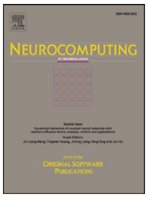
Neurocomputing 355 (2019) 82–92

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



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

Jie Shao, Yongsheng Qian*

College of Electronic and Information Engineering, Shanghai University of Electric Power, China

ARTICLE INFO

Article history:
Received 9 August 2018
Revised 9 December 2018
Accepted 6 May 2019
Available online 10 May 2019

Communicated by Rongni Yang

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

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.

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

* Corresponding author.
E-mail address: qian_yongsheng@126.com (Y. Qian).

https://doi.org/10.1016/j.neucom.2019.05.005
0925-2312/© 2019 Elsevier B.V. All rights reserved.

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 [17–19].

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 [20–22] 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 [32–36]. 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

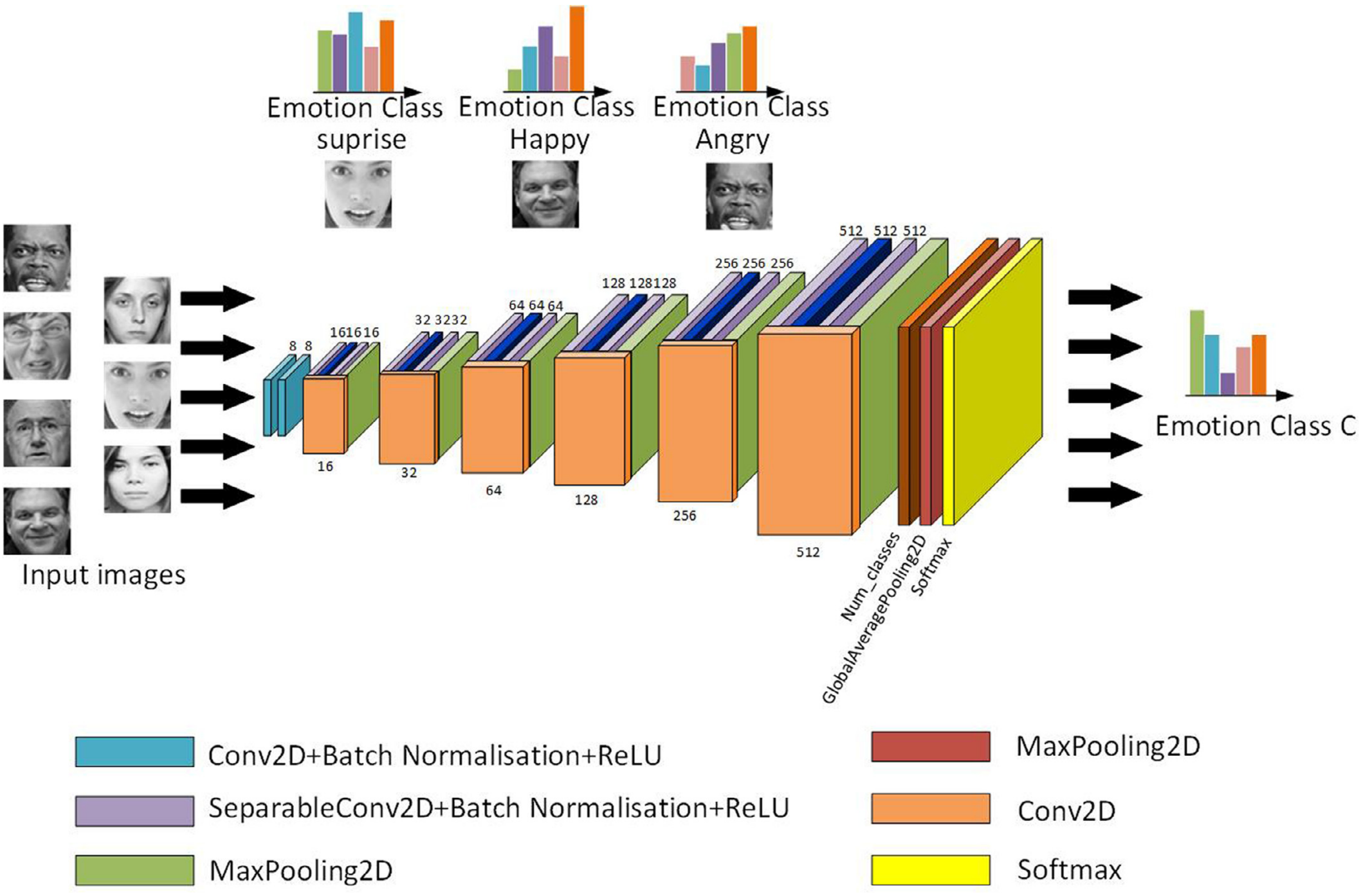


Fig. 1. The basic structure of the Light-CNN.

