Journal of Ambient Intelligence and Humanized Computing (2022) 13:721–735
https://doi.org/10.1007/s12652-020-02845-8

ORIGINAL RESEARCH

Facial expression recognition with trade-offs between data augmentation and deep learning features

Saiyed Umer[1] · Ranjeet Kumar Rout[2] · Chiara Pero[3] · Michele Nappi[3]

Received: 16 July 2020 / Accepted: 12 December 2020 / Published online: 7 January 2021

Abstract

A novel facial expression recognition system has been proposed in this paper. The objective of this paper is to recognize the types of expressions in the human face region. The implementation of the proposed system has been divided into four components. In the first component, a region of interest as face detection has been performed from the captured input image. For extracting more distinctive and discriminant features, in the second component, a deep learning-based convolutional neural network architecture has been proposed to perform feature learning tasks for classification purposes to recognize the types of expressions. To enhance the performance of the proposed system, in the third component, some novel data augmentation techniques have been applied to the facial image to enrich the learning parameters of the proposed CNN model. In the fourth component, a trade-off between data augmentation and deep learning features have been performed for fine-tuning the trained CNN model. Extensive experimental results have been demonstrated using three benchmark databases: KDEF (seven expression classes), GENKI-4k (two expression classes), and CK+ (seven expression classes). The performance of the proposed system respect for each database has been well presented and described and finally, these performances have been compared with the existing state-of-the-art methods. The comparison with competing methods shows the superiority of the proposed system.

Keywords Facial expression · Recognition · Data augmentation · Deep learning

1 Introduction

In any inter-personal communication, emotions play an important role Sandbach et al. (2012). There are various expressions in the emotions which may be noticeable or not observable. Recognizing these emotions have wide applications in Computer Vision research areas such as Human Computer interaction, security, animation, psychological patient communication, and faith criminal intent problems. The emotion recognition has been performed by face, EEG (Electroencephalography), text, and speech feature Jaimes and Sebe (2007). Among these features, the human face contributes some interesting and observable expressions such as happiness, sadness, fear, disgust, surprise, anger, and neutral. These expressions are facial expressions and are used in practice than other emotional features. Even capturing facial images are less tangible and invasive than other emotional features. Without any interruption and touching a person, the face image can be captured even he/she walks, talks, and performs activities either at far or moving at a distance Jaimes and Sebe (2007). Proenca et al. Proenca et al. (2016) had proposed an algorithm to estimate human head poses and to infer soft biometric labels based on the 3D morphology of the human head. Abate et al. Abate et al. (2019) had performed a human head pose estimation descriptor that exploits a quad-tree-based representation of facial features. Similarly, the head pose estimation for video surveillance/smart ambient scenarios had been developed by

✉ Chiara Pero
cpero@unisa.it

Saiyed Umer
saiyedumer@gmail.com

Ranjeet Kumar Rout
ranjeetkumarrout@nitsri.net

Michele Nappi
mnappi@unisa.it

[1] Department of Computer Science and Engineering, Aliah University, Kolkata, India

[2] Department of Computer Science and Engineering, National Institute of Technology, Srinagar, India

[3] Department of Computer Science, University of Salerno, Fisciano, Italy

Barra et al. Barra et al. (2020). Figure 1 demonstrates examples of some basic emotions on the human face region.

It has been observed that from the facial expressions much of the clues are obtained from the eye, mouth, and cheek regions while the other parts of the face support these regions for enhancing the expressions on the facial region. The extraction of useful information from the expressive facial regions is very important and needs better techniques for analyzing the facial images in the facial expression recognition (FER) system Tao and Tan (2005). Image analysis is one of the important tasks for any image-based recognition system. In a FER system, the input to the system is a face image, and this image may be gray-scaled ($\mathcal{F}_{M\times N}$) or color image ($\mathcal{F}_{M\times N\times 3}$), suppose $\mathcal{F}$ be an image of $M \times N$ dimension.

There are various techniques that exist for analyzing the facial images as texture and for this transformed, structural and statistical-based approaches have been employed Umer et al. (2019). Apart from transformed and structural approaches, the statistical approaches best perform for analyzing the texture patterns in facial images for the FER (facial expression recognition) system. The techniques under statistical-based approaches are Local Binary Pattern, Histogram of Oriented Gradient, Scale Invariant Feature Transform, Bag of Words, Sparse Representation, and Co-ordinate descend methods Umer et al. (2019).

Recently, the use of deep learning-based approaches with convolutional neural networks (CNN) has tremendous results for solving various Computer Vision problems such as Object detection/recognition, Person verification/identification using biometric features, scene understanding, emotion recognition, etc Khan et al. (2018). The deep learning-based approaches perform multiresolution and multilevel analysis of images and extract more discriminant features from facial texture for the FER system. During unconstrained imaging environments for facial images, the images suffer from various noise artifacts such as motion blur, illumination variations, occlusion by the hair, accessories (scarf, grass, makeup), etc.

Accepting these challenges, in this paper, we have proposed an emotion recognition system using human facial expressions where the input to the system is a gray-scaled image and the system outputs the types of expression on the facial image. The contributions of this paper are as follows:

- A deep learning-based framework using convolutional neural networks (CNN) architecture has been proposed to extract more distinctive and discriminant features that perform a remarkable result for the proposed facial expression recognition system.
- Some novel data augmentation techniques have been proposed to enhance the training capability of the CNN model as well as the task of fine-tuning for hyper-parameters in CNN, which have also been performed to increase the performance of the proposed system.
- The experiments have been performed with trade-offs between data augmentation techniques and deep learning features for the proposed FER system.
- Multiresolution and multilevel image analysis have been experimented for analyzing the facial features of the proposed FER system.

Fig. 1 Basic expressions on the human facial region

The organization of this paper is as follows: Sect. 2 discusses some recent works for facial expression recognition system. The methodology and implementation of the proposed system have been demonstrated and discussed in Sect. 3. Section 4 describes the experimental settings and discusses the results with a comparatively study for the proposed facial expression recognition system. Finally, the paper is concluded in Sect. 5.

2 Related work

In 1971, Paul Ekman Ekman and Friesen (1971) wrote one of the most important works in emotions recognition in which he defined afraid, anger, disgust, happiness, sadness and surprise as the six main emotions. Ekman, afterwards, set the standard for works on emotion recognition through a new publication called "Facial action coding system: a technique for the measurement of facial movement" (FACS Friesen and Ekman (1978)); with the subsequent inclusion of neutral, the basic emotions resulting are seven. Over the past decade numerous facial expression recognition (FER) systems have been presented in literature and a variety of approaches has been adopted. The commonality of these techniques are, respectively, the detection of the facial region and the extraction of features, which can be classified into *geometric features* and *appearance features* Ko (2018). Castrillon et al. Castrillón-Santana et al. (2017) had built a multi-expert gender recognition system using facial region based on various texture feature extraction operators. Based on depth information in facial expression recognition through RGB-D camera structured-light had been proposed by Lee et al. Lee and Lee (2019). A real-time FER for affecting identification using a multi-dimensional SVM classifier had been proposed by Meshach et al. Meshach et al. (2020). An image filter based subspace learning method had been developed by Yan et al. Yan et al. (2020) to derive the feature representation from the low-resolution face images for the FER system. Sadeghi and Raie Sadeghi and Raie (2019)

had proposed histogram distance-based metric learning for the FER system. Makhmudkhujaev et al. Makhmudkhujaev et al. (2019) had developed facial expression recognition including noise and positional variations by proposing a technique for Local Prominent Directional Pattern descriptors from facial images. A local directional maximum edge descriptor had been developed by Maheswari et al. Maheswari et al. (2020) for the FER system.

With the recent trend of deep learning methods, and more specifically convolutional neural networks for image classification, several techniques have developed yielding considerable cutting-edge results. Fan and Tjahjadi Fan and Tjahjadi (2019) had proposed a FER system where the dynamic deep learning features and handcrafted features have been fused together to perform discriminant feature learning for recognizing human facial expressions. A region of interest guided convolutional neural network (CNN) architecture based on deep learning had been proposed by Sun et al. Sun et al. (2020) for the FER system by exploiting the relationships among the extracted region of interest within facial areas. A Region-based Convolutional Fusion Network had been proposed by Ye et al. Ye et al. (2019) for the FER system. Zhang et al. Zhang et al. (2020) had proposed the FER system based on the deep convolution long short-term memory networks using the double-channel weighted mixture. A multi-scale convolutional neural network model had been designed with atrous convolutions kernels for recognizing expressions in the facial image Lai et al. (2020). Yu et al. Yu et al. (2020) had developed the multi-task global-local network to extract the global spatial appearance features and local fine-grained temporal features for the FER system.

