Journal of Ambient Intelligence and Humanized Computing (2023) 14:7747–7759
https://doi.org/10.1007/s12652-023-04586-w

ORIGINAL RESEARCH

A novel modular deep fully convolutional network for efficient low resolution facial expression recognition

Walaa Aly[1,2] · Ahmed I. Shahin[3] · Saleh Aly[1,2]

Received: 9 February 2022 / Accepted: 27 February 2023 / Published online: 15 March 2023

Abstract

Facial expression is one of the most important and natural way to express human feelings. Although deep convolutional neural networks have improved the performance of facial expression recognition (FER) systems, recognizing facial expressions from low resolution images is still a challenging task for real-time applications. A new modular deep fully convolutional neural network is designed to tackle this problem. The proposed network is composed of four modules namely, feature extraction (FE), residual spatial-channel attention (RSCA), atrous spatial pyramid pooling (ASPP), and classification module. The prominent facial regions relevant to facial expressions are extracted using FE module and then strengthened using RSCA and ASPP modules. Finally, a classification module using convolutional layers with adjusted stride parameter values is employed instead of fully connected layers to significantly reduce the number of learnable parameters. Experimental results using CK+, RAF-DB, and SFEW 2.0 datasets show that our proposed method achieves improved accuracies of 99.9%, 84.96%, and 53.0% at image resolutions of 32×32, 48×48, and 26×26, respectively.

Keywords Low resolution facial expression recognition · Deep learning · Fully Convolutional network · Atrous spatial pyramid pooling · Residual spatial-channel attention module

1 Introduction

Facial expression is considered one of the most critical manners to express emotions for human communications (Lozano-Monasor et al. 2017; Umer et al. 2021; Maheswari et al. 2021; Sikkandar and Thiyagarajan 2021; Abdullah and Abdulazeez 2021). Most facial expression recognition (FER) methods classify human emotions into one of six basic emotion classes, including anger, disgust, fear, happiness, sadness, and surprise, plus a neutral emotion class (Li and Deng 2020). Automatic facial expression recognition systems have various applications in many fields such as human-computer interaction, psychology analysis, animation, customer service, and student education monitoring (Whitehill et al. 2014; Zhao et al. 2016).

In recent years, due to the rapid development of deep learning algorithms in computer vision (Połap 2019), many researchers employed them in FER to achieve state-of-the-art results. Deep convolution neural networks (DCNN) have a powerful representation capability and outperform other classical machine learning methods based on hand-crafted features. Due to the richness of human facial features, developing appropriate and light CNN architectures plays a vital role to improve the performance of FER (Li and Deng 2020; Shao and Qian 2019; Jain et al. 2019; Rao et al. 2021; Ma et al. 2021).

Low resolution facial images is considered one of the critical FER problems that have not been resolved well. In real-life environments, surveillance cameras capture facial images with long distances, which generate various low-resolution faces as shown in Fig. 1. One severe problem

✉ Saleh Aly
s.haridy@mu.edu.sa

Walaa Aly
wa.ali@mu.edu.sa

Ahmed I. Shahin
a.shahin@mu.edu.sa

[1] Department of Information Technology, College of Computer and Information Sciences, Majmaah University, Al-Majmaah 11952, Saudi Arabia

[2] Department of Electrical Engineering, Faculty of Engineering, Aswan University, Aswan 81542, Egypt

[3] Department of Natural and Applied Sciences, Community College, Majmaah University, Al-Majmaah 11952, Saudi Arabia

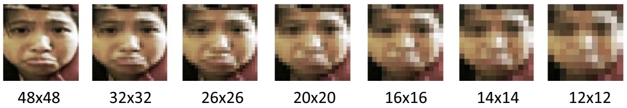

Fig. 1 Example of face images at various resolutions ranging from 48×48 to 12×12

caused by LR images is losing crucial visual information, which reduces FER systems' performance. Therefore, recognizing facial expressions using LR images is a highly challenging task.

Typical resolution of facial images in most existing databases such as CK+, RAF-DB, and FER2013 is lower than other benchmark image datasets such as the ImageNet dataset. Therefore most existing FER techniques re-scale the LR images into a canonical size (Mollahosseini et al. 2016) in order to be able to utilize the pre-trained CNN models. However, image scale normalization causes facial image blurring, and hence important facial details may be lost. The existing methods which tackled LR facial expression recognition problem could be divided into two types; super resolution (SR) (Shao and Cheng 2021), and image representation (Yan et al. 2020). Super resolution algorithms are used to recover high-resolution images from LR images (Chen et al. 2021). Shao and Cheng (2021) developed an edge-aware feedback CNN(E-FCNN) to recognize tiny facial images. Image representation techniques are used to extract discriminative features from LR images (Khan et al. 2013). Yan et al. (2020) proposed a method based on image filtering to recognize LR facial expressions. The performance of FER systems depends on the quality of the super resolution algorithms. However, our proposed method directly apply LR images without re-scaling images or applying the heavy computation super resolution algorithms.

In this paper, a novel modular deep convolutional neural network is developed to address low-resolution facial expression recognition problem. The architecture of the proposed deep CNN consists of four modules; feature extraction module, residual spatial-channel attention module, Atrous spatial pyramid pooling module, and classification module. The feature extraction module is used to extract deep facial features with adaptive stride size to compensate the information loss due to low resolution input image. Relevant facial regions and features are captured by the residual spatial channel attention module. Then, the multi-scale features that contain facial context information are captured using Atrous spatial pyramid pooling module. Finally, the classification module using three convolutional layers is employed to identify the class of the input image. The proposed model depends only on convolution operation without relying on any fully connected layers to reduce the number of learnable parameters in the network. Both spatial and context features are combined to efficiently improve image representation of low resolution facial images. To overcome the insufficient qualitative data and data imbalance problem, data augmentation and the combination of different training image datasets are employed. Experiments show that our proposed method achieves high accuracy using three public datasets in comparison with other state-of-the-art methods.

The contributions of this paper can be summarized as follows:

1. A novel lightweight fully-deep convolutional neural network framework is designed based on the combination of four modules to improve the performance of low resolution facial expression recognition systems.
2. A new effective residual spatial-channel attention module is introduced to jointly focus on discriminating features at relevant facial expression regions.
3. The Atrous spatial pyramid pooling module is utilized to capture the facial context features from relevant regions.
4. The new classification module is utilized to keep the local facial spatial information instead of the traditional fully-connected layers.
5. The proposed FER framework is evaluated using three public facial expression datasets, including one lab controlled dataset (CK+) and two in-the-wild datasets (RAF-DB and SFEW 2.0).

The rest of the paper is organized as follows: Sect. 2 reviews the recent works related to facial expression recognition, Sect. 3 explains the details of the proposed framework modules. Section 4 reports the experimental results. Finally, Sect. 4.7 draws the conclusion.

2 Related works

Recently, facial expression recognition become an interesting research topic due to its important applications. Several studies are introduced to recognize facial expressions (Chu et al. 2016; Zhang et al. 2017; Li and Deng 2020). FER systems can be divided into three stages: face detection, feature extraction, and classification. Several survey research papers (Sariyanidi et al. 2014; Kumari et al. 2015; Ben et al. 2021) explained the approaches used in each stage.

