## Title: Three convolutional neural network models for facial expression recognition in the wild

## Outline

## Cited by (148)

## Figures (8)

## Tables (6)

## Neurocomputing
Volume 355, 25 August 2019, Pages 82-92

# Three convolutional neural network models for facial expression recognition in the wild

Author links open overlay panelJie Shao, Yongsheng Qian
Show more

https://doi.org/10.1016/j.neucom.2019.05.005Get rights and content

## Abstract

Facial expression recognition (FER) in the wild is a novel and challenging topic in the field of human emotion perception. Different kinds of convolutional neural network (CNN) approaches have been applied to this topic, but few of them ever considered what kind of architecture was better for the FER research. In this paper, we proposed three novel CNN models with different architectures. The first one is a shallow network, named the Light-CNN, which is a fully convolutional neural network consisting of six depthwise separable residual convolution modules to solve the problem of complex topology and over-fitting. The second one is a dual-branch CNN which extracts traditional LBP features and deep learning features in parallel. The third one is a pre-trained CNN which is designed by transfer learning technique to overcome the shortage of training samples. Extensive evaluations on three popular datasets (public CK+, multi-view BU-3DEF and FER2013 datasets) demonstrated that our models were competitive and representative in the field of FER in the wild research. We achieved significant better results with comparisons to plenty of state-of-the-art approaches. Moreover, we provided discussions on the effectiveness and practicability of CNNs with different feature types and architectures for FER in the wild as well.

* Previous article in issue
* Next article in issue

## Keywords

FER in the wild
Convolutional neural network
Shallow network
Dual-branch CNN
Pretrained CNN

## 1\. Introduction

Emotion recognition by facial expression plays an important role in intelligent social interaction. It is widely used in intelligent security [1], robotics manufacturing [2], clinical psychology [3], multimedia [4] and automotive security [5]. In most above-mentioned applications, their inputs are faces captured in the real world. Nevertheless, the main contributions of traditional facial expression recognition methods focused on expressions of the frontal faces. Those expressions were performed by actors staying in a controlled environment. As a result, facial expression recognition in the wild is a novel and challenging topic due to various poses, illumination changes, occlusions, and subtle expressions, etc.

Previous traditional methods for FER in the wild mainly focused on modeling features, e.g., Zhong et al. [6] built a two-stage multi-task sparse learning framework to discriminate facial patches. Zheng et al. [7] treated the feature extraction problem as a convex optimization problem. They both applied traditional state-of-the-art classifier for the final recognition. However, they made good performance on frontal faces, leaving much to be desired on non-frontal faces.

In recent years, machine learning techniques of convolutional neural networks have achieved great success in the field of computer vision. It is wildly used in the fields of visual object recognition [8], Natural Language Processing [9], driverless [10], and so on. It is also a promising approach for the research of FER. Different from traditional techniques, convolutional neural networks can perform tasks in an end-to-end way, associating both feature extraction and classification steps together by training. However, there are still some problems existing in the development of deep learning network:

First, typically the efficiency of a CNN is improved by increasing the number of neurons or the number of layers, so that the network is hoped to learn more complex functions. For example, the early network AlexNet has 7 layers. Then the VGG model with 16 layers appeared, followed by the GoogLeNet consisting of

22 layers. Later came the ResNet model with 152 layers, and the modified ResNet even includes thousands of layers. Although the network?s performance has been improved, their efficiency issues appeared, namely the storage problem of the model and the speed of prediction [11]. Secondly, it may not be robust for the deep CNN to extract features from images with low-resolution, high noise and various rotational changes [12], [13]. The third is the data problem. The deeper the CNN is, the more weights there need to be determined. Consequently, the network needs to be fed with thousands of samples in a larger database, then it could acquire better performance. However, it is impossible to provide large-scale samples in every application area. It means that the shallow CNNs may have better performance in various industrial applications than the deeper ones.

Referring to the first problem we mentioned above, increasing the depth brings a series of negative issues such as overfitting, gradients disappearance, and enormous computational costs. A possible solution to this problem is to create deep sparsely compressed network. Gao Et al. [14] proposed a feed-forward approach to build connections directly between each pair of convolutional layers to form a dense convolutional network (DenseNet). It made use of the short connections between the input layer and the layer close to the output to make the convolutional network deeper, so that the training process would be more precise and more effective. Unfortunately, most current GPUs and CPUs are not able to efficiently run sparse network model [15]. Therefore, in the paper we propose a shallow CNN with good performance on facial expression recognition in the wild. So that it would be suitable for practical problems currently.

At present, most CNN models for facial expression recognition use the features generated by the convolution layers using the raw pixel data as the main features. Local Binary Pattern (Local Binary Pattern, LBP) is a texture description operator which is usually used for facial expression recognition. It can effectively adapt to changes in illumination and local rotation [16]. Features extracted by convolutional neural network may not be robust to the image rotation changes. We wanted to explore if there was any way to apply LBP features along with raw pixels to a network and observe the performance of the model when it had a combination of two different features.

The data problem, the third one we mentioned above is the most troublesome problem we met. Facial expressions in the wild have hundreds of thousands of variations referring to different poses, human races, genders, conventions and environments. On the contrary, datasets of facial expression in the wild are quite limited. Some datasets only have hundreds of samples, so it is difficult for deeper CNN models to learn as good results as they have in some other fields. A study on transfer learning of facial expression recognition seems to offer a better chance of producing more accurate predictions [17], [18], [19].

Based on the above discussion, in this paper, we proposed three kinds of convolutional neural networks for facial expression recognition in the wild. The first one is a shallow CNN named Light-CNN. The second one is a dual-branch CNN, which is an attempt to integrate traditional features with the original data in a uniform network. The third method is a pre-trained CNN, which is a deep network. We elaborated their architectures and conducted comprehensive experiments on three public facial expression datasets: CK+, BU-3DFE and FER2013. Plenty of comparisons were made among our three CNNs and other state-of-the-art methods. We demonstrated that our proposed methods are significantly better than the previous methods. Meanwhile, we also provided a discussion on the merits and shortcomings of our three network architectures.

Our contributions are as follows:

* ?

We elaborately designed three representative CNN models for facial expression recognition in the wild, in order to discuss their advantages and disadvantages, and to provide possible solutions for problems of over-fitting, high computational complexity, and lack of training samples et al. in FER in the wild by deep learning.

* ?

A large number of experiments are implemented on different facial expression datasets, including CK+, BU-3DFE and FER2013. CK+ is a traditional facial expression dataset. BU-3DFE has samples with different poses but captured in a lab-controlled environment. FER2013 includes face samples captured in the real world.

* ?

We made comparisons among the three CNN architectures, as well as the comparison between our three methods and the state-of-the-art methods. We provide conclusions about different network structures and demonstrated that our proposed methods are competitive with state-of-the-art methods.