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 pretrained 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 $64 \times 64 \times 1$ 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

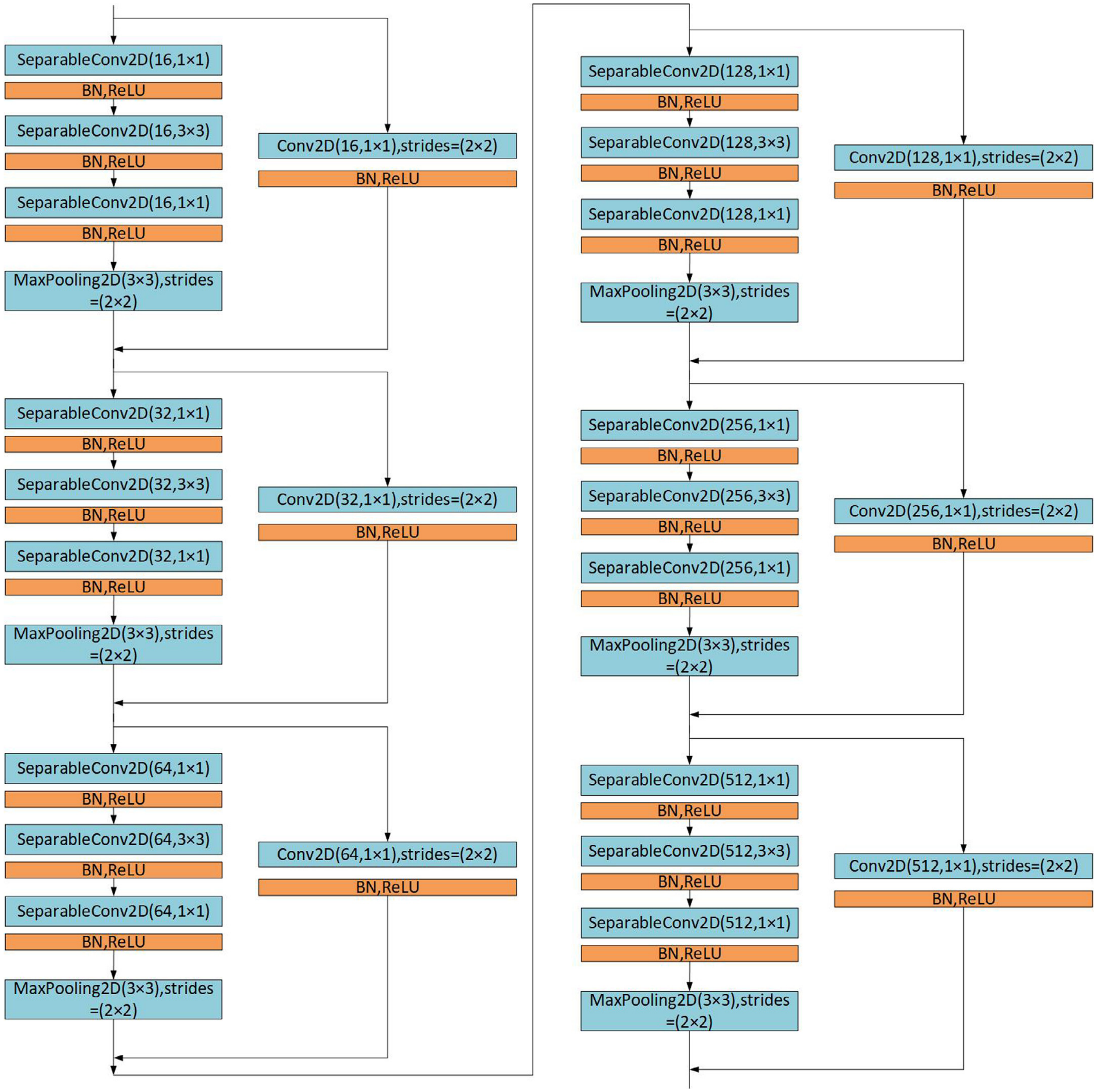


Fig. 2. The structure of 6 depthwise separable residual convolution modules.