A few research works have studied the effect of employing low resolution images on FER systems. For example, Khan et al. (2013) introduced a novel facial descriptor model for facial feature analysis by combining the pyramid of local binary pattern (PLBP) approach with the LBP descriptor which improved the performance of the FER for LR facial images. In their work, they first extracted PLBP features from salient regions in facial images. Then they divided the extracted features into two groups. Finally, they used five different classifiers to classify the facial expression according to these groups. Their approach had better accuracy in LR images as they used a new face descriptor model. Liu et al. (2020) proposed a new facial expression restoration model using an improved graph convolutional network (IGCN). Other approaches on LR images focus on the facial image representation aspect. For example, Yan et al. (2020) introduced a novel technique of image filtering based on subspace learning (IFSL) to recognize LR facial expression images. To solve the LR problem, they used the discriminative image filter (DIF) based on two-class linear discriminate analysis approach that reduced inta-class variations and maximized inter-class variations between LR facial images.

Most researches in LR facial images have employed super-resolution methods to produce high-resolution images from LR images. For example, Cheng et al. (2017) developed a novel encoder-decoder model for FER of videos for LR facial expression recognition. Shao and Cheng (2021) proposed an edge-aware feedback CNN (E-FCNN) to recognize tiny facial expressions. In their method, the tiny facial images were enlarged using super resolution network, then the facial expressions were recognized using classification network. Nan et al. (2021) proposed a multi scale super resolution method for LR images recognize facial expressions without restoring high resolution facial images. In their approach, features were extracted from LR facial images using a pre-trained model. The low-resolution features were then converted to their corresponding high-resolution features using the generator network. The generator network was trained on features from LR images and their corresponding high resolution images. This approach improved the accuracy of LR facial images by narrowing the distance between high resolution images and LR images.

3 Proposed facial expression recognition system

The proposed Facial Expression Recognition (FER) system consists of two major stages: face detection and alignment stage, and facial expression classification stage. The block diagram of the proposed facial expression recognition system is illustrated in Fig. 2. The detailed structure of each stage is explained in the following.

3.1 Face detection an alignment stage

In the first stage, the input image captured from a charged-coupled device (CCD) surveillance camera is processed to detect the facial region (Luo et al. 2020) and localize essential facial feature points such as mouth corners, nose tips, and eye centers. Then, the facial landmark points are utilized to align the cropped face images into a canonical face where eye centers lie in a fixed location. Multi-Task Cascaded Convolutional Networks (MTCNN) (Zhang et al. 2016) is considered as one of the most successful deep learning models used to detect and align small face images. The architecture of MTCNN framework contains three convolutional network stages that detect faces and localize five important landmark locations: two eye centers, a nose tip, and two mouth corners. MTCNN integrates both detection and alignment tasks using multi-task learning. The first stage uses a shallow fully CNN to quickly produce candidate windows. The second stage refines the proposed candidate windows through a more complex CNN. Furthermore, the third stage uses a third CNN, more complex than first and second networks, to further refine the result and output facial landmark positions.

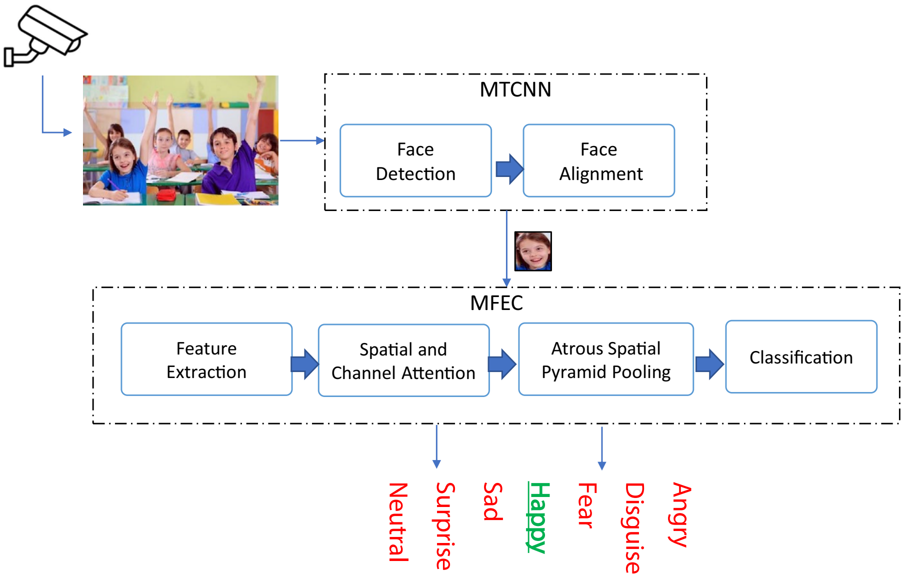

Fig. 2 Proposed framework for low resolution facial expression recognition system

3.2 Proposed modular facial expression classification

The proposed modular facial expression classification (MFEC) stage classifies the extracted discriminative features from low-resolution input facial images. Most of the commonly used CNNs such as VGG16 and ResNet50 assume that the input image has a high resolution of 224 × 224 which is not applicable for face images captured at a distance from surveillance cameras. Furthermore, the captured low-resolution facial images lose much important information, hindering the utilization of existing deep convolutional neural network architectures as a backbone feature extraction. The proposed MFEC is designed to preserve all spatial features extracted from the low-resolution input facial images. Since local and context features are two critical cues to efficiently classifying human emotions from facial images, proposed method exploits the advantages of combining channel-spatial attention and Atrous spatial pyramid pooling to achieve this target. The proposed modular facial expression classification stage is a fully convolutional network (FCN) consisting of four main modules, namely, feature extraction (FE) module, residual spatial-channel attention (RSCA) module, atrous spatial pyramid pooling (ASPP) module, and classification (CL) module. The block diagram of the proposed method is illustrated in Fig. 3. We explain in more detail these modules in the following.