Ji et al. Ji et al. (2019) had proposed a fused network for recognizing facial expressions in cross databases where the network consisted of an Intra-category common feature representation channel and an Inter-category distinction feature representation channel for FER system. A comprehensive comparison of four different strategies for the design of the ensemble of CNNs for the FER system had been developed by Renda et al. Renda et al. (2019). The gravitational search algorithm based optimized deep learning model with a diverse set of features had also been employed for facial expression recognition Alenazy and Alqahtani (2020). Hence inspired by these methods, in this paper a deep learning-based framework has been designed for the proposed FER system which has been discussed and described at the next section.

3 Proposed system

In this section, we have discussed our proposed facial expression recognition (FER) system. The input to this system is an image $\mathcal{I}_{\mathcal{M}\times\mathcal{N}}$ which contains the face region. The objective of the proposed system is to predict the types of expression (happiness, sadness, fear, disgust, surprise, anger, and neutral) on the face region of the input image $\mathcal{I}$. So, during implementation, the proposed system has been divided into three components: (a) face detection, (b) feature learning for classification through deep learning-based approach, (c) finding the effectiveness with trade-offs between data augmentation techniques and deep learning features and (d) prediction for unknown types of expression on the facial image. The block diagram of the proposed FER system has been demonstrated in Fig. 2.

3.1 Face detection

In this work, the images suffer from various noise artifacts such as motion blur, pose invariant (profile to frontal), illumination variations, occlusion by the hair, occlusion by accessories, variations in light, etc De Marsico et al. (2012). Accepting these challenges, the images are processed for detecting the facial region from an input image $\mathcal{I}$. For this purpose, a Tree-Structured Part Model (TSPM) Zhu and Ramanan (2012) method has been employed which performs searching for topological changes due to different viewpoints of a face region in the captured image. The topological changes have been obtained by computing the image gradient using a histogram of oriented gradient and then the topological changes are considered as templates that predict the landmark points over the facial region. Hence, for the frontal face, it computes sixty-eight landmark points while for profile faces, thirty-nine landmark points are obtained. These landmark points are used to compute the facial region $\mathcal{F}$ from the input image $\mathcal{I}$. The preprocessing steps of face detection for the

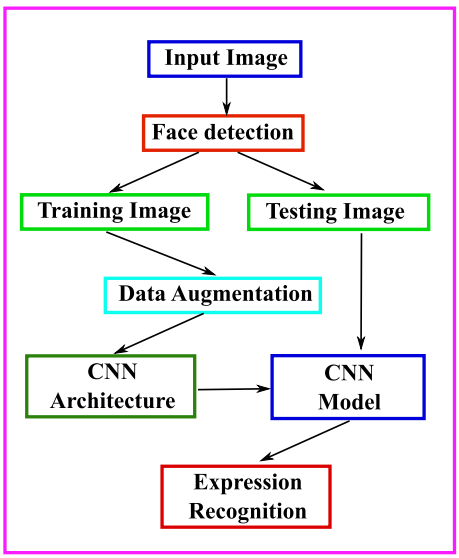

Fig. 2 Block diagram of the proposed FER system

proposed FER system have been shown in Fig. 3. Some of the detected face regions with different poses and expressions employed in the proposed FER system have also been shown in Fig. 4.

3.2 Image augmentation

It is a technique in machine learning that synthetically increases the amount of data by applying some transformation methods on the existing data Hernández-García and König (2018). There are several benefits of data augmentation which are (a) performing overtraining the convolution neural networks (CNN) with weight decay and dropout facilities, (b) solving the over-fitting problems, and (c) fine-tuning the hyper-parameters in the CNN for increasing the performance. The image augmentation techniques have a great impact on image recognition problems as these techniques increase the number of samples corresponds to each image without changing the visual quality and image fidelity of the images Perez and Wang (2017). Hence, the image augmentation techniques enhance the learning parameters of CNN architecture during training the CNN model using augmented images and obtain the best model for the recognition or classification problems. The employed data augmentation techniques for the proposed system, are as follows:

- *Bilateral filtering* This filtering technique smooths the image $\mathcal{F}$ and reduces the noise artifacts by obtaining Gaussian blurred technique Paris et al. (2009) while preserving the gradients in $\mathcal{F}$. The Bilateral Filtering is mathematically defined as:

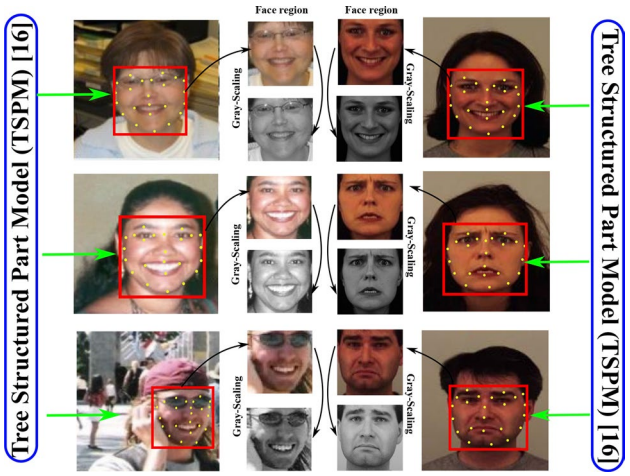

Fig. 3 Face detection steps in the proposed FER system

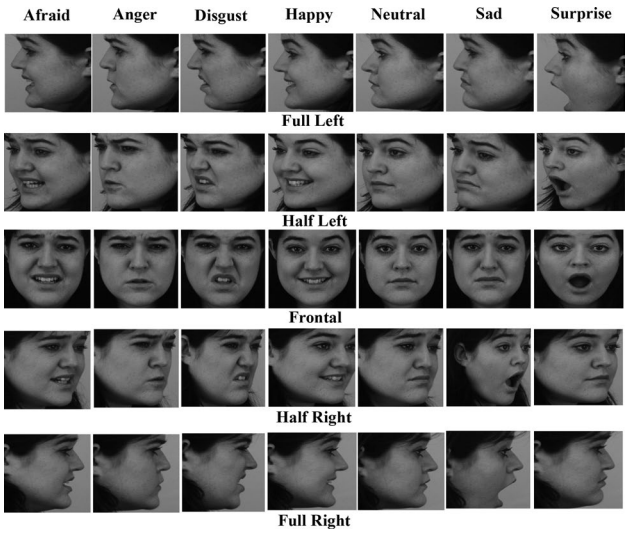

Fig. 4 Detected facial region for the proposed FER system

$$\begin{aligned} \mathcal{G}(\mathcal{F}_i) &= \sum_{j \in S} G_\sigma(||i-j||)\mathcal{F}_j \\ \mathcal{A}_1(\mathcal{F}_i) &= \frac{1}{W_i} \sum_{j \in S} G_{\sigma_u}(||i-j||) G_{\sigma_v}(||\mathcal{F}_i - \mathcal{F}_j||)\mathcal{F}_j \end{aligned} \tag{1}$$

where G_{σ_u} is the width of the neighbor, G_{σ_v} be the minimum amplitude of an edge (smaller the value of G_{σ_v} sharper the edge), $\mathcal{G}(\mathcal{F}_i)$ be the Gaussian blur and $\frac{1}{W_i}$ be the normalization factor. The term defined as $\sum_{j \in S} \mathcal{G}_{\sigma_u}(||i-j||)F_j$ denotes the linear filtering with the Gaussian blur $G_{\sigma_v}(||\mathcal{I}_i - \mathcal{F}_j||)\mathcal{F}_j$. Fig. 5 shows the effects of Bilateral filtering on the facial expressive images. Hence for each image $\mathcal{F}$, one Bilateral filtered image ($\mathcal{A}_1$) is obtained.