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.

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 $x^{(i)} \in \Re^{n+1}$, and the sample tag is $y^{(i)}$, so the training set $S = \left\{ \left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(m)}, y^{(m)}\right) \right\}$ of the supervised learning constitutes the classification layer. Then the function and cost function forms are as follows:

$$h_\theta\left(x^{(i)}\right) = \begin{bmatrix} p\left(y^{(i)} = 1 | x^{(i)}; \theta\right) \\ p\left(y^{(i)} = 2 | x^{(i)}; \theta\right) \\ \vdots \\ p\left(y^{(i)} = k | x^{(i)}; \theta\right) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_k^T x^{(i)}} \end{bmatrix} \tag{1}$$

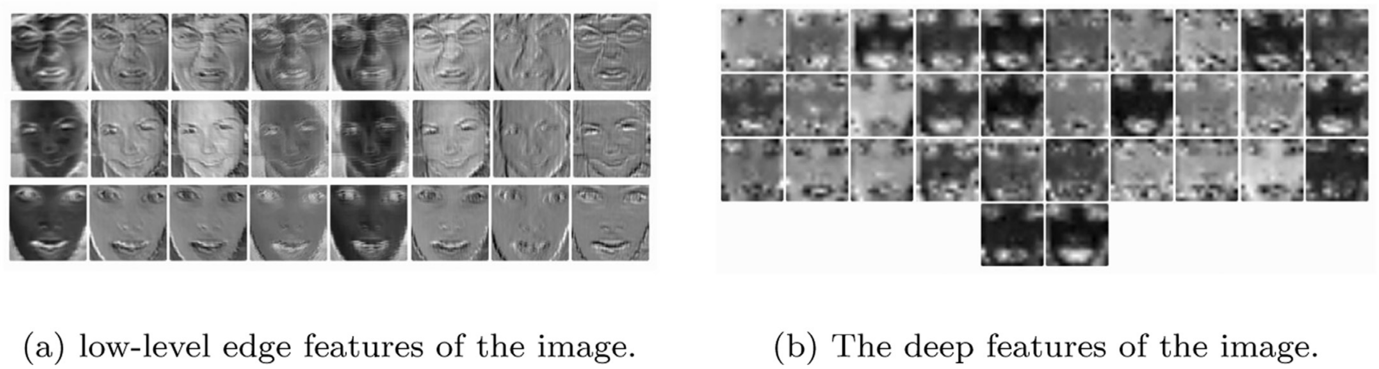

(a) low-level edge features of the image.
(b) The deep features of the image.

Fig. 3. Feature visualization of the image.