3.2.1 Feature extraction module

The feature extraction module consists of three convolutional blocks to extract facial features at multiple levels. Each block contains two convolution layers of size 3 × 3 followed by the Batch Normalization (BN) and the Exponential Linear unit (ELU) (Clevert et al. 2016) activation function. The number of filters in the first, second and third blocks are set to C, $4C$ and $8C$, respectively. Where C is the base number of channels which is set to 64 in this paper. After each of the first and second convolutional blocks, a max pooling operations using 2 × 2 window size and variable stride parameter values is employed. Most existing CNN architectures utilize pooling operations with fixed stride sizes to reduce the spatial dimensionality of the input feature maps. However, in the proposed low-resolution facial expression module, the stride parameter value is adapted according to the resolution of the input face image. Using stride with value 1 preserves all spatial information of the low-resolution feature maps for further processing in the following modules. The first convolutional layer applies the convolution operation on the input facial image of $h \times w$ image size. The output features from the first block capture the low-level features passed to the second block to obtain middle feature maps. The final convolutional block in the feature extraction module generates high-level features fed to the next attention module. The resolution of the feature maps is controlled by the stride value of the max-poling layer in which stride 1 is employed

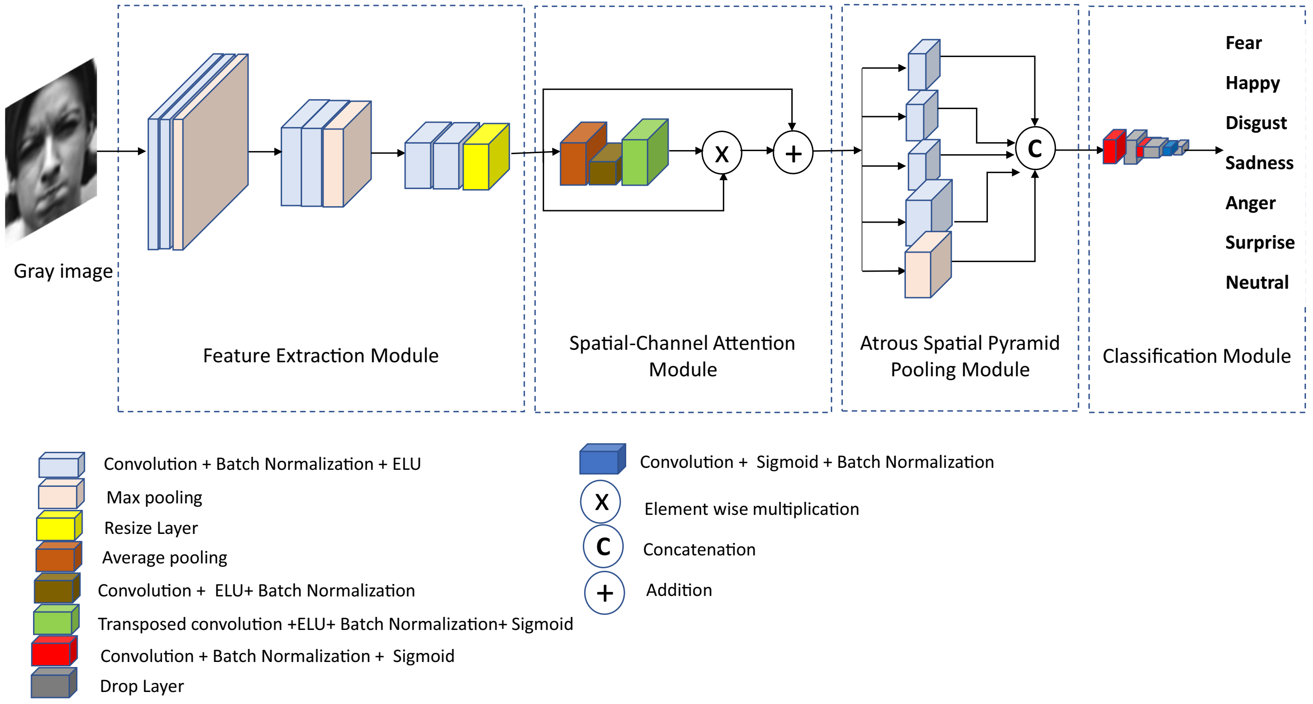
Gray image
Feature Extraction Module
X
+
Spatial-Channel Attention Module
C
Atrous Spatial Pyramid Pooling Module
Fear
Happy
Disgust
Sadness
Anger
Surprise
Neutral
Classification Module
Convolution + Batch Normalization + ELU
Max pooling
Resize Layer
Average pooling
Convolution + ELU+ Batch Normalization
Transposed convolution +ELU+ Batch Normalization+ Sigmoid
Convolution + Batch Normalization + Sigmoid
Drop Layer
Convolution + Sigmoid + Batch Normalization
X Element wise multiplication
C Concatenation
+ Addition

Fig. 3 Proposed modular low resolution facial expression classification (MFEC) network

for images with resolution less than 32 × 32 pixels and stride 2 used for 48 × 48 image resolution. The Batch Normalization (BN) operation followed by the Exponential Linear Unit (ELU) activation function is employed after each convolutional operation to speed up the learning process and achieve better generalization.

The ELU activation function is utilized instead of the commonly used ReLU in order to keep the identity of positive response values and utilize non-zero values for negative ones. Mathematically, ELU activation function can be defined as follows:

$$ELU(x) = \begin{cases} x & x \geq 0 \\ \alpha(exp(x) - 1) & x < 0 \end{cases} \tag{1}$$

where α is a hyper-parameter utilized to control the value of negative responses. The values computed by ELU activation function push the mean of the activation to be closed to zero which allows faster training and improve the generalization of the network. The dimension of the feature maps for a given input image I of size $h \times w$ generated from the last convolutional block is $h/(s1 \times s2) \times w/(s1 \times s2) \times 8C$, where $s1$, $s2$ denote the adaptive stride values of the first and second max pooling layers. The convolution operation at each level of the feature extraction module can be formulated as:

$$F_i^L = F_i^{L-1} * f \tag{2}$$

where F_i^L and F_i^{L-1} denotes the features maps at L and $L - 1$ layers of i input image and f denotes the kernel. Convolutional operations learn only features of local details at the object level. However, due to the various challenges existing in facial images caused by intra-class variations, a discriminative facial expression representation requires a combination of local, global and context features. For example, the facial expression of the happy class has different appearance between individuals; some of them laugh and others are smile, which causes different muscle movement. Therefore, two extra modules are introduce namely, RSCA and ASPP to enrich the representation of facial expression with more robust and discriminative features to alleviate the intra-class variation problem.

3.2.2 Residual spatial-channel attention module

The proposed RSCA attention module is inspired by the squeeze-and-excitation (SE) attention block (Hu et al. 2018) but with squeezing both spatial and channel dimensions. The proposed RSCA learns the attention weights from the feature maps of the last convolutional block in the feature extraction module FM which has $8C$ channels. It begins by applying an average pooling operation with window size of 7 × 7 and stride 2 to squeeze the spatial dimension of feature maps into half. Then, a convolution layer with a kernel size of 3 × 3 and $2C$ channels followed by ELU activation and batch normalization is utilized to squeeze the input channels into quarter. A transposed convolution layer is then utilized to expand spatial and channel dimensions followed by ELU, batch normalization, and sigmoid activation function.