- *Unsharp filter* It is a sharpening operator applied to the image $\mathcal{F}$, which enhances the high-frequency components in the image $\mathcal{F}$. The working principle of Unsharpe filtering Polesel et al. (2000) is that first the image is smoothed by median filtering $\mathcal{F}_{Smooth}$ and then the second derivative operator Laplacian of Gaussian (LoG) is being applied on $\mathcal{F}_{Smooth}$ to obtain $\mathcal{F}_{LoG}$. Finally, the

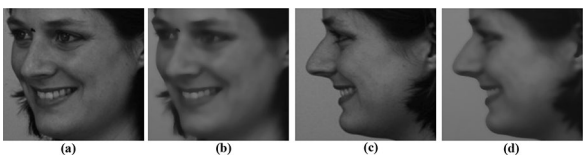

Fig. 5 Effects of bilateral filtering on the facial images: **a**, **c** are original images and **b**, **d** are bilateral filtered images of **a** and **c**, respectively

sharpened image $\mathcal{A}_2$ is obtained by subtracting $\mathcal{F}_{Smooth}$ from $\mathcal{F}$ i.e. $\mathcal{A}_2 = \mathcal{F} - \mathcal{F}_{Smooth}$. Figure 6 shows an example for Unsharp filtered images for $\mathcal{F}$. Hence, for each image $\mathcal{F}$, the corresponding augmented Unsharp filtered image $\mathcal{A}_2$ is obtained.

- *Sharpening filter* Image sharpening Pardo-Igúzquiza et al. (2006) is a method which is very sensitive to edges with finding the fine details within the image $\mathcal{F}$. This method enhances the high-frequency components such as edges, ridges, contours, and blobs in $\mathcal{F}$. Mathematically, the image sharpening is defined as $s_{ij} = x_{ij} + \lambda \times w(x_{ij})$, where x_{ij} be the pixel of original image $\mathcal{F}$ at (i, j) position, s_{ij} be the pixel of filtered image, w be the filter mask, and $\lambda \geq$ be the tuning parameter. The choice of λ depends on the grade of sharpness. The size of w may be 3×3 or 5×5 or 7×7 or 9×9 matrix, where this filter mask is convoluted with image $\mathcal{F}$ to obtain the corresponding filtered image. The image sharpening methods employed in this work are as follows:

$$w_1 = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{bmatrix}, \mathcal{A}_3 = w_1 \otimes \mathcal{F} \tag{2}$$

$$w_2 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \mathcal{A}_4 = w_2 \otimes \mathcal{F} \tag{3}$$

$$w_3 = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 10 & -1 \\ -1 & -1 & -1 \end{bmatrix}, \mathcal{A}_5 = w_3 \otimes \mathcal{F} \tag{4}$$

$$w_4(i,j) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2+j^2}{2\sigma^2}}, \mathcal{A}_6 = w_4 \in \mathbb{R}^{7\times 7} \otimes \mathcal{F} \tag{5}$$

$$w_5 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 6 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \mathcal{A}_7 = w_5 \otimes \mathcal{F} \tag{6}$$

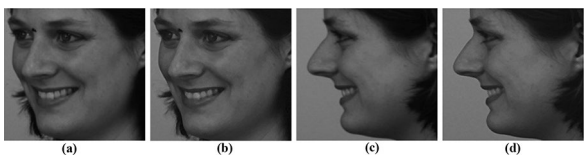

Fig. 6 Effects of unsharp filtering on the facial images: **a**, **c** are original images and **b**, **d** are unsharp filtered images of **a** and **c**, respectively

$$w_6 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 7 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \mathcal{A}_8 = w_6 \otimes \mathcal{F} \tag{7}$$

$$w_7 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 8 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \mathcal{A}_9 = w_7 \otimes \mathcal{F} \tag{8}$$

$$w_8 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -7 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \mathcal{A}_{10} = w_8 \otimes \mathcal{F} \tag{9}$$

$$w_9 = \frac{1}{8} \begin{bmatrix} -1 & -1 & -1 & -1 & -1 \\ -1 & 2 & 2 & 2 & -1 \\ -1 & 2 & 8 & 2 & -1 \\ -1 & 2 & 2 & 2 & -1 \\ -1 & -1 & -1 & -1 & -1 \end{bmatrix}, \mathcal{A}_{11} = w_9 \otimes \mathcal{F} \tag{10}$$

Figure 7 shows the enhanced images using Eqs. (2), (3), (4), (5), (6), (7), (8), (9) and (10). Hence, for each image $\mathcal{F}, \mathcal{A}_3, \cdots, \mathcal{A}_{11}$ augmented images are obtained.

- *Image rotation* It is one of the Affine transformation Asano et al. (2007) scheme applied on the image $\mathcal{F}$. The objective of the image rotation technique is to rotate image respect to an angle θ about a point p through which rotation has been performed. In this work, the image rotation is being performed on $\mathcal{F}$ respect to the angle $\theta = 20^o$ in a clockwise direction around its center point. Hence, for the image $\mathcal{F}$, one rotated image $\mathcal{A}_{12}$ is obtained (Fig. 8).
- *Image scaling* It is another affine transformation scheme applied on image F Wu et al. (2015) that refers to the resizing of F. In image scaling, the scaling operator performs the geometric transformation that may be used for zooming or shrinking the size of the image. In this work, image zooming is performed by the bilinear interpolation with some scaling factor. Figure 9 shows the effects of image scaling on the facial expressive images. Hence, for the image $\mathcal{F}$, one scaling image $\mathcal{A}_{13}$ is obtained.
- *Shear mapping* The shearing mapping technique Tanter et al. (2009) is a linear mapping that displaces each pixel position of image $\mathcal{F}$ in fixed direction respect to a line that is parallel to that direction passes through the origin. In this work, image shear mapping is being performed on image $\mathcal{F}$ and obtains the sheared image $\mathcal{A}_{14}$. Figure 10 shows the effects of shear mapping on the facial expressive images. Hence, for the image $\mathcal{F}$, one sheared mapping image $\mathcal{A}_{14}$ is obtained.
- *Image zooming* Image zooming Battiato et al. (2002) is the process of magnifying the image centered region

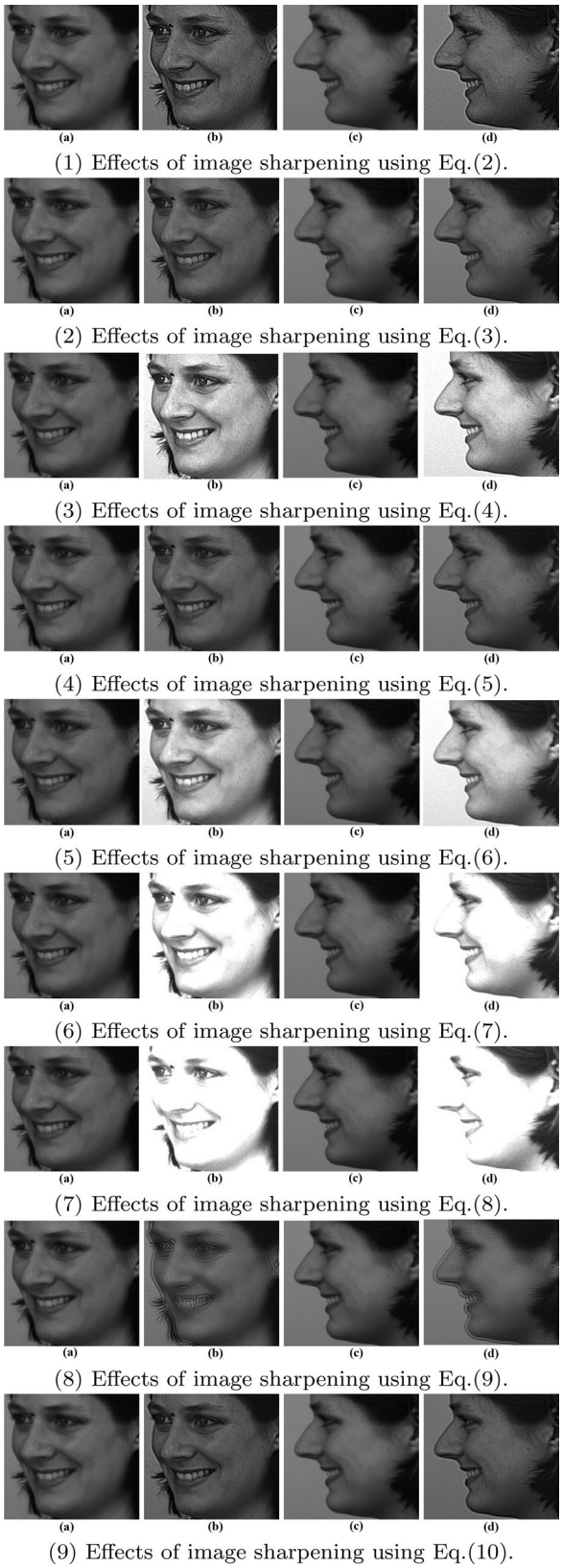

Fig. 7 **a**, **c** Original images and **b**, **d** are filtered images of **a** and **c**, respectively