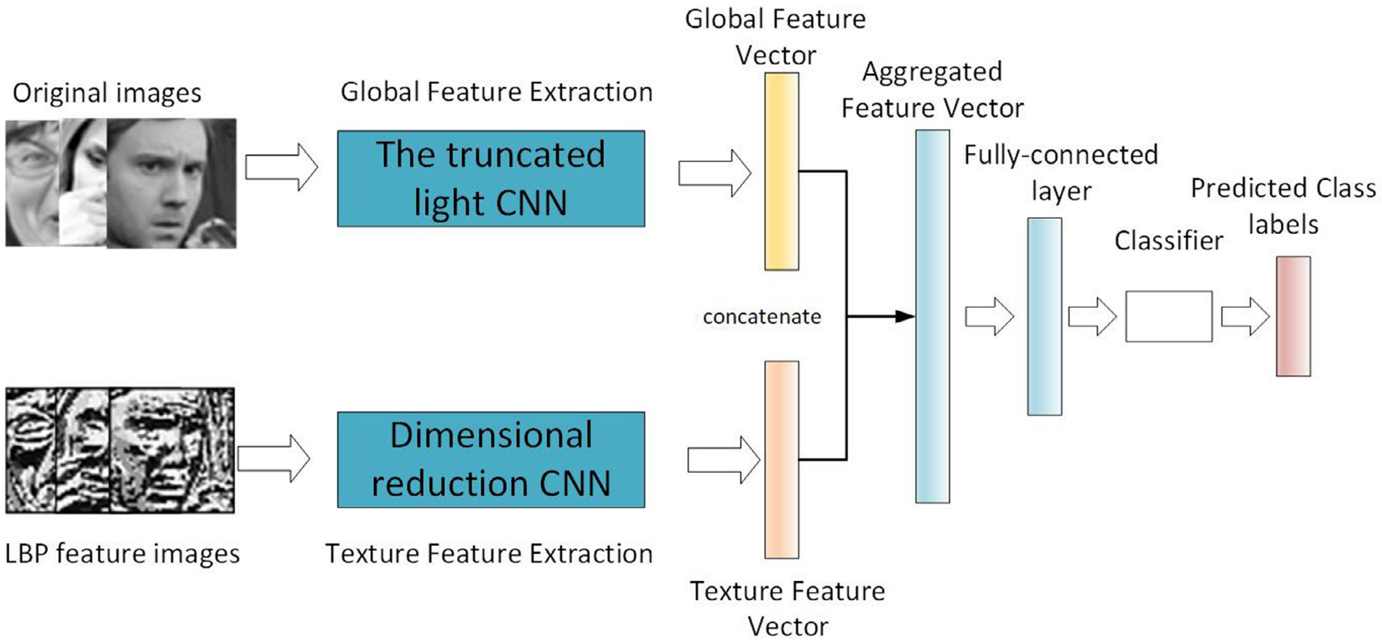


Fig. 4. Framework of the dual-branch CNN.

Where $\theta_1, \theta_2, \ldots, \theta_k \in \Re^{n+1}$ is the model parameter and $\frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}}$ is the normalization term for the probability distribution, making the sum of all probabilities equal to 1.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(i)} = j\right\} \ln \frac{\theta_j^T x^{(i)}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}\right] \tag{2}$$

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.

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.

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

Table 1
Architectures of the two branches.

The first branch		Layer1	–			Layer28	Layer29
	type	Input	–			GlobalAveragePooling2D	Flatten
	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

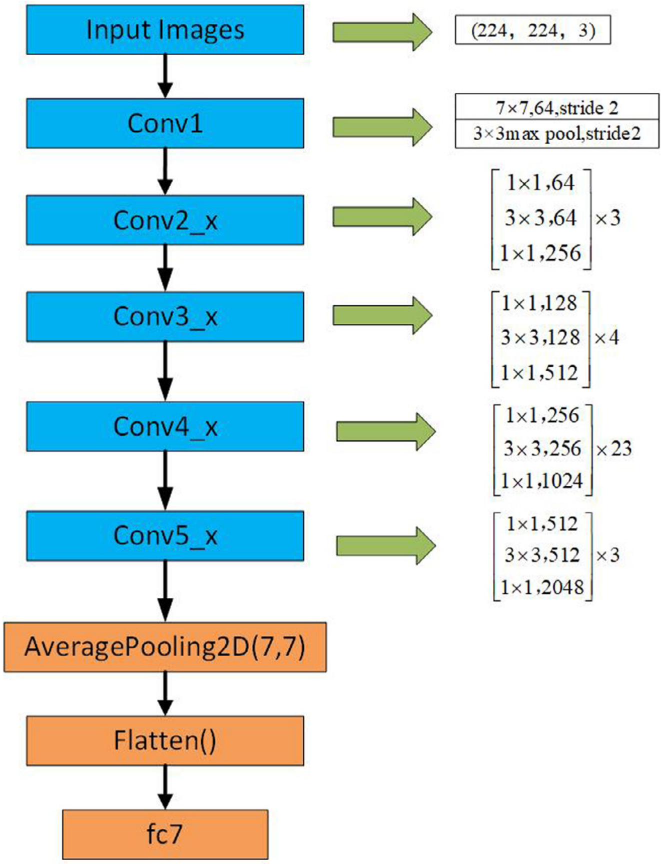


Fig. 5. The framework of the pretrained CNN.

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

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.

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.

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.