The structure of this paper is as follows: Section 2 introduces existing state-of-the-art emotion recognition approaches based on CNN, and some traditional feature extraction methods. We introduce our CNNs in details in Section 3. Section 4 describes the datasets along with details of our experiments, and then we present our results and discussions. Section 5 give a conclusion followed by a list of references.

## 2\. Related work

Traditional methods on FER can be categorized into three major steps: facial detection, feature extraction and classification, where face detection [20], [21], [22] has become a well-developed technology and been applied to the real-world applications. Extracting powerful features and designing effective classifiers are two key components of FER. For feature-based methods, hand-crafted features are often used to represent expression images. For example, Gabor wavelets [23] show good robustness through capturing image edges at different scales and orientations. Local binary pattern (LBP) is demonstrated to be useful in FER. Ying et al. [24] proposed a facial expression recognition method based on LBP and Adaboost in 2008. LBP was later extended for modeling spatio-temporal features, naming LBP-TOP [25]. Later, Qi et al. [26] proposed a new expression recognition method based on cognitive and mapping binary patterns. They applied pseudo-3D model to segment face areas into six facial sub-regions. Although the LBP operator is robust to monotonic gray-level changes and computational efficiency, there are some limitations. For example, it is sensitive to noise, and in its template, only gradients between the central pixel and its neighborhood are considered. Thus, it inevitably loses some information [27]. Other features, including Histograms of Oriented Gradients (HOG) [28], Scale-Invariant Feature Transform (SIFT), and Singular Value Decomposition (SVD) [29] have also been widely used. In [30], [31], proved that the singular value of the image can be used as the global feature with invariant scale of rotation shift. These special attributes of the singular value are used to design the compact global feature of facial image representation to improve the accuracy of low-resolution face recognition.

Facial expression images in the wild are more challenging in face detection, facial landmark location, and pose standardization than traditional facial expression images. Consequently, traditional methods are not suitable for the research on FER in the wild.

In recent years, the appearance of deep learning has significantly improved the performance of FER related tasks [32], [33], [34], [35], [36]. Then there were two trends. On the one hand, the FER problem increasingly utilized deeper and deeper neural networks to improve the ability of tackling big-data problems. Mollahossein et al. [32] proposed an in-depth neural network architecture for FER, which was inspired by GoogLeNet and AlexNet. It outperformed traditional methods based on hand-crafted features. Training deep networks with limited data may even result in poor performance due to over-fitting. To solve the problem, Zhang et al. [34] proposed a deep neural network (DNN) with the SIFT feature, which achieved the accuracy of 78.9% on the multi-view BU-3DFE dataset. To reduce the influence of various head poses, Jung et al. [35] proposed a jointly CNNs with facial landmarks and color images, which achieved the accuracy of 72.5%, but the network consisted of only three convolutional layers and two hidden layers, making it be difficult to accurately learn facial features. Lopes et al. [36] proposed a combination of Convolutional Neural Network and special image pre-processing steps (C-CNN) to recognize six expressions under head pose at 0?, whose accuracy was 90.96% on the BU-3DFE dataset. Its robustness was unknown under different head poses. On the other hand, some works preferred to aggregate different features in deep networks. They demonstrated that comprehensive feature representations had better performance than single feature. For example, Majumder et al. [37] fused LBP features and facial geometric features with a deep network-based technique for FER in the wild, and achieved good performance. Hamester et al. [38] proposed a new architecture by constructing a multi-channel convolutional neural network (MCCNN). It utilized CNN and an automatic encoder to extract features. On the contrary, Alizadeh et al. [39] claimed that hybrid feature sets did not help in improving the model accuracy. Therefore, we attempt to provide a dual-branch model solution in this paper which includes both traditional texture features and raw data.

Lack of training samples is a big problem for FER in the wild using deep CNNs. To solve this problem, some methods used pre-trained network for classification or re-trained a network model to re-initialize the weights for new datasets [40]. The techniques are regarded as ?transfer learning?. Ruiz-Garcia et al. [41] used greedy layer-wise fashion to pre-train deep CNNs as a stacked convolution auto-encoder (SCAE) for emotion recognition. Employing SCAE as a pre-training model improves not only performance but training time. Yanai et al. [42] sought a good combination of DCNN-related techniques. The fine-tuning and activation features were extracted from the pre-trained DCNN. In addition to its high classification accuracy, DCNN was very suitable for large-scale image data.

## 3. Proposed method

In this section, the proposed three CNNs: a Light-CNN, a dual-branch CNN and a Pretrained CNN are described in details.

### 3.1. The Light-CNN

The Light-CNN is a shallow CNN, its architecture is shown in Fig. 1. It is a fully convolutional neural network. It consists of 6 depthwise separable residual convolution modules whose architectures are shown in Fig. 2. The architecture of the module was inspired by the Xception and ResNet. We associated the depthwise separable module with the residual network module to build a depthwise separable residual convolution module. The depthwise separable residual convolution module has three separable convolution layers (SeparableConv2D) and one convolution layer. In the first SeparableConv2D layer, we had 16 1 × 1 filters along with batch normalization, but without max pooling. In the second SeparableConv2D layer, we had 16 3 × 3 filters along with batch normalization, but without max pooling as well. In the third SeparableConv2D layer, we had 16 1 × 1 filters along with batch normalization, as well as max pooling with a filter of size 2 × 2. The number of filters gradually increases from 16 to 512 in 6 modules, as shown in Fig. 2. Each depthwise separable residual convolution module is followed by a Rectified Linear unit (ReLU). The images are resized to be 64 × 64 × 1 pixels before being sent to the network. In the first and the second convolutional layer, we have 8 3 × 3 filters respectively, with the stride of size 1, along with batch normalization and ReLU. They extract low-level edge features of the image and retain the details. The low-level edge features are shown in Fig. 3-a. The deep features of the extracted image from first depthwise separable residual convolution modules are shown in Fig. 3-b. It can be found that the deeper it is, the more abstract the output features are.

1. Download: Download high-res image (800KB)
2. Download: Download full-size image

Fig. 1. The basic structure of the Light-CNN.

1. Download: Download high-res image (2MB)
2. Download: Download full-size image

Fig. 2. The structure of 6 depthwise separable residual convolution modules.

1. Download: Download high-res image (376KB)
2. Download: Download full-size image

Fig. 3. Feature visualization of the image.