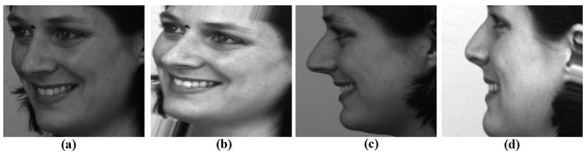

Fig. 8 Effects of image rotation on the facial images: **a**, **c** are original images and **b**, **d** are rotated images of **a** and **c**, respectively

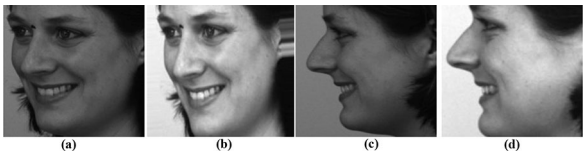

Fig. 9 Effects of image scaling on the facial images: **a**, **c** are original images and **b**, **d** are rotated images of **a** and **c**, respectively

Fig. 10 Effects of image shearing on the facial images: **a**, **c** are original images and **b**, **d** are rotated images of **a** and **c**, respectively

Fig. 11 Effects of image zooming on the facial images: **a**, **c** are original images and **b**, **d** are rotated images of **a** and **c**, respectively

and is used to extract the region based on the edge pixels using the pixel replication with the nearest neighbor interpolation method. Figure 11 demonstrates the effects of the image zooming phenomenon applied to the facial expressive images. Hence, for the image $\mathcal{F}$, one zoomed image $\mathcal{A}_{15}$ is obtained.

- *Image filling* Image filling De Queiroz (2000) is the process of adding either some rows or some columns in the image $\mathcal{F}$ by the region filling algorithms of digital image processing. Figure 12 demonstrates the effects of image filling applied on the facial expressive images. Hence, for the image $\mathcal{F}$, one image-filled $\mathcal{A}_{16}$ is obtained.
- *Image horizontal flip* It is a special case of image rotation Iliyasu et al. (2012) about the rotation of the image at an angle 180^o by applying the rotation affine transformation on the image $\mathcal{F}$. Figure 13 demonstrates the effects of image horizontal flipping applied on the facial expressive

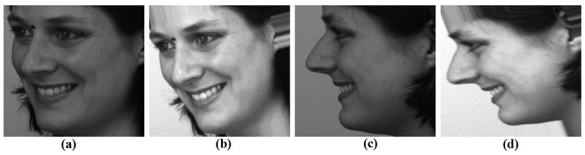

Fig. 12 Effects of image filling on the facial images: **a**, **c** are original images and **b**, **d** are rotated images of **a** and **c**, respectively

Fig. 13 Effects of image horizontal flip on the facial images: **a**, **c** are original images and **b**, **d** are rotated images of **a** and **c**, respectively

images. Hence, for the image $\mathcal{F}$, one horizontal flipped image $\mathcal{A}_{17}$ is obtained.

The above-discussed data augmentation techniques are based on the concepts and theories of digital image processing. The objective of these data augmentation techniques are as follows:

- The edge enhancement techniques such as Biletral filtering (Fig. 5), Unsharp mask filtering (Fig. 6), edge enhancement filtering (Fig. 7(9)), image sharpening (Fig. 7(1)), ⋯, Fig. 7(8), extract the silent features such as edges, blobs, ridges, contours in the facial regions. These edge enhancement techniques not only preserve the edge information but also increase the tone mapping and contrast stretching during multi-scale image decomposition for feature extraction.
- Due to the unconstrained imaging environment, the various noise artifacts have been introduced in the image and these reduce the texture information from the images. So, image smoothing (Fig. 5) technique has been employed to suppress the noise artifacts such that coarse to fine details can be addressed during image enhancement.
- Image rotation (Fig. 8) the technique provides the rotation invariant properties to images for learning the parameters in the proposed CNN architecture. Image flipping (Fig. 12) either horizontally or vertically about a line not only increases the training samples but also enhances the classification mode more robust and effective for the unknown test samples.
- Image scaling (Fig. 9), image zooming (Fig. 11), shear mapping for images (Fig. 10) techniques contribute the pixelated properties to the images where increasing the image size will add some new pixels to the images while the reduction of image size removes some unwanted pixels. Hence, the image scaling data augmentation technique is beneficial for learning the parameters of the CNN architecture.

Image sharpening, image smoothing and affine transformations (rotation, flipping, reflection, shearing, scaling) thus increase the training samples. These data augmentation techniques help the CNN architecture for fine learning the parameters of the deep neural networks to prepare the classification mode more robust and effective while increasing the performance for the challenging unknown test samples.

3.3 Expression recognition

In this work, the expression recognition is based on the images. In most of the pattern recognition problems, the first step is the detection of a region of interest from the given image. Then more distinctive and discriminant features have been extracted in the form of feature vector from the detected image region in the second step. In the third step, the collection of feature vectors from the considered training images, are used to learn the classifier to build a classification model. This classification model is used to predict the class types for the unknown test samples Branson et al. (2010). So, the classification or recognition problems are based on the techniques that well define and describe the patterns which depend on the employed feature extraction methods. For the image and video-based recognition problems, there are several structural and statistical-based approaches exist that have grained the great success for computer vision problems.

For the current state-of-the-art recognition problems in Computer Vision, these structural and statistical techniques have limited performance due to the data captured in the unconstrained environments. To extract the information from these data to obtain better performance than the existing methods, the deep learning-based approaches have been employed Mollahosseini et al. (2016). In this work, a convolutional neural network (CNN) architecture based on deep learning approaches has been proposed. This convolutional neural network architecture is based on the core building blocks of convolutional layer Vedaldi and Zisserman (2016), max-pooling layer Vedaldi and Zisserman (2016), fully connected layer Targ et al. (2016) and dense layer Targ et al. (2016). Additionally, for improving the stability, performance, and speed of the CNN, the batch normalization layer has also been included in the proposed CNN architecture.

In the convolutional layer, the input to the layer is an image that is convoluted with several distinct kernels (filter banks) and then computes the convoluted image (feature-map) respect to each filter. The convolution of feature maps has been performed by the heavy lifting of computations and it increases complexity with increasing the image size

and the number of convolutional layers. Here the parameters are the weights adjusted in the filter banks. The Maxpooling layers are used to reduce the computational overheads by decreasing the number of parameters in the network. In this layer, most of the time 2×2 filter is considered over each feature map (obtained from the previous layer) with a stride of 2-downsample and obtain the maximum or minimum or average value are 4-numbers along horizontally and then vertically.

A fully connected layer Targ et al. (2016) considers all the features from the previous layers and transforms them into a 1-dimensional vector. The dense layer is the same as a fully connected layer but the difference is that linear operations are being performed in the dense layer Huang et al. (2017). In the linear operations, every input is connected to every output and generates the probability scores at the output layer using the Softmax activation function.

So, using the above discussed deep neural network layers with their concepts and theories, in this work we have proposed a convolutional neural network (CNN) architecture which is based on the convolution layer, max-pooling layer, fully connected layer, and dense layer. The rough architecture of the proposed CNN has been shown in Fig. 14. From this figure, it has been observed that there are six blocks (where each block contains convolution layers, Batch Normalization Ioffe (2017), Activation, Maxpooling, and Dropout layers), two fully connected layers with three dense layers, where the last dense layer outcomes the probability scores for seven facial expression class types. For a better understanding of the proposed expression recognition system, the description of the CNN architecture with employed layers, the output shape of the feature maps, and parameters employed at each layer have been demonstrated in Table 1. From this table, it has been observed that there are several Convolutional layers, Maxpooling layers, Batch Normalization, Activation, and Dense layers. Here the Batch Normalization is just after the Convolutional layer.