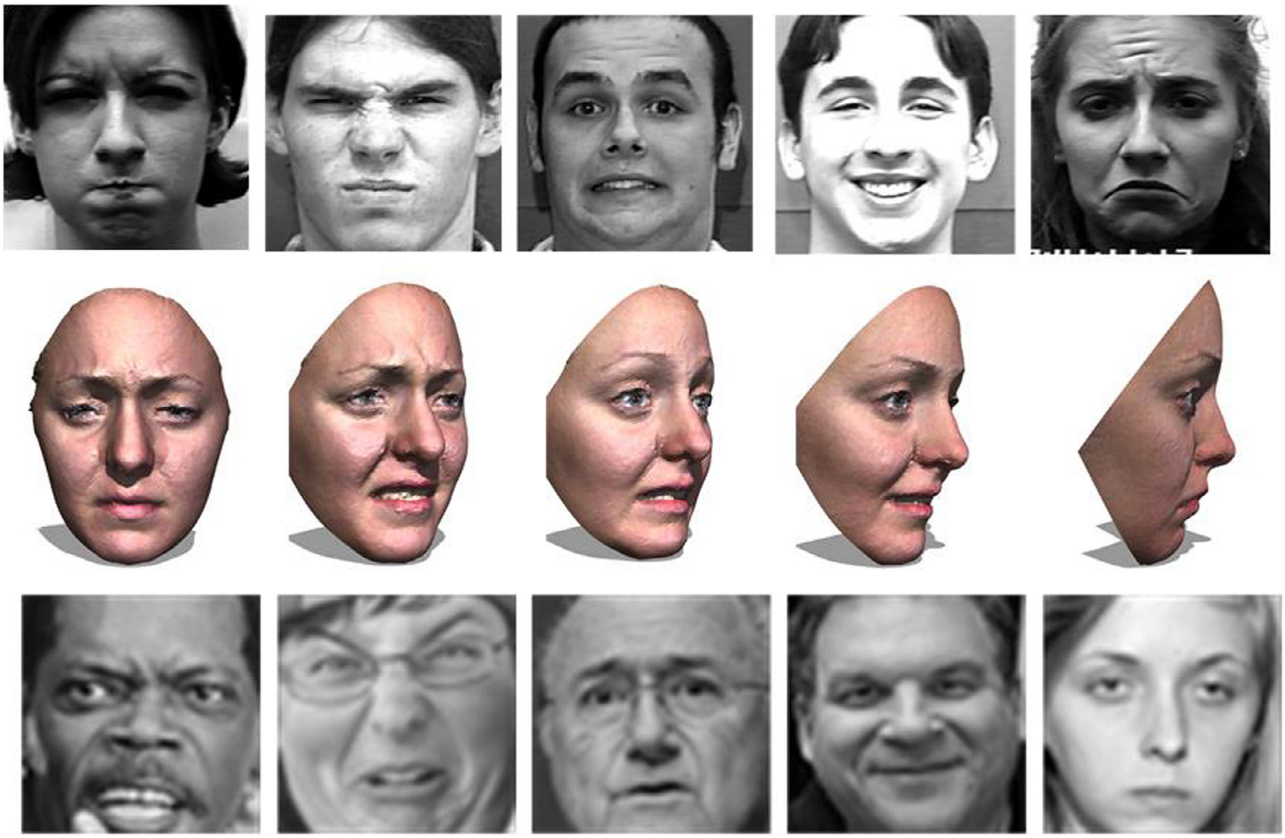


Fig. 6. Examples Images in CK+(top), BU-3DFE (middle), FER2013 (bottom) Datasets.

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.

