017, pp. 405–412.
- [22] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, 2010, pp. 94–101.
- [23] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2106–2112.
- [24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Representations*, San Diego, May. 2015.
- [25] J. Ou, X.-B. Bai, Y. Pei, L. Ma, and W. Liu, "Automatic facial expression recognition using Gabor filter and expression analysis," in *Proc. 2nd Int. Conf. Comput. Model. Simul.*, 2010, pp. 215–218.
- [26] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *Proc. IEEE Int. Conf. Image Process.*, 2005, pp. II-370, doi: 10.1109/ICIP.2005.1530069.
- [27] P. Kumar, S. Happy, and A. Routray, "A real-time robust facial expression recognition system using hog features," in *Proc. IEEE Int. Conf. Comput. Anal. Secur. Trends*, 2016, pp. 289–293.
- [28] G. S. Murty, J. SasiKiran, and V. V. Kumar, "Facial expression recognition based on features derived from the distinct LBP and GLCM," *Int. J. Image, Graph. Signal Process.*, vol. 2, no. 1, pp. 68–77, 2014.
- [29] M. Saaidia, N. Zermi, and M. Ramdani, "Facial expression recognition using neural network trained with Zernike moments," in *Proc. IEEE 4th Int. Conf. Artif. Intell. Appl. Eng. Technol.*, 2014, pp. 187–192.
- [30] Z. Fei, E. Yang, D. Li, S. Butler, W. Ijomah, and H. Zhou, "Combining deep neural network with traditional classifier to recognize facial expressions," in *Proc. IEEE 25th Int. Conf. Automat. Comput.*, 2019, pp. 1–6.
- [31] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie, "Facial expression recognition via learning deep sparse autoencoders," *Neuro-computing*, vol. 273, pp. 643–649, 2018.
- [32] T. Carvalhais and L. Magalhães, "Recognition and use of emotions in games," in *Proc. IEEE Int. Conf. Graph. Interact.*, 2018, pp. 1–8.
- [33] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 19–27.
- [34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
- [35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
- [36] M. Liu, S. Li, S. Shan, and X. Chen, "Au-aware deep networks for facial expression recognition," in *Proc. 10th IEEE Int. Conf. Workshops Automat. Face Gesture Recognit.*, 2013, pp. 1–6.
- [37] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *Int. J. Comput. Vis.*, vol. 126, no. 5, pp. 550–569, 2018.