The Batch Normalization technique Ioffe (2017) normalizes the batch of data from the previous layer by subtracting the batch mean and divide the batch by its standard deviation. So, after processing, the batch normalization adds, two trainable parameters i.e. mean (β) and standard deviation (γ) to each layer. For, Activation function, Rectified Linear Unit (ReLU) Xu et al. (2015) function has been employed. The activation functions derive the output of a neural network. Each activation function is attached with each neuron and fires the output based on the input to the neuron in the network to predict its relevance for the model. Here ReLU activation function is defined as $g(z) = max(0, z)$, z be the derived input computed after convolution and batch normalization and it is input to the activation function. Dropout technique Srivastava et al. (2014) has been used to ignore randomly selected neurons during training on the forward pass such that the weights will be not updated on those neurons during backward pass. The use of the dropout technique prevents overfitting problems and provides a way for combining efficiently the predictions of various different neural nets. Finally, the softmax activation function $\left(Softmax(y_i) = \frac{e_{y_i}}{\sum_{i=0}^{7} e_{y_i}}\right)$ has been employed on the outputs of last dense layer to get the probability scores for the facial expression classes. Hence, the employed CNN architecture has been employed for the proposed facial expression recognition system.

4 Experiments

In this section, we have performed experiments for the proposed FER system. For experimental purposes, three benchmark face databases have been employed. Here each database is partitioned randomly with 50% of data into the training set while 50% into the testing set. This partition has been performed ten times and the averaged performance has been reported herewith these experiments.

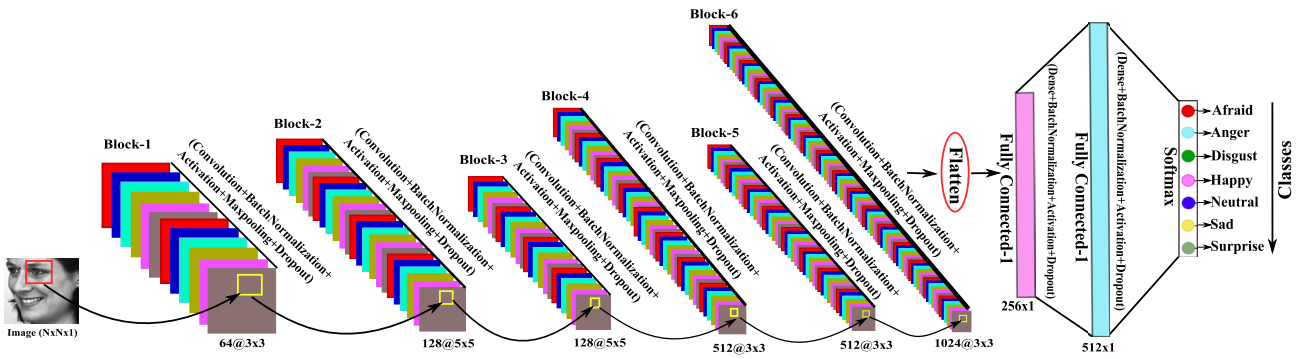

Fig. 14 Proposed CNN architecture for the FER system

Table 1 Description of proposed CNN architecture with employed layers, output shapes and parameters at each layer where the size of input image is 128×128

Layer	Output shape	Image size	Parameters	Layer	Output shape	Image size	Parameters
Block-1				Block-4			
Convolution2D (3x3)@64	(n,n,64)	(128,128,64)	((3x3)+1)x64 =640	Convolution2D (3x3)x512	$(n_3, n_3, 512)$	(16,16,512)	((3x3x128)+1) x 512 =590336
Batch Normalization	(n,n,64)	(128,128,64)	4x64=256	Batch Normalization	$(n_3, n_3, 512)$	(16,16,512)	4x512=2048
Activation Relu	(n,n,64)	(128,128,64)	0	Activation Relu	$(n_3, n_3, 512)$	(16,16,512)	0
Maxpooling2D (2x2)	$(n_1, n_1, 64) n_1 = n/2$	(64,64,64)	0	Maxpooling2D (2x2)	$(n_4, n_4, 512)$ $n_4 = n_3/2$	(8,8,512)	0
Dropout	$(n_1, n_1, 64)$	(64,64,64)	0	Dropout	$(n_4, n_4, 512)$	(8,8,512)	0
Block-2				Block-5			
Convolution2D (5x5)@128	$(n_1, n_1, 128)$	(64,64,128)	((5x5x64)+1) x 128 =204928	Convolution2D (3x3)@512	$(n_4, n_4, 512)$	(8,8,512)	((3x3x512)+1) x 512= 2359808
Batch Normalization	$(n_1, n_1, 128)$	(64,64,128)	4x128=512	Batch Normalization	$(n_4, n_4, 512)$	(8,8,512)	2048
Activation Relu	$(n_1, n_1, 128)$	(64, 64, 128)	0	Activation Relu	$(n_4, n_4.512)$	(8,8,512)	0
Maxpooling2D (2x2)	$(n_2, n_2, 128)$ $n_2 = n_1/2$	(32, 32, 128)	0	Maxpooling2D (2x2)	$(n_5, n_5, 512)$ $n_5 = n_4/2$	(4,4,512)	0
Dropout	$(n_2, n_2, 128)$	(32, 32, 128)	0	Dropout	$(n_5, n_5, 512)$	(4,4,512)	0
Block-3				Block-6			
Convolution2D (5x5)@128	$(n_2, n_2, 128)$	(32,32,128)	((5x5x128)+1) x 128 =409728	Convolution2D (3x3)@1024	$(n_5, n_5, 1024)$	(4,4,1024)	((3x3x512)+1) x 1024= 47,19,616
Batch Normalization	$(n_2, n_2, 128)$	(32,32,128)	4x128=512	Batch Normalization	$(n_5, n_5, 1024)$	(4,4,1024)	4,096
Activation Relu	$(n_2, n_2, 128)$	(32,32,128)	0	Activation Relu	$(n_5, n_5, 1024)$	(4,4,1024)	0
Maxpooling2D (2x2)	$(n_3, n_3, 128)$ $n_3 = n_2/2$	(16,16,128)	0	Maxpooling2D (2x2)	$(n_6, n_6, 1024)$ $n_6 = n_5/2$	(2,2,1024)	0
Dropout	$(n_3, n_3, 128)$	(16,16,128)	0	Dropout	$(n_6, n_6, 1024)$	(2,2,1024)	0
Layer		Output shape		Image size		Parameter	
Flatten		$(1, n_6 \times n_6 \times 1024)$		(1,4096)		0	
Dense		(1,256)		(1,256)		(4096+1)x256= 1048832	
Batch Normalization		(1,256)		(1,256)		1024	
Activation Relu		(1,256)		(1,256)		0	
Dropout		(1,256)		(1,256)		0	
Dense		(1,512)		(1,512)		(256+1)x512= 131584	
Batch Normalization		(1,512)		(1,512)		2048	
Activation Relu		(1,512)		(1,512)		0	
Dropout		(1,512)		(1,512)		0	
Dense		(1,7)		(1,7)		(512+1)x7= 3591	
Total Parameters for Image ($\mathcal{I}$) size (128 x 128)						9481607	

[a]The input image that is convoluted with several distinct kernels in the convolution layer

[b]The computational overheads have been reduced in the Maxpooling layer.

[c]The conversion of features in 1-dimensional vector is performed by the fully connected layer.

[d]The ReLU is an activation function which is defined as $g(z) = max(0, z)$.

[e]The Softmax is an activation function which is used by output layer.

[f]The Batch Normalization technique normalizes the batch of data from the previous layer

4.1 Database used

The databases used for the proposed system are Karolinska directed emotional faces (KDEF) Lundqvist et al. (1998), GENKI-4K Jain and Crowley (2013) and Cohn-Kanade Extended (CK+) Lucey et al. (2010).