After 6 depth wise separable residual convolution modules, we designed a convolutional layer following with a global average pooling layer to reduce the number of features, and to regularize the entire network to prevent overfitting. The output of the layer is a vector whose dimension is the number of expressions. A softmax layer is at the bottom, which is a generalization of the logistic regression model for multi-classification problems. In the multi-classification problem, k possibilities are predicted (k is the number of sample tags). Assume that the input feature is x(i)??n+1, and the sample tag is _y_(_i_), so the training set S={(x(1),y(1)),(x(2),y(2)),?,(x(m),y(m))} of the supervised learning constitutes the classification layer. Then the function and cost function forms are as follows:(1)h?(x(i))=[p(y(i)=1|x(i);?)p(y(i)=2|x(i);?)?p(y(i)=k|x(i);?)]=1?j=1ke?jTx(i)[e?1Tx(i)e?2Tx(i)?e?kTx(i)]
Where ?1,?2,?,?k??n+1 is the model parameter and 1?j=1ke?jTx(i) is the normalization term for the probability distribution, making the sum of all probabilities equal to
1.(2)J(?)=?1m[?i=1m?j=1k1{y(i)=j}ln?jTx(i)?l=1ke?lTx(i)]
Among them, 1{} = 1 is an indicative function whose value rule is: when the expression in the curly braces is true, the result of the function is 1, otherwise the result is 0.

### 3.2. The dual-branch CNN

The dual-branch CNN is designed to simultaneously estimate the global features and local texture features. Fig. 4 illustrates its flowchart. The architecture consists of three modules: two individual CNN branch modules and a fusion module. The first branch takes the entire image as input and extract global features. The other branch takes the texture feature image preprocessed by LBP as input. Finally, the third module is a fusion network that takes as input the global and texture features. The global feature is intended to represent the integrity of the expression, while the texture feature focuses on the details of the description of the local area, which can directly indicate some active expression areas on the face. These two separate branches represent expressions from two different aspects. They are complementary and both are of interest.

1. Download: Download high-res image (533KB)
2. Download: Download full-size image

Fig. 4. Framework of the dual-branch CNN.

The Light-CNN, we introduced in the previous sub-section, is truncated to apply for the first branch. The dimensional reduction CNN for the second branch consists of two Convolutional layers. It reduces the dimension of LBP features and facilitates the following combination of the two features. The architecture details of the two branches are shown in Table I. We omitted the architecture details between layer1 and layer28 of the first branch in Table 1.

Table 1. Architectures of the two branches.

The first branch	Empty Cell	Layer1	?	Layer28	Layer29
Empty Cell	type	Input	?	GlobalAveragePooling2D	Flatten
Empty Cell	size	224 × 224 × 1	?	4096	4096

The second branch		Layer1	Layer2	Layer3	Layer4	Layer5	Layer6
	type	Input	Conv2D	Max Pooling	Conv2D	Max Pooling	Flatten
	size	64 × 64 × 1	4 × 4 × 32	2 × 2	4 × 4 × 16	2 × 2	2704

### 3.3. The pretrained CNN

To observe the effect of deeper CNN, the ResNet101[43] network was exploited to construct our pretrained network model. As we didn?t have big databases to train the network, we directly used the model which was previously trained by ImageNet [44] dataset. ImageNet has thousands of different face images, so we could retain the most original network parameters for initialization. Then we trained the model and performed fine-tuning on some of the layers to extract more specific features. Fig. 5 shows the architecture of the pretrained CNN. The original network consists of five convolution modules. Then average pooling is followed by a flatten layer. The output of the full connection layer is 1000. We modified the full connection layer from 1000 to 6 or 7, according to the number of expression categories.

1. Download: Download high-res image (497KB)
2. Download: Download full-size image

Fig. 5. The framework of the pretrained CNN.

## 4\. Experimental results

We evaluated the proposed methods on three publicly available facial expression datasets. Some image samples are shown in Fig. 6. Images from the CK+ Database are in the top row. Images from the BU-3DFE Database are in the middle row. Images from the FER2013 Database are in the bottom row. The experimental details will be described in this section.

1. Download: Download high-res image (851KB)
2. Download: Download full-size image

Fig. 6. Examples Images in CK+(top), BU-3DFE (middle), FER2013 (bottom) Datasets.

### 4.1. Databases and Protocols

CK+ Database: The Extended Cohn-Kanade (CK+) database [45] includes 593 facial expression video sequences recorded from 123 subjects ranging from 18 to 30 years old in lab-controlled environment. Most are frontal faces. We only retained the final frames with peak expression of video sequences in our experiments. Totally we got 327 static expression images with seven emotion labels (anger, contempt, disgust, fear, happy, sadness, surprise). We divided the CK+ dataset into a training set with 90% samples and a validation set with the other 10% samples.

BU-3DFE Database: The BU-3DFE multi-view facial expression database [46] contains 100 subjects of different ethnicities, including 56 females and 44 males. Six facial expressions (anger, disgust, fear, happiness, sadness, and surprise) are elicited by various manners and head poses. Each of them includes 4 levels of intensities. The images are also captured in the lab-controlled environment. These models are comprised by both 3D geometrical shapes and color textures with 83 feature points. We use 3D facial models to restore 2D facial images of multiple viewing angles (0?, 30?, 45?, 60? and 90?). We divided the dataset into a training set with 90% samples and a validation set with the other 10% samples as well.

FER2013 database: The FER2013 dataset [47] is a static real world facial expression database, which consists of 35,887 48 × 48 gray face images. The image is processed in such a way that the face is centered and the occupancy of each face in the image is approximately the same. Each image is divided into one of seven categories that express different facial emotions. These facial emotions have been categorized as: anger, disgust, fear, happy, sad, surprise and neutral. Besides the image category, images are divided into three different sets, a training set, a validation set, and a test set. There are approximately 29,000 training images, 4,000 verification images and 4000 images for testing. For the purpose of data enhancement, we make a mirror image by horizontally flipping the image in the training set.

### 4.2. Experimental parameters

The experimental platform consists of AMD Ryzen 5 1600(6 × 3.2 GHz processor), 16GB memory, GTX1080 and Ubuntu 16.04 operation system. The deep learning framework Keras is exploited. The parameter settings of the Light-CNN, the dual-branch CNN and the pretrained CNN are presented in Table 2.

Table 2. The parameter setting of Light-CNN, Dual-Branch Network and Pre-trained network.

Models	Parameters	Values
The Light-CNN	Optimizer	Adam
	Image size	224 × 224
The dual-branch CNN	Optimizer	SGD
	Learning rate	1e?3
	Momentum	0.9
	Learning decaying factor	1e?6
	Image size	224 × 224
The pretrained CNN	Optimizer	SGD
	Learning rate	1e-3
	Momentum	0.9
	Learning decaying factor	1e?6
	Image size	224 × 224

In preprocessing, we applied Multi-Task Convolutional Neural Network (MT-CNN)[48] for face detection. Then, the cropped face and five facial landmarks were detected. The five landmarks indicate the centers of two eyes, the end of the nose and two corners of the mouth. All face images were resized to 224 × 224 pixels and aligned based on three landmarks (two center points of eyes and the center point of mouth). In the dual-branch network, LBP feature images were resized to 64 × 64 pixels.