Given an input feature map $FM \in \Re^{8C \times h' \times w'}$, the mathematical operations of the RSCA mechanism can be written a follows:

$$F_{avg} = AP_2^{7\times7}(FM) \tag{3}$$

$$FM_{down}^{2C} = BN(\delta(Conv_1^{3\times3}(F_{avg}))) \tag{4}$$

$$S = \sigma(BN(\delta(TConv_2^{2\times2}(FM_{down}^{2C})))) \tag{5}$$

where $AP_2^{7\times7}$ denotes average pooling operation in the spatial domain with window size 7 × 7 and stride 2. $Conv_1^{3\times3}$ is 2D convolutional operation with a kernel size 3 × 3, $2C$ channels, and stride 1. BN, δ, and σ denote the batch normalization operation, sigmoid and ELU activation function, respectively. $TConv_2^{2\times2}$ is the transposed convolution operation with up-sampling value equal 2 and $8C$ channels. The attention feature maps FM_{Att} can be calculated as follows:

$$FM_{Att} = S \odot FM + FM \tag{6}$$

where $\odot$ is an element-wise multiplication operation of the attention weights S and the feature maps FM. The dimension of the S attention weights is the same as the dimension of FM denoted as $\Re^{8C \times h' \times w'}$

3.2.3 Atrous spatial pyramid pooling module

The intra-class variation issue significantly affects the classification of facial expressions as it needs both local and context information at different scales. Therefore, Atrous Spatial Pyramid Pooling (ASPP) (Chu et al. 2016) module is employed after the attention module to improve the performance of facial expression recognition. ASPP captures multi-scale context information to alleviate intra-class variation issues. Atrous/Dilated convolutional operation is used to control the field of view of the input image by changing the dilation rate. The number of filter learnable parameters can be fixed while the receptive field of the filter is enlarged by interleaving intermediate pixels by zeros. Atrous convolutions operation can be generalized as follows:

$$y[i] = \sum_k x[i + r.k]w[k] \tag{7}$$

where the Atrous rate r determines the dilation rate which sampling the input image x. Standard convolution is a special case of Atrous using rate r = 1. Atrous/Dilated convolution permits the network to learn larger feature details instead of traditional convolution by increasing the dilation rate parameter. Simply, Atrous convolution is similar to the standard convolution except that the weights (w) of an Atrous convolution kernel are sparse, i.e., the kernel of dilated convolution layers are spaced r locations apart. A dilated 3×3 kernel with rate r = 3 will have the same field of view of 7×7 kernel using only 9 parameters. By controlling the rate parameter, we can arbitrarily control the receptive fields of the convolution layer. This allows the convolution filter to learn features from larger areas of the input without increasing the kernel size. Figure 4 shows examples of dilated convolutional filters of size 3×3 at rates r = 1, 3, and 5.

The proposed ASPP module contains five parallel blocks including three Atrous convolutional filters with kernel size of 3×3 and dilation rates (r = 1, 3, and 5), one convolution with 1×1 kernel size, and max pooling with 2×2 window size. The multi-scale context features are generated by concatenating all these branches and fed to a convolutional layer with 1×1 kernel size and $8C$ filters to reduce the size of concatenated channels.

3.2.4 Proposed classification module

The structure of proposed classification module is light and very simple as it consists of only three convolution layers of kernel 3×3 without relying on fully connected operations (Li et al. 2021). The number of channels in each layer is set to: 4C, 2C, and 7, respectively. This number is gradually decreased until it reaches the number of facial expression classes. The first and second convolutional layers are followed by batch normalization, ELU activation, and dropout layer, while the third convolution is followed by ELU, batch normalization, dropout, and softmax activation. The stride values of each convolution layer is adjusted so that the final feature maps have spatial dimension of 1×1 and number of channels equal to number of facial expression classes (i.e. 7). Using convolutional layers instead of fully connected layers significantly reduces the number of learnable parameters. The stride values of the convolutional layers (sx, sy) are adaptively adjusted according to the input image size and stride values used in the feature extraction module $s1$ and $s2$. The following formulas are used to calculate strides in the x and y directions (i.e., sx and sy) of the convolutional layers in the classification module:

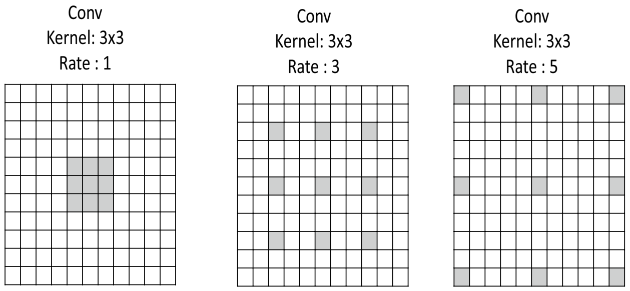

Fig. 4 Atrous convolution with kernel size 3×3 and different dilation rates. Standard convolution corresponds to Atrous convolution with rate = 1. Employing large value of Atrous rate enlarges the model's field-of-view, enabling facial encoding at multiple scales

$$sx = \lceil log(w/(s1 \times s2)/log(3)) \rceil \tag{8}$$

$$sy = \lceil log(h/(s1 \times s2)/log(3)) \rceil \tag{9}$$

where w and h denote the width and height of the input image, $s1$ and $s2$ denote the stride of first and second max pooling in the feature extraction module. For example, when the image size is 48×48, and stride values of first and second pooling layer in FEM s1 = 2 and s2 = 1, the spatial resolution of feature maps fed to classification module (CL) is 24×24. In order to reduce spatial resolution to 1×1, CL apply stride operation in each of the three convolutional layers, the stride values (sx,sy) calculated as: $\lceil log(48/2)/log(3) \rceil$ which equal to 3. Hence, the spatial resolutions of the feature maps generated from CL module will be gradually decreased into 8×8, 3×3, and 1×1.

4 Experimental results

In this section, extensive experiments are conducted to evaluate the effectiveness of proposed method using two in-the-wild datasets including RAF-DB (Li and Deng 2018) and SFEW 2.0 (Dhall et al. 2015) and another lab controlled datasets, namely CK+ (Lucey et al. 2010) dataset.

4.1 Facial expression recognition datasets

Various facial expression datasets are required to evaluate the performance of proposed FER method. Real world FER systems require both large scale and unconstrained datasets. Therefore, in-the-wild datasets such as Real-world Affective Face Database (RAF-DB) (Li and Deng 2018) and SFEW 2.0 (Dhall et al. 2015), produce accurate model for real world FER, unlike lab controlled datasets such as Extended Cohn Kanade dataset (CK+) (Lucey et al. 2010). The accuracy of the proposed method is calculated as the overall mean accuracy of the seven expressions in the FER datasets. Each dataset contains a different distribution of facial expressions inside it as shown in Table 1. Sample images from each dataset are shown Fig. 5.

4.2 Implementation details

The proposed system is implemented using Matlab software package running on Windows machine with Intel core i7 processor, 16 GB RAM and NVIDIA RT 2080 GPU. The

Table 1 Number of samples in the training portions of the used benchmark databases

Databases	Surprise	Fear	Disgust	Happy	Sad	Anger	Neutral/contempt
RAF-DB	1290	281	717	4772	1982	705	2524
SFEW	96	98	66	198	172	178	150
CK+	249	75	177	207	84	135	54

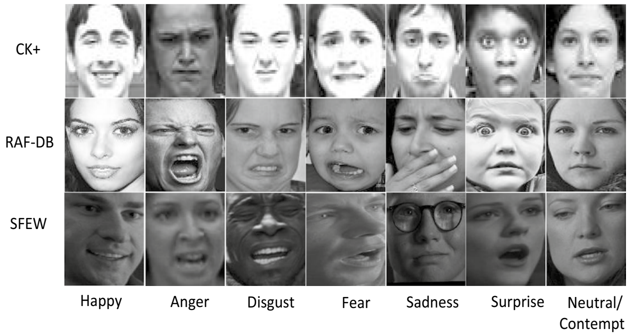

Fig. 5 Example of images from three different databases

number of base filters *C* in all models is set to 64. Proposed models are trained using Adam optimizer with initial learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is decayed by multiplying its value with 0.8 every 10 epochs. Mini-batch size is set to 64 samples and number of epochs is 80, and L2 regularization parameter was set to 0.0001. The drop factor used in the dropout layers after each convolutional layer of the classification modules was 0.3.