The KDEF database is composed of 4900 emotion images of human facial expressions about seven classes (Happy, Anger, Sad, Surprise, Neutral, Disgust, and Fear). These images are the collection of seventy different individuals (thirty-five female and thirty-five male). During experimentation, we have employed 1222 images of seven classes for training dataset while 1225 images of seven classes for testing dataset. The images of KDEF database is shown in Fig. 15. The GENKI-4K database is composed of 4000 facial images which are labeled as 'Happy' and 'Non-Happy'. During experimentation, 2000 images are randomly selected as training set while the remaining 2000 images have been considered for the testing set. Figure 16 demonstrates the images from GENKI-4K database. The CK+ database is composed of 593 short videos from 123 subjects with varying illumination and age. Each video has 10 to 60 frames, the starting frame is neutral while the last frame contains high intensity expressive image. For this experiment we have selected 981 images of seven classes (Happy, Anger, Sad, Surprise, Contempt, Disgust, and Fear) where 500 images for training purposes while remaining 481 images for testing purpose. Figure 17 demonstrates some images from CK+ database. Here each database has been experimented individually and the results are reported accordingly.

4.2 Result and discussion

Here the proposed facial expression recognition (FER) system has been implemented in Python on Ubuntu 16.04 LTS O/S version with Intel Core i7 processor 3.20GHz and 32GB RAM. During implementation we have used several packages from Keras Chollet et al. (2015) and Theano to build the proposed CNN architecture. Accuracy is the metric adopted for evaluate the proposed system. Informally, it represents the ratio of the number of correct predictions to

Fig. 16 Examples of images from GENKI-4K database

the total number of samples of the model. During face detection, we have extracted the facial region $\mathcal{F}_{\mathcal{N}\times\mathcal{N}}$ using TSPM model from the input image $\mathcal{I}$. Then each detected facial region $\mathcal{F}_{\mathcal{N}\times\mathcal{N}}$ undergoes to the proposed data augmentation techniques. Hence, from each image $\mathcal{F}_{\mathcal{N}\times\mathcal{N}}$, $\mathcal{A}_1, \cdots, \mathcal{A}_{17}$ augmented images have been derived using techniques defined in Sect. 3.2.

The data augmentation techniques have been applied only on the training images to learn the parameters of the proposed CNN architecture described and defined in Sect. 3.3. In the proposed FER system, for expression recognition, we have performed experiments for the effectiveness of multiresolution and multiscaling images with varying sizes such as $\mathcal{F}_{48\times48}$, $\mathcal{F}_{64\times64}$, $\mathcal{F}_{96\times96}$ and $\mathcal{F}_{128\times128}$. In training the CNN architecture, the batch size and number of epochs are very much important. The batch size refers to the number of training images utilized at one iteration whereas the epochs refer to the number of times that the learning algorithm

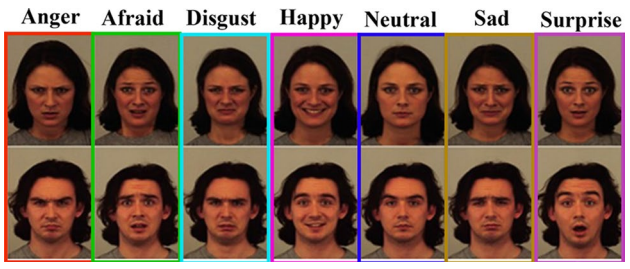

Fig. 15 Examples of images from KDEF database

Fig. 17 Examples of images from CK+ database

will work on the entire training dataset. So, in this work we have utilized the Mini-Batch Gradient Descent technique Ioffe and Szegedy (2015) with varying batch sizes such as $\{20, 30, 40\}$ and number of epochs are $\{50, 100, 200, 500\}$. The performance of the proposed FER with the trade-off between batch sizes and the number of epochs without data augmented training images have been shown in Fig. 18. Here Fig. 18a–d show the performance due to the multiresolution images $\mathcal{F}_{48\times48}$, $\mathcal{F}_{64\times64}$, $\mathcal{F}_{96\times96}$ and $\mathcal{F}_{128\times128}$ respectively.

Here the performances are shown for test images with respective image sizes. From Fig. 18, it has been observed that increasing the number of epochs may lead to increase in performance but increasing the batch size may not correlate with the increased performance. So, for further experiments we have employed 40 batch size and 500 epochs. Now we have applied the data augmentation techniques on $\mathcal{F}_{128\times128}$ images for both KDEF (7-class problems) and GENKI-4k (2-class problems) database. The purpose of this experiment is to show the effectiveness of the proposed data augmentation techniques on the FER system where these techniques have been applied in 10 segments $(D_0, \cdots, D_9)$, where D_0 segment shows the performance without applying data augmentation, D_1 for 'Bilateral Filtering', D_2 for 'Bilateral + Unsharp Filtering', D_3 for 'Bilateral + Unsharp + Sharpening Filter', D_4 for 'Bilateral + Unsharp + Sharpening + Image rotation', similarly D_5, D_6, D_7, D_8 are so on and finally, D_9 for 'Bilateral + Unsharp + Sharpening + Image rotation + Image scaling + Shear mapping + Image zooming + Image filling + Image horizontal flip' augmentation techniques (described at Sect. 3.2). Figure 19 demonstrates the effectiveness of the employed data augmentation techniques on the proposed FER system.

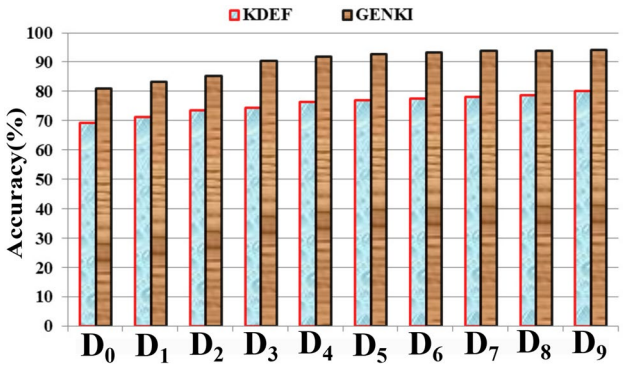

Fig. 19 Effectiveness of the employed data augmentation techniques in the proposed FER system

From Fig. 19 it has been observed that the proposed FER system has obtained better performance for D_9 segment of data augmentation. So, for further experiments, we have employed a D_9 segment of data augmentation on the FER

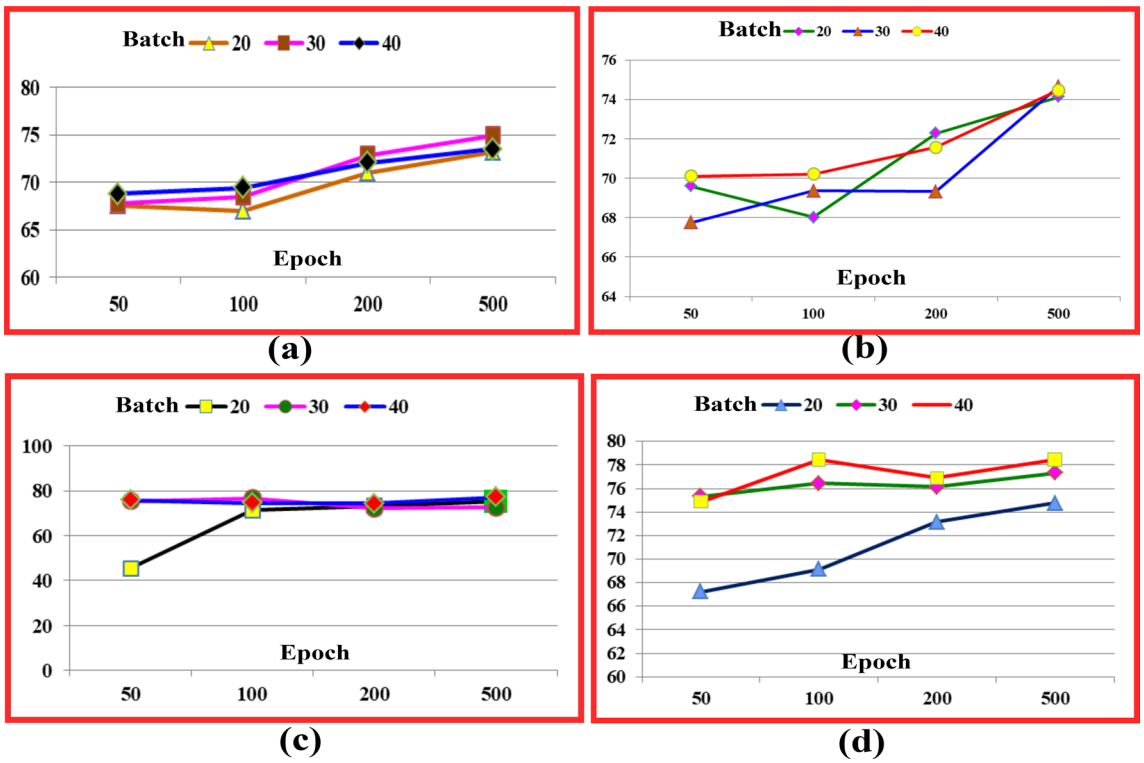

Fig. 18 Performance comparison for the trade-off between epoch and batch size employed in the proposed CNN model