For image enhancement, we used a series of random transformations to ?enhance? the image so that the model would not be fed with two identical images [49]. It would effectively improve image utilization. The transformations included rotation, flipping, scaling, and panning. In this paper, the width and height displacement were used. The shifting range of width and height were set under 20%. The random rotation range was 0?20?. Both the shear range and the zoom range were [0?0.1]. We flipped the images horizontally and applied the fill pattern strategy to fill the newly created pixels as well.

### 4.3. Experiments on three popular datasets

We tested our methods on three widely used FER datasets: CK+, BU-3DFE and FER2013. The CK+ dataset includes expressions of seven labels: anger, contempt,disgust, fear, happy, sadness, surprise. The BU-3DFE dataset has six labels: anger, disgust, fear, happiness, sadness, and surprise. The FER2013 dataset has seven labels: anger, disgust, fear, happy, sad, surprise and neutral.

To evaluate the overall performance, the confusion matrices of our methods on three datasets are illustrated in Fig. 7. Fig. 7(a)?(c) are experimental results on CK+, implemented by the Light-CNN, the dual-branch CNN and the pretrained CNN respectively. Fig. 7(d)?(f) are experimental results on BU-3DFE dataset with three models. Fig. 7(g)?(i) are experimental results on FER2013. As demonstrated in these figures, the pretrained CNN resulted in higher accuracy for most of the labels. All the three models performed well on CK+ datasets especially the Light-CNN and the pretrained CNN, as CK+ is a dataset with facial expression samples captured in a lab-controlled environment. It is interesting to see that the happy label has the highest accuracy in CK+ and FER2013 datasets, which implies that the features of a happy face are more distinguishable than other expressions. Besides, the sadness and the surprise expressions are relatively easier to be recognized from an acted face than from a face in the real world. Because the sad and surprise labels have high accuracy on CK+ and BU-3DFE datasets, but fail to be good on FER2013. Their matrices also reveal which labels are likely to be confused by the trained networks. For example, we can see the correlation of angry label with the fear and surprise labels. There are lots of instances that their true label is angry but the classifier has misclassified them as fear or surprise. These mistakes are consistent with what we see when looking at images in the dataset; even as a human, it can be difficult to recognize whether an angry expression is actually surprise or angry. This is due to the fact that people do not all express emotions in the same way.

1. Download: Download high-res image (2MB)
2. Download: Download full-size image

Fig. 7. Confusion matrices for three networks on three expression databases. (a)?(c) are confusion matrices for the Light-CNNs, dual-branch CNN and the pretrained CNN on CK+, (d)?(f) are confusion matrices for the three networks on BU-3DFE, (g)?(i) are confusion matrices for the three networks on FER2013.

Moreover, we plotted the obtained accuracy of FER2013 using the Light-CNN, the dual-branch CNN and the pretrained CNN during epochs in Fig. 8. As seen in Fig. 8, the pretrained CNN has the best validation accuracy. The performance of the Light-CNN is close to the best one. Furthermore, one can observe that the Light-CNN has less overfitting behavior than the others. We also provided the number of parameters in networks and their running time on FER2013 for comparison in Table 3. The Light-CNN has the least parameters, and it runs much faster than the others. By integrating the results shown in Fig. 8 and Table 3, we concluded that LBP features were not helpful in deep network. With the development of the architecture, CNNs adopting raw pixel data is strong enough to extract sufficient information for facial expression in the wild. Besides, the Light-CNN got good scores on all three datasets and its performances are quite close to those of the pretrained CNN. Besides, it runs much faster than the pretrained CNN, which is beneficial for practical applications.

1. Download: Download high-res image (425KB)
2. Download: Download full-size image

Fig. 8. Comparison of parameters and running time on FER2013 dataset.

Table 3. Comparison of parameters in three networks and their running time on

FER2013 dataset.

Models	Parameters	Running time (h)
The Light-CNN	1,108,151	12.7
The dual-branch CNN	64,629,847	23.3
The pretrained CNN	7,128,327	18.8

### 4.4. Comparisons with the state-of-the-art methods

To evaluate the performance of the proposed algorithm with other algorithms, Table 4 and 5 list the accuracy of our proposed and the state-of-the-art algorithms on the CK+ and BU-3DFE databases. LBP, HOG and Gabor filters are traditional feature descriptors in facial expression recognition and have been widely used. However, the recognition accuracy of most traditional methods are lower than that of deep learning.

Table 4. The accuracy (%) of different methods on CK+ dataset.

Methods	# of Expression	Accuracy(%)
LBP [16]	6+neutral	87.20
HOG [28]	6+neutral	89.70
Gabor filter [50]	7	84.80
Poursaberi et al. [51]	6	92.02
Deepak Ghimire [52]	6	94.10
AU-DNN [33]	6+neutral	92.05
JFDNN [35]	6	97.3
CNN [32]	6	93.2(Top-1)
C-CNN [36]	6	91.64
Our Light-CNN	**7**	**92.86**
Our dual-branch CNN	**7**	**85.71**
Our pre-trained CNN	**7**	**95.29**

For the CK+ database, the accuracy of our algorithm is superior to most of the other advanced algorithms. The best performance of the existing deep learning methods is 97.3%, which is achieved by Jung [35]. His network consists of three convolutional layers and two hidden layers. The filter size in the three convolutional layers is 5 × 5, and the numbers of hidden nodes is set to 100 and 600 respectively. But his results declined to 92.35% without joint fine-tuning. By contrast our proposed method does not use any geometric features or temporal video information, and improved the accuracy to 95.29% under seven expressions.

The comparison results on BU-3DFE dataset are shown in Table 5, the accuracy of method[53] based on HOG is 54.64%. The accuracy of multi-class SVM with LBP and LGBP in [54] is 71.1%. Dapogny et. al. [55] proposed PCRF to capture low-level expression transition patterns on the condition of head pose estimation for multi-view dynamic facial expression recognition. Their average accuracy reached 76.1%. The JFDNN [35] reaches only 72.5%, which used to get the best result in CK+ dataset. The higher accuracies are achieved with SIFT feature using GSRRR and DNN-Driven methods proposed in [56] and [34], which are 78.9% and 80.1%, respectively. In addition, Lopes et al. [36] used intensity features to recognize six expressions with frontal poses and achieved an average accuracy of 90.96%. Our best result reaches 86.5%, which is competitive with the above method.

Table 5. Accuracy (%) using different methods on BU-3DFE dataset.