The three benchmark datasets have a limited number of samples which may cause over-fitting when using CNN models. Therefore, data augmentation mechanism was used to increase the number of sampling and improve the performance of our framework. In our data augmentation technique, each image is either, reflected horizontally with 50% probability, or scaled within range of [0.8, 1.2], or translated horizontal or vertically within range of $[-2, 2]$ pixels.

4.3 Experiments on RAF-DB

To evaluate the impact of each module on the performance of the proposed method, we investigate the performance of these models with various input image resolutions used in Nan et al. (2021). Furthermore, we conduct several experiments with different combination of the proposed modules shown in Table 2. The combinations of the modules generate four different CNN models with different number of parameters ranged from 1.45 to 2.85 Millions. All proposed models are considered as light-weight in comparisons with other existing deep CNN models.

4.3.1 Ablation study

First, we examine the combination of feature extraction and classification modules (FE-CL), which contains 1.45 Million parameters. Then, we evaluate the effect of adding the proposed attention module on the basic model which called FE-Att-CL model containing 1.67 Million parameters. The contribution of adding the multi-scale atrous spatial pyramid pooling module is also examined using FE-ASPP-CL model with 2.64 Million parameters. Finally, the combination of all modules is evaluated using FE-Att-ASPP-CL model which contains 2.85 Million parameters.

Table 3 introduces the result of total accuracies for each image resolution on the proposed models. Results show that the FE-Att-ASPP-CL model has achieved the highest accuracy with input image resolution 48×48 with value of 84.96%. On the other hand, the FE-Att-CL model has achieved the lowest accuracy value of 70.57% among other proposed models with input image resolution 12×12. The performance of our proposed models at low resolution facial images outperforms other state-of-the-art methods.

From this table, it is observed that using the attention module does not significantly improve the performance of the baseline. However, adding the ASPP module improves the results of accuracy for some image resolutions including 14 and 16. It is also observed that the combinations of all

Table 2 Settings of different proposed models

Model/module	Feature extraction (FE)	Attention (RSCA)	Multi-scale (ASPP)	Classification (CL)	Parameters (M)
FE-CL	✓			✓	1.45
FE-Att-CL	✓	✓		✓	1.67
FE-ASPP-CL	✓		✓	✓	2.64
FE-Att-ASPP-CL	✓	✓	✓	✓	2.85

Table 3 A comparison of our proposed models vs. the previous methods on RAF-DB dataset based on total accuracy (%)

	12 × 12	14 × 14	16 × 16	20 × 20	26 × 26	32 × 32	48 × 48
IFSL-SVM (Yan et al. 2020)	–	–	–	–	–	76.90	–
Multi-class LDA-SVM (Yan et al. 2020)	–	–	–	–	–	50.30	–
E-FCNN (Shao and Cheng 2021)	–	–	–	–	79.04	81.68	83.21
FSER-FER (Nan et al. 2021)	65.97	70.86	73.31	76.66	80.02	81.78	84.03
FE-CL	71.12	73.74	75.64	**78.22**	79.43	80.64	83.71
FE-Att-CL	70.57	73.41	75.70	77.60	79.60	80.15	83.36
FE-ASPP-CL	71.03	**73.94**	**76.39**	77.44	80.51	82.18	83.78
FE-Att-ASPP-CL	**71.42**	72.83	75.38	77.76	**80.71**	**82.67**	**84.96**

modules improve the performance of most image resolutions compared with other models which include attention or ASPP modules separately. It is noticed that increasing the input image resolution is highly impact the performance of all models. The accuracy in increased by 13.54% when image resolution increased from 12 × 12 into 48 × 48. Finally, we have concluded that it is better to process low input images size at the same scale instead of up-scaling the low resolution images to fit the input image size of pre-trained models. In addition, training CNN models with low resolution facial images is better than using super-resolution strategies in Shao and Cheng (2021) and Nan et al. (2021).

4.3.2 Impact of changing stride size in the feature extraction module

In this experiment, the impact of changing the stride size of max pooling layers in feature extraction module is investigated. The lowest and the highest resolutions is chosen to calculate the accuracy of FE-Att-ASPP-CL model at various stride sizes. Four different strides (s1,s2) are examined including, (1, 1), (2, 1), (1, 2), and (2, 2). Employing stride of (2, 1) and (2, 2) in the feature extraction module will reduce the dimension of feature maps into half and quarter the dimension of input image, respectively. From the obtained results, it is beneficial to keep enough spatial resolution information for low resolution facial image classification. Results reported in Fig. 6 show that using stride (1, 1) is better for low resolutions of 12, 14, 16, 20, 26, 32 while stride (2, 1) is better for image with 48 × 48 resolution.

4.3.3 Impact of employing fully-connected layers in the classification module

In this experiment, we study the impact of replacing the convolutional layers used in the proposed classification module (CL) with fully-connected layers. A new model named FE-Att-ASPP-FC is created and compared with FE-Att-ASPP-CL model. We have selected the lowest and highest resolutions to investigate the performance. As shown in Fig. 7, the proposed CL module with convolutional layers increases the performance of 12 × 12 resolution image by 3.8 %. Furthermore, the proposed classification module have increased the performance of 48 × 48 resolution image with 3.4 %. Results reveal that using proposed CL module with convolutional layers is better than fully connected layers for low resolution facial images.

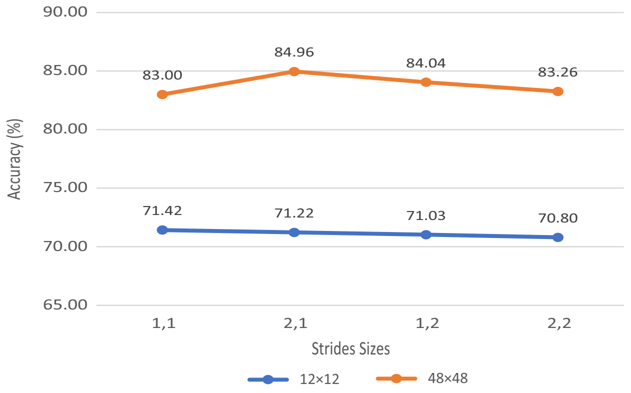

Fig. 6 study the effect of strides sizes in the features extraction module

4.4 Experiments on CK+

In this experiment, we investigate the performance of proposed models using CK+ dataset. Since CK+ dataset has no standard splitting sets, we follow the 10 folds-cross-Validation splitting technique utilized in Shao and Qian (2019), Rao et al. (2021), Chirra et al. (2021), Li and Deng (2018) and Cai et al. (2018). The superior performance of our proposed models is noticed in Table 4. The FE-ASPP-CL model achieves the highest accuracy of 99.9 % with image resolution 32 × 32. On the other hand, the FE-CL achieves the lowest accuracy value of 99.0% with the image resolution 12 × 12. It is noticed that the performance of all proposed models is superior with the resolution size 48 × 48. The high performance of the CK+ dataset are due to several reasons such as it is lab-controlled dataset, the faces inside images in the CK+ dataset are well-aligned with no pose variations, and the appearance of all facial expressions are distinct.