Table 2 Performance of the proposed FER system in accuracy (%) due to varying image sizes

Image-size	KDEF	GENKI	CK+
$\mathcal{F}_{48\times48}$	73.67	80.34	87.23
$\mathcal{F}_{64\times64}$	75.89	84.78	91.87
$\mathcal{F}_{96\times96}$	78.92	89.45	94.35
$\mathcal{F}_{128\times128}$	82.79	94.33	97.69

system. Moreover, it has also been observed that due to data augmentation techniques, there is a great enhancement of improved performance of the proposed FER system. Further using the full data augmentation techniques (D_9), the learned CNN model is fine-tuned to reach to some better performance for all KDEF, GENKI-4K and CK+ databases and these performance are shown in Table 2 using multiresolution images $\mathcal{F}_{48\times48}$, $\mathcal{F}_{64\times64}$, $\mathcal{F}_{96\times96}$ and $\mathcal{F}_{128\times128}$. Here the performance is for test images with respective image sizes. From the Table 2, it has been observed that the proposed FER system has attained outstanding performance for image size $\mathcal{F}_{128\times128}$ and these are 82.79% for KDEF, 94.33% for GEMKI-4k and 97.69% for CK+ database.

For comparison purposes, we have compared the performance of the proposed system with some existing state-of-the-art methods such as Rao et al. Rao et al. (2015), Zavare et al. Zavarez et al. (2017) and Sun et al. Sun et al. (2017) under the same training-testing protocols. These methods had obtained the performance only for 980 frontal images for the KDEF database. We have copied the performance from the respective papers of these methods and have reported the performance in Table 3. Moreover, we have also compared the performance of the proposed system with some well-known CNN models such as Vgg16 Simonyan and Zisserman (2014), Res-Net50 Szegedy et al. (2016), and Inception-v3 Szegedy et al. (2017) under the same training-testing protocols and we have fine-tuned these models using $\mathcal{F}_{128\times128}$ images with 40 batch size and 500 epochs. From Table 3 it has been observed that the proposed FER system outperforms the other existing methods reported in Table 3 for KDEF database.

For the GENKI-4K database, the proposed system has been compared with some existing state-of-the-art methods such as An et al. An et al. (2015), Zhang et al. Zhang et al. (2015), and Gao et al. Gao et al. (2016) and the performance of these methods are copied from the respective paper of these competing methods and reported in Table 4. Here also We have compared the performance of the proposed system with Vgg16 Simonyan and Zisserman (2014), ResNet50 Szegedy et al. (2016), and Inception-v3 Szegedy et al. (2017) CNN models which have been fine-tuned using $\mathcal{F}_{128\times128}$ images with 40 batch size and 500 epochs for two expression classes. The performance of these methods is shown in Table 4. From Table 4 it has been observed that the proposed system outperforms the other exiting methods in Table 4. For the CK+ database,

Table 3 Performance comparison of the proposed system with some existing state of the art methods for KDEF database

Method	Accuracy (%)	Remarks
Vgg16 Simonyan and Zisserman (2014)	65.08	Images used (980), Expression class (7), Train/Test Split
ResNet50 Szegedy et al. (2016)	72.32	Images used (980), Expression class (7), Train/Test Split
Inception-v3 Szegedy et al. (2017)	75.04	Images used (980), Expression class (7), Train/Test Split
Rao et al. Rao et al. (2015)	74.05	Images used (720), Expression class (6)
		Images type (Frontal), 10-fold cross validation
Zavare et al. Zavarez et al. (2017)	72.55	Images used (980), Expression class (7)
		Images type (Frontal), 10-fold cross validation
Sun et al. Sun et al. (2017)	82.24	Images used (490), Expression class (7)
Proposed	83.43	Images used (980), Proposed CNN for 7 expression classes

Table 4 Performance comparison of the proposed system with some existing state of the art methods for GENKI-4k database

Method	Accuracy (%)	Remarks
Vgg16 Simonyan and Zisserman (2014)	72.08	VGG16 CNN for 7 expression classes
ResNet50 Szegedy et al. (2016)	82.30	ResNet 50 CNN for 7 expression classes
Inception-v3 Szegedy et al. (2017)	85.38	Inception-v3 CNN for 7 expression classes
An et al. An et al. (2015)	88.50	Feature (HOG), Classifier (ELM)
Zhang et al. Zhang et al. (2015)	94.21	Feature (CNN), Classifier (Softmax)
Gao et al. Gao et al. (2016)	94.33	Feature (Ensemble), Classifier (Ensemble)
Proposed	94.67	Proposed CNN for 7 expression classes

Table 5 Performance comparison of the proposed system with some existing state of the art methods for CK+ database

Method	Accuracy (%)	Remarks
Sun et al. Sun et al. (2020)	94.67	Images used (510), Expression class (7), k-fold cross-validation
Zhanga et al. Zhang et al. (2020)	97.50	VGG16 network with 10-fold cross validation
ResNet50 Szegedy et al. (2016)	92.45	Images used (981), Expression class (7), Train/Test Split
Inception-v3 Szegedy et al. (2017)	94.23	Images used (981), Expression class (7), Train/Test Split
Proposed	97.69	Images used (981), Proposed CNN for 7 expression classes

the performance has been compared with Sun et al. Sun et al. (2020), Zhanga et al. Zhang et al. (2020), ResNet50 Szegedy et al. (2016), Inception-v3 Szegedy et al. (2017) methods under the same training testing protocols. From this comparison (shown in Table 5), it has been observed that the proposed FER system has obtained better performance for CK+ database.

5 Conclusion

This paper represents a facial expression recognition system. The images captured in the unconstrained environment have been considered here. During implementation from each captured image, a face region is detected. Then to speed up the processes, the color facial image is transformed into its gray-scaled image. The purpose of employing these images is to recognize the types of expression in the images and for this a convolutional neural network (CNN) based framework is designed. To enhance the learning parameters of the proposed CNN model, some data augmentation techniques have been applied to the training images. The proposed data augmentation techniques not only solve the problem of overfitting by fine-tuning the hyper-parameters but also increases the performance of the proposed system. The trade-off between employed data augmentation and deep learning-based features affect the FER system to accept more challenging to recognize the expressions in the unknown test samples. Finally, the proposed system has been tested with three benchmark databases: KDEF, GENKI-4k, and CK+. The performance due to these databases shows the superiority of the proposed FER system compared with the state-of-the-art methods of respect to these databases.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest. The funding agency had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

Abate AF, Barra P, Bisogni C, Nappi M, Ricciardi S (2019) Near real-time three axis head pose estimation without training. IEEE Access 7:64256–64265

Alenazy WM, Alqahtani AS (2020) Gravitational search algorithm based optimized deep learning model with diverse set of features for facial expression recognition. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-020-02235-0

An L, Yang S, Bhanu B (2015) Efficient smile detection by extreme learning machine. Neurocomputing 149:354–363

Asano T, Bitou S, Motoki M, Usui N (2007) In-place algorithm for image rotation. In: International symposium on algorithms and computation. Springer, pp 704–715