Methods	Poses	Accuracy (%)
HOG [53]	5	54.64
LBP and LGBP [54]	7	71.1

JFDNN [35]	5	72.5
PCRF [55]	5	76.1
CGPR [57]	5	76.5
GSRRR [56]	5	78.90
DNN-Driven [34]	5	80.10
C-CNN [36]	1 (frontal)	90.96
Our Light-CNN	**5**	**86.20**
Our dual-branch CNN	**5**	**48.17**
Our pre-trained CNN	**5**	**86.50**

Besides, Table 6 shows the results achieved by the competing methods on the FER2013 database, which is the most challenging database in our experiment. There was a leaderboard of facial expression recognition challenge on FER2013 dataset. The number one method is the RBM. Our Light-CNN model achieved the accuracy of 68%, which is ranked #5 in the list, and the pretrained model ranked #2 among all the participating teams. It has almost the same accuracy with the first team.

Table 6. Accuracy (%) using different methods on FER2013 dataset.

Methods	Accuracy (%)
RBM [58]	71.16
Kim et al. [59]	70.58
Jeon et al. [60]	70.47
Devries et al. [61]	67.21
CNN [32]	66.4 (Top-1)
Liu et al. [62]	65.03
Shen et al. [63]	61.86
Ergen et al. [64]	57.10
Our Light-CNN	**68**
Our dual-branch CNN	**54.64**
Our pre-trained CNN	**71.14**

### 4.5. Discussion

Above all, Our CNN models achieved state-of-the-art performance without using additional training data or functions, comprehensive data enhancement or facial registration. It is predictable that it will success in processing larger database in the future. Under the same conditions, the performance of deeper pre-trained CNN was better than the others. Our experimental results demonstrated the potential to significantly improve FER performance using pre-trained deep network structures, which could solve the problems of the lack of training samples and over-fitting. The Light-CNN overcome the challenge of overfitting, and kept good performance in all the popular FER datasets (see Table 3) as well. In addition, in our dual-branch CNN, learning features and manual features were put into the final fusion layers to explore whether the combination of features can improve the classification effect. The results showed that the effect of learning deep features was not improved under the guidance of traditional features.

## 5\. Conclusions and future works

We developed three CNN models for facial expression recognition in the wild and evaluated their performances using different analyzing and visualization techniques. The results demonstrated that the deeper model has better performance on facial feature learning and emotion classification. However, the experiments implemented by the Light-CNN proved that a shallow CNN could also achieve good scores in facial expression recognition in the wild. In addition, mixing feature sets do not help to improve accuracy, which means that convolutional neutral networks can learn key facial features simply by using raw pixel data. In future work, we will use more efficient hand-crafted features to join our dual-branch CNN and change the fusion mode. Moreover, we will use cross-database training network parameters to get better generalization capabilities.

## Conflict of interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China Youth Science Foundation project (Nos. 61802250, 61401268) and Shanghai Natural Science Foundation of China (15ZR1418400).

Recommended articles

## References

1. [1]
Wang R., Fang B.
Affective computing and biometrics based HCI surveillance system
Proceedings of the International Symposium on Information Science and Engineering (2008), pp. 192-195
View in ScopusGoogle Scholar
2. [2]
W. Weiguo, Qingmei M., Yu W.
Development of the humanoid head portrait robot system with flexible face and expression
Proceedings of the 2004 IEEE International Conference on Robotics and Biomimetics (2004), pp. 757-762, 10.1109/ROBIO.2004.1521877
Google Scholar
3. [3]
Su M.H., Wu C.H., Huang K.Y., Hong Q.B., Wang H.M.
Exploring microscopic fluctuation of facial expression for mood disorder classification
Proceedings of the International Conference on Orange Technologies (2017), pp. 65-69
CrossrefView in ScopusGoogle Scholar
4. [4]
M.B. Mariappan, M. Suk, B. Prabhakaran, Facefetch: a user emotion driven multimedia content recommendation system based on facial expression recognition, Proceedings of the 2012 IEEE International Symposium on Multimedia(2012) 84?87.
Google Scholar
5. [5]
S.A. Patil, P.J. Deore
Local binary pattern based face recognition system for automotive security
Proceedings of the International Conference on Signal Processing, Computing and Control (2016), pp. 13-17
Google Scholar
6. [6]
Zhong L., Liu Q., Yang P., Liu B., Huang J., D.N. Metaxas
Learning multiscale active facial patches for expression analysis
Proceedings of the Computer Vision and Pattern Recognition (2012), pp. 2562-2569
View in ScopusGoogle Scholar
7. [7]
Zheng H., Geng X., Tao D., Jin Z.
A multi-task model for simultaneous face identification and facial expression

recognition
Neurocomputing, 171 (C) (2016), pp. 515-523
View PDFView articleView in ScopusGoogle Scholar
8. [8]
Qi T., Yong X., Quan Y., Wang Y., Ling H.
Image-based action recognition using hint-enhanced deep neural networks
Neurocomputing, 267 (2017), pp. 475-488
View PDFView articleView in ScopusGoogle Scholar
9. [9]
M. Morchid
Parsimonious memory unit for recurrent neural networks with application to natural language processing
Neurocomputing, 314 (2018), pp. 48-64
View PDFView articleView in ScopusGoogle Scholar
10. [10]
Yu L., Shao X., Yan X.
Autonomous overtaking decision making of driverless bus based on deep q-learning method
Proceedings of the 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO) (2017), pp. 2267-2272
View in ScopusGoogle Scholar
11. [11]
Yang K., Gong X., Liu Y., Li Z., Xing T., Chen X., Fang D.
CDEEPARCH: a compact deep neural network architecture for mobile sensing
Proceedings of the 2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON) (2018), pp. 1-9
View in ScopusGoogle Scholar
12. [12]
M.D. Zeiler, R. Fergus
Visualizing and understanding convolutional networks
Proceedings of the ECCV (2014)
Google Scholar
13. [13]
A. Azulay, Y. Weiss, Why do deep convolutional networks generalize so poorly to small image transformations? 2018, arXiv preprint arXiv:1805.12177.
Google Scholar
14. [14]
Huang G., Liu Z., L. van der Maaten, K.Q. Weinberger
Densely connected convolutional networks
Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017), pp. 2261-2269
View in ScopusGoogle Scholar
15. [15]
G. Pleiss, Chen D., Huang G., Li T., L. van der Maaten, K.Q. Weinberger
Memory-efficient implementation of densenets
(2017)
, arXiv preprint arXiv:1707.06990
Google Scholar
16. [16]
Shan C., Gong S., P.W. Mcowan
Robust facial expression recognition using local binary patterns
Proceedings of the IEEE International Conference on Image Processing (2005)
Google Scholar
17. [17]