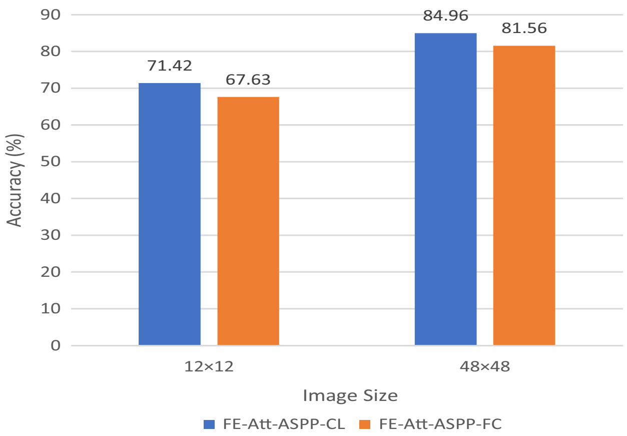

Fig. 7 Study the effect of utilizing our classification module vs. fully-connected layers

We have concluded that the classification of CK+ dataset with our proposed low resolution CNN models is better than using pre-trained model in Shao and Qian (2019). Moreover, utilizing of the single stream network is more better than using dual-branch CNN which achieved low performance in Rao et al. (2021). This can be explained because of the importance of facial context information captured by ASPP module. Furthermore, the utilization of super-resolution strategy is very complex compared to our proposed models and achieved low performance (Rao et al. 2021).

4.5 Experiments on SFEW 2.0

This experiment studies the performance of the proposed models using a small and challenging dataset. SFEW 2.0 dataset not only suffer from class imbalance problem but also there are few number of samples for each class. Since deep learning algorithms require large dataset to be able to adapt its learnable parameters, a combination form RAF-DB, SFEW, and FER2013 training datasets is created to train our proposed FE-Att-ASPP-CL model. We also compare the performance of our model with other state-of-the-art low resolution methods. The result of the experiment is reported in Table 5. The obtained results reveal that proposed models outperform the state-of-the-art methods for images of size 12 × 12, 16 × 16 and 26 × 26. The poor performance of SFEW 2.0 dataset can be explained by that it has several non-face and poor quality images as shown in Fig. 8.

4.6 Visualization of the network activation maps

This experiment visualize the activation maps of sample images from CK+ dataset captured for same person with six different emotions for image size 48 × 48 pixels. Then, the visualization of proposed FE-Att-ASPP-CL model at various levels of the feature extraction module is shown in Fig. 9. We apply the input image samples to activate both early and deep layers in the feature extraction module. The visualized activation maps are calculated from the maximum activation features map of the first and second convolutional blocks. The obtained activation maps illustrate the prominent facial

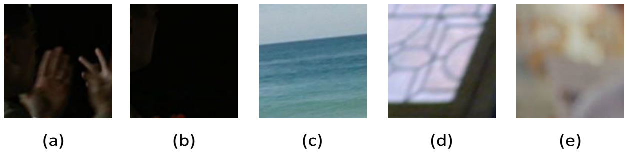

Fig. 8 Visualization of non-face samples inside SFEW 2.0 FER dataset

Table 4 A comparison of our proposed models vs. the previous methods on CK+ dataset based on total accuracy (%)

	12 × 12	14 × 14	16 × 16	20 × 20	26 × 26	32 × 32	48 × 48
IFSL-SVM (Yan et al. 2020)	–	–	–	–	–	98.70	–
Multi-class LDA-SVM (Yan et al. 2020)	–	–	–	–	–	87.10	–
Light-CNN (Shao and Qian 2019)	–	–	–	–	–	–	92.86
Dual-brach CNN (Shao and Qian 2019)	–	–	–	–	–	–	85.71
Pre-trained CNN (Shao and Qian 2019)	–	–	–	–	–	–	95.29
E-FCNN (Rao et al. 2021)	–	–	–	–	–	–	94.95
DCNN (Chirra et al. 2021)	–	–	–	–	–	–	97.77
DCNN-VC (Chirra et al. 2021)	–	–	–	–	–	–	99.04
DLP-CNN (Li and Deng 2018)	–	–	–	–	–	–	95.78
IL-CNN (Cai et al. 2018)	–	–	–	–	–	–	94.35
Pre-trained CNN (Shao and Qian 2019)	–	–	–	–	–	–	95.29
FE-CL	99.00	99.30	99.40	99.30	99.50	99.60	**99.70**
FE-Att-CL	**99.50**	99.20	**99.60**	99.40	99.60	99.60	**99.70**
FE-ASPP-CL	**99.50**	99.40	**99.60**	**99.70**	**99.70**	**99.90**	**99.70**
FE-Att-ASPP-CL	99.40	**99.70**	99.30	**99.70**	**99.70**	99.70	**99.70**

Table 5 A comparison of our proposed models vs. state-of-the-art methods on SFEW 2.0 dataset based on total accuracy (%)

Model	12 × 12	14 × 14	16 × 16	20 × 20	26 × 26	32 × 32	48 × 48
IFSL-SVM (Yan et al. 2020)	–	–	–	–	–	46.50	–
Multi-class LDA-SVM (Yan et al. 2020)	–	–	–	–	–	39.30	–
FSER-FER (Nan et al. 2021)	40.01	**45.28**	45.28	47.35	49.64	**52.39**	**55.14**
FE-CL	40.00	40.75	42.75	44.00	**53.00**	45.25	51.25
FE-ATT-CL	40.00	43.00	**47.75**	45.00	49.50	48.50	51.50
FE-ASPP-CL	40.50	43.00	44.00	45.75	47.75	46.75	48.75
FE-Att-ASPP-CL	43.50	**50.50**	49.75	48.00	51.00	**41.25**	

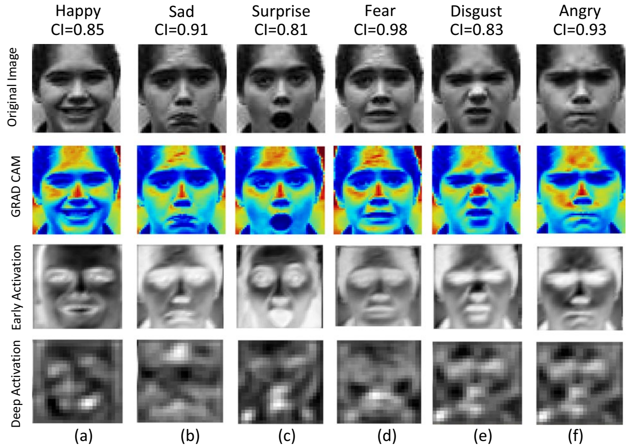

Fig. 9 Visualization of early and deep activation maps of proposed FE-Att-ASPP-CL model for several facial expression samples with its corresponding confidence value

regions which impact the classification results of proposed model.