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

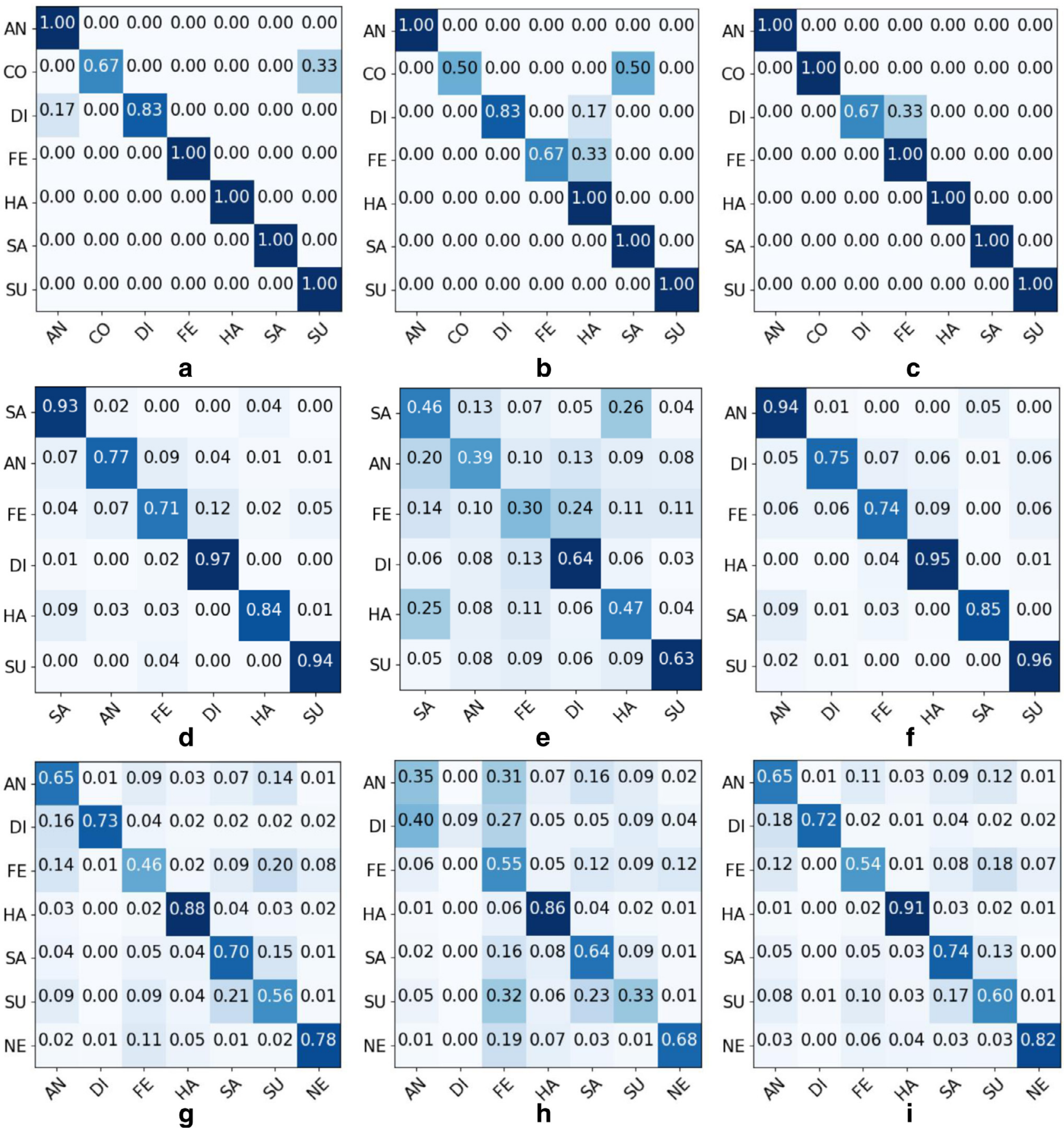


Fig. 7. Confusion matrices for three networks on three expression databases. (a)–(c) are confusion matrices for the Light-CNNs, dual-branch CNN and the pretrained CNN on CK+, (d)–(f) are confusion matrices for the three networks on BU-3DFE, (g)–(i) are confusion matrices for the three networks on FER2013.

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

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.

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

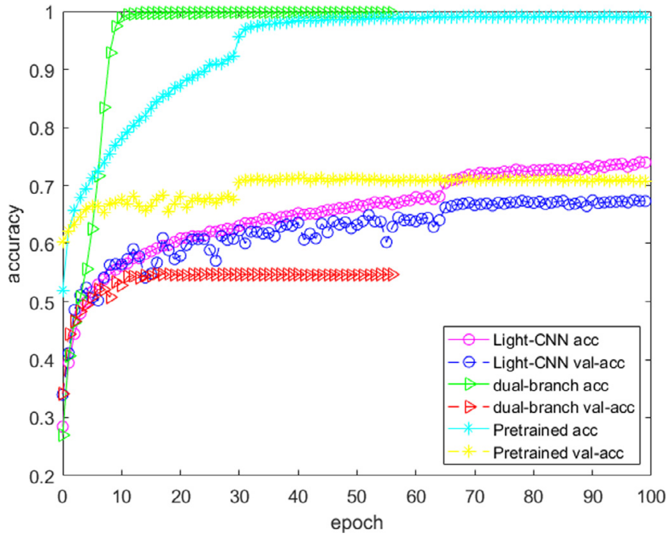


Fig. 8. Comparison of parameters and running time on FER2013 dataset.

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**

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**

used. However, the recognition accuracy of most traditional methods are lower than that of deep learning.