Xu M., Cheng W., Zhao Q., Ma L., Xu F.
Facial expression recognition based on transfer learning from deep convolutional networks
Proceedings of the International Conference on Natural Computation (2016), pp. 702-708
Google Scholar
18. [18]
J. Luttrell, Zhou Z., Zhang Y., Zhang C., Gong P., Yang B., Li R.
A deep transfer learning approach to fine-tuning facial recognition models
Proceedings of the 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA) (2018), pp. 2671-2676, 10.1109/ICIEA.2018.8398162
View in ScopusGoogle Scholar
19. [19]
Peng M., Wu Z., Zhang Z., Chen T.
From macro to micro expression recognition: deep learning on small datasets using transfer learning
Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (2018), pp. 657-661
View in ScopusGoogle Scholar
20. [20]
Jian M., Lam K.M., Dong J.
Facial-feature detection and localization based on a hierarchical scheme
Inf. Sci., 262 (3) (2014), pp. 1-14
View PDFView articleView in ScopusGoogle Scholar
21. [21]
G. Sikander, S. Anwar, Y.A. Djawad
Facial feature detection: a facial symmetry approach
Proceedings of the International Symposium on Computational and Business Intelligence (2017), pp. 26-31
View in ScopusGoogle Scholar
22. [22]
Zhang K., Zhang Z., Li Z., Qiao Y.
Joint face detection and alignment using multitask cascaded convolutional networks
IEEE Signal Process. Lett., 23 (10) (2016), pp. 1499-1503
View in ScopusGoogle Scholar
23. [23]
S.C. Bakchy, M.J. Ferdous, A.H. Sathi, K.C. Ray, F. Imran, M.M. Ali
Facial expression recognition based on support vector machine using Gabor wavelet filter
Proceedings of the 2017 2nd International Conference on Electrical Electronic Engineering (ICEEE) (2017), pp. 1-4, 10.1109/CEEE.2017.8412888
View in ScopusGoogle Scholar
24. [24]
Ying Z., Fang X.
Combining LBP and adaboost for facial expression recognition
Proceedings of the 9th international conference on signal processing (2008), pp. 1461-1464
Google Scholar
25. [25]
Wang Y., Yu H., B. Stevens, Liu H.
Dynamic facial expression recognition using local patch and LBP-top
Proceedings of the International Conference on Human System Interactions (2015), pp. 362-367

CrossrefView in ScopusGoogle Scholar

26. [26]
Qi C., Li M., Wang Q., Zhang H., Xing J., Gao Z., Zhanga H.
Facial expressions recognition based on cognition and mapped binary patterns
IEEE Access, PP (99) (2018)
Google Scholar
1?1.

27. [27]
Huang D., M. Ardabilian, Wang Y., Chen L.
Asymmetric 3D/2D face recognition based on LBP facial representation and canonical correlation analysis
Proceedings of the IEEE International Conference on Image Processing (2010), pp. 3289-3292
Google Scholar

28. [28]
P. Kumar, S.L. Happy, A. Routray
A real-time robust facial expression recognition system using hog features
Proceedings of the International Conference on Computing, Analytics and Security Trends (2017), pp. 289-293
CrossrefGoogle Scholar

29. [29]
Jian M., Lam K.M.
Face-image retrieval based on singular values and potential-field representation
Signal Process., 100 (7) (2014), pp. 9-15
View PDFView articleView in ScopusGoogle Scholar

30. [30]
Jian M., Lam K.M., Dong J.
A novel face-hallucination scheme based on singular value decomposition
Pattern Recognit., 46 (11) (2013), pp. 3091-3102
View PDFView articleView in ScopusGoogle Scholar

31. [31]
Jian M., Lam K.M.
Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition
IEEE Trans. Circuits Syst. Video Technol., 25 (11) (2015), pp. 1761-1772
View in ScopusGoogle Scholar

32. [32]
A. Mollahosseini, Chan D., M.H. Mahoor
Going deeper in facial expression recognition using deep neural networks
Proceedings of the Applications of Computer Vision (2016), pp. 1-10
Google Scholar

33. [33]
Liu M., Li S., Shan S., Chen X.
Au-aware deep networks for facial expression recognition
Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (2013), pp. 1-6
View in ScopusGoogle Scholar

34. [34]
Zhang T., Zheng W., Cui Z., Zong Y., Yan J., Yan K.
A deep neural network driven feature learning method for multi-view facial expression recognition
IEEE Trans. Multimed., 18 (12) (2016), pp. 2528-2536
View in ScopusGoogle Scholar

35. [35]
Jung H., Lee S., Yim J., S. Park, Kim J.
Joint fine-tuning in deep neural networks for facial expression recognition
Proceedings of the IEEE International Conference on Computer Vision (2015), pp. 2983-2991
View in ScopusGoogle Scholar
36. [36]
A.T. Lopes, E.D. Aguiar, A.F.D. Souza, T. Oliveira-Santos
Facial expression recognition with convolutional neural networks: coping with few data and the training sample order
Pattern Recognit., 61 (2017), pp. 610-628
View PDFView articleView in ScopusGoogle Scholar
37. [37]
A. Majumder, L. Behera, V.K. Subramanian
Automatic facial expression recognition system using deep network-based data fusion
IEEE Trans. Cybern., 48 (1) (2017), pp. 103-114
Google Scholar
38. [38]
D. Hamester, P. Barros, S. Wermter
Face expression recognition with a 2-channel convolutional neural network
Proceedings of the International Joint Conference on Neural Networks (2015), pp. 1-8
View in ScopusGoogle Scholar
39. [39]
S. Alizadeh, A. Fazel
Convolutional neural networks for facial expression recognition
(2016)
, arXiv preprint arXiv:1704.06756
Google Scholar
40. [40]
J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, Zhang N., Tzeng E., T. Darrell
Decaf: a deep convolutional activation feature for generic visual recognition
Proceedings of the International Conference on International Conference on Machine Learning (2014), pp. I-647
Google Scholar
41. [41]
A. Ruiz-Garcia, M. Elshaw, A. Altahhan, V. Palade
Stacked deep convolutional auto-encoders for emotion recognition from facial expressions
Proceedings of the International Joint Conference on Neural Networks (2017), pp. 1586-1593
View in ScopusGoogle Scholar
42. [42]
Yanai K., Y. Kawano
Food image recognition using deep convolutional network with pre-training and fine-tuning
Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (2015), pp. 1-6
Google Scholar
43. [43]
He K., Zhang X., Ren S., Sun J.
Deep residual learning for image recognition
Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern

Recognition (CVPR) (2016), pp. 770-778, 10.1109/CVPR.2016.90
Google Scholar

44. [44]
O. Russakovsky, Deng J., Su H., J. Krause, S. Satheesh, Ma S., Huang Z., A. Karpathy, A. Khosla, M. Bernstein
Imagenet large scale visual recognition challenge
Int. J. Comput. Vis., 115 (3) (2015), pp. 211-252
CrossrefGoogle Scholar