Finally, the confusion matrices is shown in Tables 6, 7 and 8 of the proposed FE-Att-ASPP-CL model for each testing dataset. Most of the classification errors are due to the confusion of fear, anger, and disgust classes. The three fear, anger, and disgust facial expression classes have few number of samples with similar appearance characteristics. For example, most of the disgust samples are misclassified as anger and vice verse. This mis-classification is due to the similarity of muscle deformation around the mouth in disgust and anger classes. Moreover, most fear samples is misclassified as neutral and many fear samples is misclassified as neutral. One drawback of the proposed method is the large confusion between classes such as: fear-surprise and disgust-neutral-sad due to the low resolution of input face image which cause a loss of substantial information to discriminate between these classes. Another drawback is that

Table 6 Confusion matrix in percentage using FE-Att-ASPP-CL model on RAF-DB dataset with 48 × 48 image size

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	**81.99**	1.86	2.48	5.59	3.11	3.11	1.86
Disgust	9.38	**47.50**	0.63	8.13	13.13	18.13	3.13
Fear	4.05	2.70	**54.05**	5.41	6.76	9.46	17.57
Happy	0.51	0.59	0.25	**93.74**	3.30	1.18	0.42
Neutral	1.04	1.48	0.15	3.56	**83.38**	8.31	2.08
Sad	1.89	2.31	0.21	3.35	7.76	**84.07**	0.42
Surprise	2.43	0.61	1.82	2.43	6.08	2.13	**84.50**

Table 7 Confusion matrix in percentage using FE-Att-ASPP-CL model on CK+ dataset with 48 × 48 image size

	Anger	Contempt	Disgust	Fear	Happy	Sad	Surprise
Anger	**100**	0	0	0	0	0	0
Contempt	0	**95**	0	0	0	0	5
Disgust	0	0	**100**	0	0	0	0
Fear	0	0	0	**100**	0	0	0
Happy	0	0	0	0	**100**	0	0
Sad	0	0	0	0	0	**100**	0
Surprise	0	0	0	0	0	0	**100**

Table 8 Confusion matrix in percentage using FE-RSCA-ASPP-CL model on SFEW 2.0 dataset with 48 × 48 image size

	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	**51.3**	5.3	18.4	0.0	9.2	2.6	13.2
Disgust	0.0	**25.0**	0.0	0.0	50.0	25.0	0.0
Fear	14.3	0.0	**14.3**	0.0	0.0	0.0	71.4
Happy	8.9	3.3	6.7	**65.6**	4.4	5.6	5.6
Neutral	10.1	5.9	8.4	2.5	**46.2**	18.5	8.4
Sad	14.7	10.7	9.3	0.0	17.3	**44.0**	4.0
Surprise	10.3	0.0	24.1	0.0	3.4	6.9	**55.2**

as we increase the resolution of the input images, our model requires large amount of training data.

4.7 Computational time analysis

The recognition time of our proposed method is tested on a PC with Intel core i7 processor, 16 GB RAM and NVIDIA RT 2080 GPU. The computational time of our proposed method is based on CK+ dataset. Figure 10 shows the computational time of proposed models at various image resolutions. We also compare the computational time of our proposed method with the latest existing traditional and deep learning methods as shown in Table 9.

From Table 9, we have noticed that our proposed method has achieved the highest recognition accuracy with the lowest input image resolution.The computational time of our proposed methods has achieved the fourth rank among all traditional and CNN-based methods. We have also noticed that employing traditional machine learning methods can significantly decrease the computation time (Gogić et al. 2020; Happy and Routray 2014) as compared with CNN-based methods. However, these methods have achieved the lowest accuracy. Moreover, we have investigated the computational time of our four proposed models as shown in Fig. 10. We noticed that FE-CL module has achieved the lowest computational in all image resolution. However, FE-Att-ASPP-CL model has achieved the highest computational time for all image resolution. This can be explained by that FE-Att-ASPP-CL module contains the highest number of parameters among the four proposed models.

Table 9 Computational time comparison with other state-of-the-art methods

Method	Image size	Computation time (ms)	Accuracy (%)
VGG19 (Mandal et al. 2019)	224 × 224	1441.95	78.71 (7 classes)
ResNet50 (Mandal et al. 2019)	224 × 224	1343.89	87.31 (7 classes)
LBF-NN (Gogić et al. 2020)	–	1	96.48 (7 classes)
Facial Patches + SVM (Happy and Routray 2014)	96 × 96	295.5	94.14 (6 classes)
CNN (Lopes et al. 2017)	640 × 480	10	95.75 (7 classes)
FN2EN (Ding et al. 2017)	256 × 256	3	98.6 (6 classes)
DRADAP (Mandal et al. 2019)	120 × 20	763.57	84.60 (7 classes)
FE-ASPP-CL	32 × 32	14.81	99.9 (7 classes)

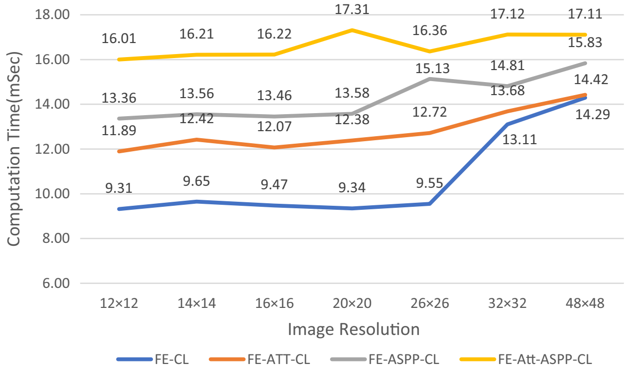

Fig. 10 Computational time of proposed models at various image resolution

5 Conclusions

In this paper, we proposed a new modular full-convolutional deep neural network to classify the facial expressions at various low-resolution images into seven expression categories. A lightweight deep CNN architecture with adaptive strides, spatial and channel attention, and Atrous convolutions is designed to handle low-resolution images. The performance of the proposed modular facial expression system achieves competitive results with other low-resolution state-of-the-art approaches using three public benchmark databases. Our proposed model achieves a relatively a high accuracy of 99% for CK+ dataset in the lowest resolution 12 × 12 and 99.9% in the 32 × 32 resolution. Furthermore, the proposed model has higher accuracy than the state-of-the-art methods in recognizing the RAF-DB dataset. It achieved 71.42% and 84.96% accuracies in the 12 × 12 and 48 × 48 resolutions, respectively. Moreover, the proposed model achieved a competitive performance with state-of-the-art methods for SFEW 2.0 dataset.The recognition accuracies were improved for 12 × 12, 16 × 16, 20 × 20, and 26 × 26 image resolutions. In addition, the proposed method can be generalized to solve other low-resolution image recognition problems instead of relying on super-resolution up-scaling strategies.