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 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**

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.

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).

References

[1] R. Wang, B. Fang, Affective computing and biometrics based HCI surveillance system, in: Proceedings of the International Symposium on Information Science and Engineering, 2008, pp. 192–195.

[2] W. Weiguo, M. Qingmei, W. Yu, Development of the humanoid head portrait robot system with flexible face and expression, in: Proceedings of the 2004 IEEE International Conference on Robotics and Biomimetics, 2004, pp. 757–762, doi:10.1109/ROBIO.2004.1521877.

[3] M.H. Su, C.H. Wu, K.Y. Huang, Q.B. Hong, H.M. Wang, Exploring microscopic fluctuation of facial expression for mood disorder classification, in: Proceedings of the International Conference on Orange Technologies, 2017, pp. 65–69.

[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.

[5] S.A. Patil, P.J. Deore, Local binary pattern based face recognition system for automotive security, in: Proceedings of the International Conference on Signal Processing, Computing and Control, 2016, pp. 13–17.

[6] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, D.N. Metaxas, Learning multiscale active facial patches for expression analysis, in: Proceedings of the Computer Vision and Pattern Recognition, 2012, pp. 2562–2569.

[7] H. Zheng, X. Geng, D. Tao, Z. Jin, A multi-task model for simultaneous face identification and facial expression recognition, Neurocomputing 171 (C) (2016) 515–523.

[8] T. Qi, X. Yong, Y. Quan, Y. Wang, H. Ling, Image-based action recognition using hint-enhanced deep neural networks, Neurocomputing 267 (2017) 475–488.

[9] M. Morchid, Parsimonious memory unit for recurrent neural networks with application to natural language processing, Neurocomputing 314 (2018) 48–64.

[10] L. Yu, X. Shao, X. Yan, 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.

[11] K. Yang, X. Gong, Y. Liu, Z. Li, T. Xing, X. Chen, D. Fang, 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.

[12] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of the ECCV, 2014.

[13] A. Azulay, Y. Weiss, Why do deep convolutional networks generalize so poorly to small image transformations? 2018, arXiv preprint arXiv:1805.12177.

[14] G. Huang, Z. Liu, 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.

[15] G. Pleiss, D. Chen, G. Huang, T. Li, L. van der Maaten, K.Q. Weinberger, Memory-efficient implementation of densenets, 2017, arXiv preprint arXiv:1707.06990.

[16] C. Shan, S. Gong, P.W. Mcowan, Robust facial expression recognition using local binary patterns, in: Proceedings of the IEEE International Conference on Image Processing, 2005.

[17] M. Xu, W. Cheng, Q. Zhao, L. Ma, F. Xu, Facial expression recognition based on transfer learning from deep convolutional networks, in: Proceedings of the International Conference on Natural Computation, 2016, pp. 702–708.

[18] J. Luttrell, Z. Zhou, Y. Zhang, C. Zhang, P. Gong, B. Yang, R. Li, A deep transfer learning approach to fine-tuning facial recognition models, in: Proceedings of the 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2018, pp. 2671–2676, doi:10.1109/ICIEA.2018.8398162.

[19] M. Peng, Z. Wu, Z. Zhang, T. Chen, From macro to micro expression recognition: deep learning on small datasets using transfer learning, in: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, 2018, pp. 657–661.

[20] M. Jian, K.M. Lam, J. Dong, Facial-feature detection and localization based on a hierarchical scheme, Inf. Sci. 262 (3) (2014) 1–14.

[21] G. Sikander, S. Anwar, Y.A. Djawad, Facial feature detection: a facial symmetry approach, in: Proceedings of the International Symposium on Computational and Business Intelligence, 2017, pp. 26–31.

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

[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, in: Proceedings of the 2017 2nd International Conference on Electrical Electronic Engineering (ICEEE), 2017, pp. 1–4, doi:10.1109/CEEE.2017.8412888.

[24] Z. Ying, X. Fang, Combining LBP and adaboost for facial expression recognition, in: Proceedings of the 9th international conference on signal processing, 2008, pp. 1461–1464.

[25] Y. Wang, H. Yu, B. Stevens, H. Liu, Dynamic facial expression recognition using local patch and LBP-top, in: Proceedings of the International Conference on Human System Interactions, 2015, pp. 362–367.

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

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

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

[29] M. Jian, K.M. Lam, Face-image retrieval based on singular values and potential-field representation, Signal Process. 100 (7) (2014) 9–15.