45. [45]
P. Lucey, J.F. Cohn, T. Kanade, J. Saragih
The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression
Proceedings of the Computer Vision and Pattern Recognition Workshops (2010), pp. 94-101
View in ScopusGoogle Scholar

46. [46]
Yin L., Wei X., Sun Y., Wang J., M.J. Rosato
A 3D facial expression database for facial behavior research
Proceedings of the International Conference on Automatic Face and Gesture Recognition (2006), pp. 211-216
View in ScopusGoogle Scholar

47. [47]
I.J. Goodfellow, D. Erhan, C.P. Luc, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Tang Y., D. Thaler, Lee D.H.
Challenges in representation learning: a report on three machine learning contests.
Neural Netw, 64 (2015), pp. 59-63
View PDFView articleView in ScopusGoogle Scholar

48. [48]
Xiang J., Zhu G.
Joint face detection and facial expression recognition with mtcnn
Proceedings of the International Conference on Information Science and Control Engineering (2017), pp. 424-427
CrossrefView in ScopusGoogle Scholar

49. [49]
L. Perez, Wang J.
The effectiveness of data augmentation in image classification using deep learning
(2017)
, arXiv preprint arXiv:1712.04621
Google Scholar

50. [50]
M. Stewart, B.G. Littlewort, I. Fasel, J.R. Movellan
Real time face detection and facial expression recognition: Development and
Proceedings of the Computer Vision and Pattern Recognition Workshop, 2003. CVPRW ?03. Conference on (2003)
53?53
Google Scholar

51. [51]
A. Poursaberi, H.A. Noubari, M. Gavrilova, S.N. Yanushkevich
Gauss?Laguerre wavelet textural feature fusion with geometrical information for facial expression identification
EURASIP J. Image Video Process., 2012 (1) (2012), pp. 1-13
Google Scholar

52. [52]
D. Ghimire, Jeong S., Yoon S., Choi J., Lee J.
Facial expression recognition based on region specific appearance and geometric features
Proceedings of the Tenth International Conference on Digital Information Management (2016), pp. 142-147
Google Scholar

53. [53]
Hu Y., Zeng Z., Yin L., Wei X.
Multi-view facial expression recognition
Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (2008), pp. 1-6
Google Scholar

54. [54]
S. Moore, R. Bowden
Local binary patterns for multi-view facial expression recognition
Comput. Vis. Image Underst., 115 (4) (2011), pp. 541-558
View PDFView articleView in ScopusGoogle Scholar

55. [55]
A. Dapogny, K. Bailly, S. Dubuisson
Dynamic pose-robust facial expression recognition by multi-view pairwise conditional random forests
IEEE Trans. Affect. Comput., PP (99) (2016)
1?1
Google Scholar

56. [56]
Zheng W.
Multi-view facial expression recognition based on group sparse reduced-rank regression
IEEE Trans. Affect. Comput., 5 (1) (2014), pp. 71-85
View in ScopusGoogle Scholar

57. [57]
O. Rudovic, I. Patras, M. Pantic
Coupled gaussian process regression for pose-invariant facial expression recognition
Proceedings of the European Conference on Computer Vision (2010), pp. 350-363
CrossrefView in ScopusGoogle Scholar

58. [58]
Tang Y.
Deep learning using linear support vector machines
(2013)
, arXiv preprint arXiv:1306.0239
Google Scholar

59. [59]
Kim B.K., Roh J., Dong S.Y., Lee S.Y.
Hierarchical committee of deep convolutional neural networks for robust facial expression recognition
J. Multimod. User Interfaces, 10 (2) (2016), pp. 1-17
View in ScopusGoogle Scholar

60. [60]
Jeon J., J.-C. Park, Jo Y., Nam C., Bae K.-H., Hwang Y., Kim D.-S.
A real-time facial expression recognizer using deep neural network
Proceedings of the International Conference on Ubiquitous Information Management & Communication (2016), pp. 1-4, 10.1145/2857546.2857642

Google Scholar
 61. [61]
T. Devries, K. Biswaranjan, G. W. Taylor
Multi-task learning of facial landmarks and expression
Proceedings of the Computer & Robot Vision (2014), pp. 98-103, 10.1109/CRV.2014.21
View in ScopusGoogle Scholar
 62. [62]
Liu K., Zhang M., Pan Z.
Facial expression recognition with CNN ensemble
Proceedings of the International Conference on Cyberworlds (2016), pp. 163-166
CrossrefView in ScopusGoogle Scholar
 63. [63]
Zeng G., Zhou J., Jia X., Xie W., Shen L.
Hand-crafted feature guided deep learning for facial expression recognition
Proceedings of the 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018) (2018), pp. 423-430, 10.1109/FG.2018.00068
View in ScopusGoogle Scholar
 64. [64]
V. Tumen, O.F. Soylemez, B. Ergen
Facial emotion recognition on a dataset using convolutional neural network
Proceedings of the Artificial Intelligence and Data Processing Symposium (2017), pp. 1-5
Google Scholar

## Cited by (148)

 * ### Facial expression recognition with grid-wise attention and visual transformer
2021, Information Sciences
Citation Excerpt :
Kumar et al. [19] used CNN-based Kinect APIs to extracted 71 facial points to represent facial expressions for gesture recognition. Shao et al. [33] combined three kinds of convolutional neural networks, i.e., shallow CNN, dual branch CNN, and transfer-learning-based CNN, for robust facial expression recognition in the wild. Hossain et al. [15] used a pretrained CNN model and two deep sparse auto-encoders to extract facial and speech features, and employed a support vector machine to determine a corresponding emotion label under a secure edge and cloud computing environment.
Show abstract
 _F_ acial _E_ xpression _R_ ecognition (FER) has achieved remarkable progress as a result of using _C_ onvolutional _N_ eural _N_ etworks (CNN). Relying on the spatial locality, convolutional filters in CNN, however, fail to learn long-range inductive biases between different facial regions in most neural layers. As such, the performance of a CNN-based model for FER is still limited. To address this problem, this paper introduces a novel FER framework with two attention mechanisms for CNN-based models, and these two attention mechanisms are used for the low-level feature learning the high-level semantic representation, respectively. In particular, in the low-level feature learning, a grid-wise attention mechanism is proposed to capture the dependencies of different regions from a facial expression image such that the parameter update of convolutional filters in low-level feature learning is regularized. In the high-level semantic representation, a visual transformer attention mechanism uses a sequence of visual semantic tokens (generated from pyramid features of high convolutional layer blocks) to learn the global representation. Extensive experiments have been conducted on three public facial expression datasets, CK+, FER+, and RAF-DB. The results show that our

FER-VT has achieved state-of-the-art performance on these datasets, especially with a 100% accuracy on CK + datasets without any extra training data.