Acknowledgements Walaa Aly would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under Project No. R-2023-103.

Availability of data and materials The datasets used in the current study are available in the below links. RAF-DB: http://www.whdeng.cn/raf/model1.html. SFEW 2.0: https://cs.anu.edu.au/few/emotiw2015.html. CK+: https://www.pitt.edu/~emotion/ck-spread.htm. FER-2013: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge.

References

Ben X, Ren Y, Zhang J, Wang S-J, Kpalma K, Meng W, Liu Y-J (2021) Video-based facial micro-expression analysis: a survey of datasets, features and algorithms. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2021.3067464

Cai J, Meng Z, Khan AS, Li Z, O'Reilly J, Tong Y (2018) Island loss for learning discriminative features in facial expression recognition. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, pp 302–309

Chen Y, Phonevilay V, Tao J, Chen X, Xia R, Zhang Q, Yang K, Xiong J, Xie J (2021) The face image super-resolution algorithm based on combined representation learning. Multim Tools Appl 80(20):30839–30861

Cheng B, Wang Z, Zhang Z, Li Z, Liu D, Yang J, Huang S, Huang TS (2017) Robust emotion recognition from low quality and low bit rate video: a deep learning approach. In: 2017 seventh international conference on affective computing and intelligent interaction (ACII). IEEE, pp 65–70

Chirra VRR, Uyyala SR, Kolli VKK (2021) Virtual facial expression recognition using deep cnn with ensemble learning. J Ambient Intell Humaniz Comput 12:10581–10599. https://doi.org/10.1007/s12652-020-02866-3

Chu W-S, De la Torre F, Cohn JF (2016) Selective transfer machine for personalized facial expression analysis. IEEE Trans Pattern Anal Mach Intell 39(3):529–545

Clevert D-A, Unterthiner T, Hochreiter S (2016) Fast and accurate deep network learning by exponential linear units (elus)

Dhall A, Ramana Murthy OV, Goecke R, Joshi J, Gedeon T (2015) Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp 423–426

Ding H, Zhou SK, Chellappa R (2017) Facenet2expnet: regularizing a deep face recognition net for expression recognition. In: 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017). IEEE, pp 118–126

Gogić I, Manhart M, Pandžić IS, Ahlberg J (2020) Fast facial expression recognition using local binary features and shallow neural networks. Vis Comput 36(1):97–112

Happy SL, Routray A (2014) Automatic facial expression recognition using features of salient facial patches. IEEE Trans Affect Comput 6(1):1–12

Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141

Jain DK, Shamsolmoali P, Sehdev P (2019) Extended deep neural network for facial emotion recognition. Pattern Recognit Lett 120:69–74. https://doi.org/10.1016/j.patrec.2019.01.008

Jyoti Kumari R, Rajesh R, Pooja KM (2015) Facial expression recognition: a survey. Procedia Comput Sci 58(1):486–491

Khan RA, Meyer A, Konik H, Bouakaz S (2013) Framework for reliable, real-time facial expression recognition for low resolution images. Pattern Recognit Lett 34(10):1159–1168

Li S, Deng W (2018) Reliable crowd sourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Trans Image Process 28(1):356–370

Li S, Deng W (2020) Deep facial expression recognition: a survey. IEEE Trans Affect Comput. ISSN: 1949-3045. https://doi.org/10.1109/TAFFC.2020.2981446

Li X, He M, Li H, Shen H (2021) A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection. IEEE Geosci Remote Sens Lett 19:1–5

Liu Z, Li L, Wu Y, Zhang C (2020) Facial expression restoration based on improved graph convolutional networks. In: International conference on multimedia modeling. Springer, pp 527–539

Lopes AT, de Aguiar E, De Souza AF, Oliveira-Santos T (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recognit 61:610–628

Lozano-Monasor E, López MT, Vigo-Bustos F, Fernández-Caballero A (2017) Facial expression recognition in ageing adults: from lab to ambient assisted living. J Ambient Intell Humaniz Comput 8(4):567–578

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

Luo J, Liu J, Lin J, Wang Z (2020) A lightweight face detector by integrating the convolutional neural network with the image pyramid. Pattern Recognit Lett 133:180–187

Ma Y, Wang X, Wei L (2021) Multi-level spatial and semantic enhancement network for expression recognition. Appl Intell 1–14

Uma Maheswari V, Varaprasad G, Viswanadha Raju S (2021) Local directional maximum edge patterns for facial expression recognition. J Ambient Intell Humaniz Comput 12(5):4775–4783

Mandal M, Verma M, Mathur S, Vipparthi SK, Murala S, Kumar DK (2019) Regional adaptive affinitive patterns (radap) with logical operators for facial expression recognition. IET Image Process 13(5):850–861

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

Nan F, Jin W, Tian F, Zhang J, Chao K-M, Hong Z, Zheng Q (2021) Feature super-resolution based facial expression recognition for multi-scale low-resolution images. Knowl Based Syst 107678

Połap D (2019) Analysis of skin marks through the use of intelligent things. IEEE Access 7:149355–149363

Rao T, Li J, Wang X, Sun Y, Chen H (2021) Facial expression recognition with multiscale graph convolutional networks. IEEE Multim 28(2):11–19

Saleem Sharmeen M, Abdullah A, Abdulazeez AM (2021) Facial expression recognition based on deep learning convolution neural network: a review. J Soft Comput Data Min 2(1):53–65

Sariyanidi E, Gunes H, Cavallaro A (2014) Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans Pattern Anal Mach Intell 37(6):1113–1133

Shao J, Cheng Q (2021) E-fcnn for tiny facial expression recognition. Appl Intell 51(1):549–559

Shao J, Qian Y (2019) Three convolutional neural network models for facial expression recognition in the wild. Neurocomputing 355:82–92

Sikkandar H, Thiyagarajan R (2021) Deep learning based facial expression recognition using improved cat swarm optimization. J Ambient Intell Humaniz Comput 12(2):3037–3053

Umer S, Rout RK, Pero C, Nappi M (2021) Facial expression recognition with trade-offs between data augmentation and deep learning features. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-020-02845-8

Whitehill J, Serpell Z, Lin Y-C, Foster A, Movellan JR (2014) The faces of engagement: automatic recognition of student engagement from facial expressions. IEEE Trans Affect Comput 5(1):86–98. https://doi.org/10.1109/TAFFC.2014.2316163

Yan Y, Zhang Z, Chen S, Wang H (2020) Low-resolution facial expression recognition: a filter learning perspective. Signal Process 169:107370

Zhang K, Huang Y, Yong D, Wang L (2017) Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans Image Process 26(9):4193–4203

Zhang K, Zhang Z, Li Z, Qiao Yu (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503

Zhao R, Gan Q, Wang S, Ji Q (2016) Facial expression intensity estimation using ordinal information. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, pp 3466–3474

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