[30] M. Jian, K.M. Lam, J. Dong, A novel face-hallucination scheme based on singular value decomposition, Pattern Recognit. 46 (11) (2013) 3091–3102.

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

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

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

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

[35] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2983–2991.

[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) 610–628.

[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) 103–114.

[38] D. Hamester, P. Barros, S. Wermter, Face expression recognition with a 2-channel convolutional neural network, in: Proceedings of the International Joint Conference on Neural Networks, 2015, pp. 1–8.

[39] S. Alizadeh, A. Fazel, Convolutional neural networks for facial expression recognition, 2016, arXiv preprint arXiv:1704.06756.

[40] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: a deep convolutional activation feature for generic visual recognition, in: Proceedings of the International Conference on International Conference on Machine Learning, 2014, pp. I–647.

[41] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, V. Palade, Stacked deep convolutional auto-encoders for emotion recognition from facial expressions, in: Proceedings of the International Joint Conference on Neural Networks, 2017, pp. 1586–1593.

[42] K. Yanai, Y. Kawano, Food image recognition using deep convolutional network with pre-training and fine-tuning, in: Proceedings of the IEEE International Conference on Multimedia & Expo Workshops, 2015, pp. 1–6.

[43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi:10.1109/CVPR.2016.90.

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

[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, in: Proceedings of the Computer Vision and Pattern Recognition Workshops, 2010, pp. 94–101.

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

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

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

[49] L. Perez, J. Wang, The effectiveness of data augmentation in image classification using deep learning, 2017, arXiv preprint arXiv:1712.04621.
[50] M. Stewart, B.G. Littlewort, I. Fasel, J.R. Movellan, Real time face detection and facial expression recognition: Development and, in: Proceedings of the Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on, 2003 53–53.
[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) 1–13.
[52] D. Ghimire, S. Jeong, S. Yoon, J. Choi, J. Lee, Facial expression recognition based on region specific appearance and geometric features, in: Proceedings of the Tenth International Conference on Digital Information Management, 2016, pp. 142–147.
[53] Y. Hu, Z. Zeng, L. Yin, X. Wei, Multi-view facial expression recognition, in: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, 2008, pp. 1–6.
[54] S. Moore, R. Bowden, Local binary patterns for multi-view facial expression recognition, Comput. Vis. Image Underst. 115 (4) (2011) 541–558.
[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.
[56] W. Zheng, Multi-view facial expression recognition based on group sparse reduced-rank regression, IEEE Trans. Affect. Comput. 5 (1) (2014) 71–85.
[57] O. Rudovic, I. Patras, M. Pantic, Coupled gaussian process regression for pose-invariant facial expression recognition, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 350–363.
[58] Y. Tang, Deep learning using linear support vector machines, 2013, arXiv preprint arXiv:1306.0239.
[59] B.K. Kim, J. Roh, S.Y. Dong, S.Y. Lee, Hierarchical committee of deep convolutional neural networks for robust facial expression recognition, J. Multimod. User Interfaces 10 (2) (2016) 1–17.
[60] J. Jeon, J.-C. Park, Y. Jo, C. Nam, K.-H. Bae, Y. Hwang, D.-S. Kim, A real-time facial expression recognizer using deep neural network, in: Proceedings of the International Conference on Ubiquitous Information Management & Communication, 2016, pp. 1–4, doi:10.1145/2857546.2857642.
[61] T. Devries, K. Biswaranjan, G. W. Taylor, Multi-task learning of facial landmarks and expression, in: Proceedings of the Computer & Robot Vision, 2014, pp. 98–103, doi:10.1109/CRV.2014.21.
[62] K. Liu, M. Zhang, Z. Pan, Facial expression recognition with CNN ensemble, in: Proceedings of the International Conference on Cyberworlds, 2016, pp. 163–166.
[63] G. Zeng, J. Zhou, X. Jia, W. Xie, L. Shen, Hand-crafted feature guided deep learning for facial expression recognition, in: Proceedings of the 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), 2018, pp. 423–430, doi:10.1109/FG.2018.00068.
[64] V. Tumen, O.F. Soylemez, B. Ergen, Facial emotion recognition on a dataset using convolutional neural network, in: Proceedings of the Artificial Intelligence and Data Processing Symposium, 2017, pp. 1–5.

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.