* ### Attention mechanism-based CNN for facial expression recognition
2020, Neurocomputing
Citation Excerpt :

In this paper, for each kind of these six expressions, we select the last three frames with peak information as our new dataset. The comparison results of our method and some representative methods [20,30,33,35,37] on this dataset are listed in Table 5, which indicate that our method performs better than most of the methods except Zhang et al. [37]. However, the multitask network in [37] learns from some auxiliary attributes like gender, age and head pose, except for facial expression images.

Show abstract

Facial expression recognition is a hot research topic and can be applied in many computer vision fields, such as human?computer interaction, affective computing and so on. In this paper, we propose a novel end-to-end network with attention mechanism for automatic facial expression recognition. The new network architecture consists of four parts, i.e., the feature extraction module, the attention module, the reconstruction module and the classification module. The LBP features extract image texture information and then catch the small movements of the faces, which can improve the network performance. Attention mechanism can make the neural network pay more attention to useful features. We combine LBP features and attention mechanism to enhance the attention model to obtain better results. In addition, we collected and labelled a new facial expression dataset of seven expressions from 35 subjects aged from 20 to 25. For each subject, we captured both RGB images and depth images with a Microsoft Kinect sensor. For each image type, there are 245 image sequences, each of which contains 110 images, resulting in 26,950 images in total. We apply the newly proposed method to our own dataset and four representative expression datasets, i.e., JAFFE, CK+, FER2013 and Oulu-CASIA. The experimental results demonstrate the feasibility and effectiveness of the proposed method.

* ### Deep convolution network based emotion analysis towards mental health care
2020, Neurocomputing

Show abstract

Facial expressions play an important role during communications, allowing information regarding the emotional state of an individual to be conveyed and inferred. Research suggests that automatic facial expression recognition is a promising avenue of enquiry in mental healthcare, as facial expressions can also reflect an individual's mental state. In order to develop user-friendly, low-cost and effective facial expression analysis systems for mental health care, this paper presents a novel deep convolution network based emotion analysis framework to support mental state detection and diagnosis. The proposed system is able to process facial images and interpret the temporal evolution of emotions through a new solution in which deep features are extracted from the Fully Connected Layer 6 of the AlexNet, with a standard Linear Discriminant Analysis Classifier exploited to obtain the final classification outcome. It is tested against 5 benchmarking databases, including JAFFE, KDEF,CK+, and databases with the images obtained ?in the wild? such as FER2013 and AffectNet. Compared with the other state-of-the-art methods, we observe that our method has overall higher accuracy of facial expression recognition. Additionally, when compared to the state-of-the-art deep learning algorithms such as Vgg16, GoogleNet, ResNet and AlexNet, the proposed method demonstrated better efficiency and has less device requirements. The experiments presented in this paper demonstrate that the proposed method outperforms the other methods in terms of accuracy and efficiency which suggests it could act as a smart, low-cost, user-friendly cognitive aid to detect, monitor, and diagnose the mental health of a patient through automatic facial expression analysis.

* ### Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion
2023, IEEE Transactions on Affective Computing
* ### Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild
2022, IEEE Access
* ### FER-net: facial expression recognition using deep neural net
2021, Neural Computing and Applications

View all citing articles on Scopus

Jie Shao received the B.S. and M.S. degree in Nanjing University of Aeronauticsand Astronautics. Thenshe got her Ph.D. in Tongji University. At present, she is an associate professor in Shanghai University of Electric Power. Her currentresearch interest includes computer vision, video surveillance, and human emotion analysis.

Yongsheng Qian received his bachelor?s degree in electrical engineering and automation from Hubei University for Nationalities in 2015. He is currently a graduate student in the department of electronics and information engineering in Shanghai University of Electric Power, Shanghai, China. His research interest includes facial expression recognition and deep learning.

View Abstract

## Recommended articles

* ### OFF-ApexNet on micro-expression recognition system
Signal Processing: Image Communication, Volume 74, 2019, pp. 129-139
Y.S. Gan, ?, Lit-Ken Tan
View PDF
* ### Facial expression recognition via learning deep sparse autoencoders
Neurocomputing, Volume 273, 2018, pp. 643-649
Nianyin Zeng, ?, Abdullah M. Dobaie
View PDF
* ### Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks
Journal of Visual Communication and Image Representation, Volume 59, 2019, pp. 176-185
Min Hu, ?, Ronggui Wang
View PDF
* ### Facial expression recognition with local prominent directional pattern
Signal Processing: Image Communication, Volume 74, 2019, pp. 1-12
Farkhod Makhmudkhujaev, ?, Oksam Chae
View PDF
* ### MiniExpNet: A small and effective facial expression recognition network based on facial local regions
Neurocomputing, Volume 462, 2021, pp. 353-364
Xing Jin, Zhong Jin
View PDF
* ### Anatomical and neurophysiological basis of face recognition
Revue Neurologique, Volume 178, Issue 7, 2022, pp. 649-653
F. Sellal
View PDF

Show 3 more articles

## Article Metrics

Citations

 * Citation Indexes: 148
Captures
 * Readers: 135
Mentions
 * News Mentions: 1
View details

Cookies are used by this site. Cookie Settings

## Cookie Preference Center

We use cookies which are necessary to make our site work. We may also use additional cookies to analyse, improve and personalise our content and your digital experience. For more information, see our Cookie Policy and the list of Google Ad-Tech Vendors.

You may choose not to allow some types of cookies. However, blocking some types may impact your experience of our site and the services we are able to offer. See the different category headings below to find out more or change your settings.

Allow all

### Manage Consent Preferences

#### Strictly Necessary Cookies

Always active

These cookies are necessary for the website to function and cannot be switched off in our systems. They are usually only set in response to actions made by you which amount to a request for services, such as setting your privacy preferences, logging in or filling in forms. You can set your browser to block or alert you about these cookies, but some parts of the site will not then work. These cookies do not store any personally identifiable information.

Cookie Details List?

#### Functional Cookies

Functional Cookies

These cookies enable the website to provide enhanced functionality and personalisation. They may be set by us or by third party providers whose services we have added to our pages. If you do not allow these cookies then some or all of these services may not function properly.

Cookie Details List?

#### Performance Cookies

Performance Cookies

These cookies allow us to count visits and traffic sources so we can measure and improve the performance of our site. They help us to know which pages are the most and least popular and see how visitors move around the site.

Cookie Details List?

#### Targeting Cookies

Targeting Cookies

These cookies may be set through our site by our advertising partners. They may be used by those companies to build a profile of your interests and show you relevant adverts on other sites. If you do not allow these cookies, you will experience less targeted advertising.

Cookie Details List?

Back Button

### Cookie List

Search Icon

Filter Icon

Clear

checkbox label label

Apply Cancel

Consent Leg.Interest

checkbox label label

checkbox label label

checkbox label label

Confirm my choices