Barra P, Barra S, Bisogni C, De Marsico M, Nappi M (2020) Web-shaped model for head pose estimation: An approach for best exemplar selection. IEEE Trans Image Process 29:5457–5468

Battiato S, Gallo G, Stanco F (2002) A locally adaptive zooming algorithm for digital images. Image Vis Comput 20(11):805–812

Branson S, Wah C, Schroff F, Babenko B, Welinder P, Perona P, Belongie S (2010) Visual recognition with humans in the loop. In: European conference on computer vision. Springer, pp 438–451

Castrillón-Santana M, De Marsico M, Nappi M, Riccio D (2017) Meg: texture operators for multi-expert gender classification. Comput Vis Image Underst 156:4–18

Chollet F (2015) Keras: Deep learning library for theano and tensorflow. https://keras.io/

De Marsico M, Nappi M, Riccio D, Wechsler H (2012) Robust face recognition for uncontrolled pose and illumination changes. IEEE Trans Syst Man Cybern Syst 43(1):149–163

De Queiroz RL (2000) On data filling algorithms for MRC layers. In: Proceedings 2000 international conference on image processing (Cat. No. 00CH37101), vol 2. IEEE, pp 586–589

Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Personal Soc Psychol 17(2):124

Fan Xijian, Tjahjadi Tardi (2019) Fusing dynamic deep learned features and handcrafted features for facial expression recognition. J Vis Commun Image Represent 65:102659

Friesen E, Ekman P (1978) Facial action coding system: a technique for the measurement of facial movement. Palo Alto 3

Gao Y, Liu H, Pingping W, Wang C (2016) A new descriptor of gradients self-similarity for smile detection in unconstrained scenarios. Neurocomputing 174:1077–1086

Hernández-García A, König P (2018) Further advantages of data augmentation on convolutional neural networks. In: International conference on artificial neural networks. Springer, pp 95–103

Huang G, Liu Z, Der Maaten LV, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708

Iliyasu AM, Le PQ, Dong F, Hirota K (2012) Watermarking and authentication of quantum images based on restricted geometric transformations. Inf Sci 186(1):126–149

Ioffe Sergey (2017) Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In: Advances in neural information processing systems, pp 1945–1953

Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv :1502.03167

Jaimes Alejandro, Sebe Nicu (2007) Multimodal human–computer interaction: a survey. Comput Vis Image Underst 108(1–2):116–134

Jain V, Crowley JL (2013) Smile detection using multi-scale gaussian derivatives

Ji Y, Hu Y, Yang Y, Shen F, Shen HT (2019) Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network. Neurocomputing 333:231–239

Khan S, Rahmani H, Shah SAA, Bennamoun M (2018) A guide to convolutional neural networks for computer vision. Synth Lect Comput Vis 8(1):1–207

Ko BC (2018) A brief review of facial emotion recognition based on visual information. Sensors 18(2):401

Lai Z, Chen R, Jia J, Qian Y (2020) Real-time micro-expression recognition based on resnet and atrous convolutions. J Ambient Intell Humaniz Comput 1–12

Lee K, Lee EC (2019) Comparison of facial expression recognition performance according to the use of depth information of structured-light type RGB-D camera. J Ambient Intell Humaniz Comput 1–17

Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended Cohn–Canade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops. IEEE, pp 94–101

Lundqvist D, Flykt A, Öhman A (1998) The karolinska directed emotional faces (kdef). CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, vol 91, no 630, p 2

Maheswari VU, Varaprasad G, Viswanadha RS (2020) Local directional maximum edge patterns for facial expression recognition. J Ambient Intell Humaniz Comput 1–9

Makhmudkhujaev F, Abdullah-Al-Wadud M, Iqbal MTB, Ryu B, Chae O (2019) Facial expression recognition with local prominent directional pattern. Signal Process Image Commun 74:1–12

Meshach WT, Hemajothi S, Anita EAM (2020) Real-time facial expression recognition for affect identification using multi-dimensional SVM. J Ambient Intell Humaniz Comput 1–11

Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 1–10

Pardo-Igúzquiza E, Chica-Olmo M, Atkinson PM (2006) Downscaling cokriging for image sharpening. Remote Sens Environ 102(1–2):86–98

Paris Sylvain, Kornprobst Pierre, Tumblin Jack, Frédo D (2009) Theory and applications. Bilateral filtering. Now Publishers Inc., Norwell

Perez L, Wang J (2017) The effectiveness of data augmentation in image classification using deep learning. arXiv:1712.04621

Polesel A, Ramponi G, Mathews VJ (2000) Image enhancement via adaptive unsharp masking. IEEE Trans Image Process 9(3):505–510

Proenca H, Neves JC, Barra S, Marques T, Moreno JC (2016) Joint head pose/soft label estimation for human recognition in-the-wild. IEEE Trans Pattern Anal Mach Intell 38(12):2444–2456

Rao Q, Qu X, Mao Q, Zhan Y (2015) Multi-pose facial expression recognition based on surf boosting. In: 2015 international conference on affective computing and intelligent interaction (ACII). IEEE, pp 630–635

Renda A, Barsacchi M, Bechini A, Marcelloni F (2019) Comparing ensemble strategies for deep learning: an application to facial expression recognition. Expert Syst Appl 136:1–11

Sadeghi H, Raie AA (2019) Histogram distance metric learning for facial expression recognition. J Vis Commun Image Represent 62:152–165

Sandbach G, Zafeiriou S, Pantic M, Yin L (2012) Static and dynamic 3d facial expression recognition: a comprehensive survey. Image Vis Comput 30(10):683–697

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556

Srivastava Nitish, Hinton Geoffrey, Krizhevsky Alex, Sutskever Ilya, Salakhutdinov Ruslan (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958

Sun Xiao, Xia Pingping, Zhang Luming, Shao Ling (2020) A ROI-guided deep architecture for robust facial expressions recognition. Inf Sci 522:35–48

Sun Zhe, Zheng-Ping Hu, Wang Meng, Zhao Shu-Huan (2017) Discriminative feature learning-based pixel difference representation for facial expression recognition. IET Comput Vis 11(8):675–682

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826

Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence

Tanter Mickaël, Touboul David, Gennisson Jean-Luc, Bercoff Jeremy, Fink Mathias (2009) High-resolution quantitative imaging of cornea elasticity using supersonic shear imaging. IEEE Trans Med Imaging 28(12):1881–1893

Tao J, Tan T (2005) Affective computing: a review. In: International conference on affective computing and intelligent interaction. Springer, pp 981–995

Targ S, Almeida D, Lyman K (2016) Resnet in resnet: generalizing residual architectures. arXiv:1603.08029

Umer S, Dhara BC, Chanda B (2019) Face recognition using fusion of feature learning techniques. Measurement 146:43–54

Vedaldi A, Zisserman A (2016) VGG convolutional neural networks practical. Department of Engineering Science, University of Oxford, p 66

Wu R, Yan S, Yi S, Dang Q, Sun G (2015) Deep image: scaling up image recognition 7(8). arXiv:1501.02876

Xu B, Wang N, Chen T, Li M (2015) Empirical evaluation of rectified activations in convolutional network. arXiv:1505.00853

Yan Yan, Zhang Zizhao, Chen Si, Wang Hanzi (2020) Low-resolution facial expression recognition: a filter learning perspective. Signal Process 169:107370

Ye Yingsheng, Zhang Xingming, Lin Yubei, Wang Haoxiang (2019) Facial expression recognition via region-based convolutional fusion network. J Vis Commun Image Represent 62:1–11

Mingjing Yu, Zheng Huicheng, Peng Zhifeng, Dong Jiayu, Heran Du (2020) Facial expression recognition based on a multi-task global-local network. Pattern Recognit Lett 131:166–171

Zavarez MV, Berriel RF, Oliveira-Santos T (2017) Cross-database facial expression recognition based on fine-tuned deep convolutional network. In: 2017 30th SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). IEEE, pp 405–412

Zhang Hepeng, Huang Bin, Tian Guohui (2020) Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. Pattern Recognit Lett 131:128–134

Zhang Kaihao, Huang Yongzhen, Wu Hong, Wang Liang (2015) Facial smile detection based on deep learning features. In: 2015 3rd

IAPR Asian conference on pattern recognition (ACPR). IEEE, pp 534–538
Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 2879–2886