

Attention mechanism-based CNN for facial expression recognition

Jing Li^a, Kan Jin^a, Dalin Zhou^b, Naoyuki Kubota^c, Zhaojie Ju^{b,*}

^a School of Information Engineering, Nanchang University, Nanchang 330031, China

^b School of Computing, University of Portsmouth, Portsmouth PO1 3HE, UK

^c Graduate School of Systems Design, Tokyo Metropolitan University, China

ARTICLE INFO

Article history:

Received 5 January 2020

Revised 23 April 2020

Accepted 2 June 2020

Available online 12 June 2020

Communicated by Y. Liu

Keywords:

Facial Expression Recognition

Convolutional Neural Network

Attention Mechanism

Local Binary Pattern

Image Classification

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.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

Facial expression is one of the most direct signals to express inner feelings in people's daily communication. The physical or mental state of a person at one time can be obtained by analyzing facial expressions. Therefore, facial expression recognition is of great significance in autopilot, human–computer interaction, medical treatment and other fields related to facial expression, and has gradually become a more and more important research direction. In machine learning, a variety of facial expression recognition algorithms have been proposed. Due to the complexity, diversity, occlusion, lighting and other challenges in facial expression recognition, the recognition accuracy in practical applications is still unsatisfactory.

In this paper, our goal is to design a recognition model that can automatically and accurately recognize different expressions in various types of images. Generally, the process of facial expression recognition consists of the following steps: i) pre-processing of the facial expression data; ii) feature extraction of facial expressions; and iii) classification of facial expressions. The process is depicted

in Fig. 1. We usually consider two kinds of features, namely, facial features and face model features. The facial features are specific points on the face, like eyes, mouth, and eyebrows; the face model features are the features used to model the face. Therefore, there are several ways for facial representation, like using the whole face to get the holistic representation, using specific points for local representation, and combining different points to get a hybrid approach. The final step is to define some set of categories to which the expression belongs.

When dealing with expression recognition as a classification problem, traditional methods often use hand-crafted features such as Local Binary Patterns (LBP) and traditional machine learning algorithms such as Support Vector Machine (SVM) to classify. These methods may work well on datasets collected under laboratory conditions, but with the introduction of more challenging expression datasets in uncontrollable environments (e.g., FER2013), they cannot effectively achieve this task. Fortunately, deep learning has made a breakthrough in convenience and effectiveness since it has been used to deal with the image classification problem.

The attention mechanism has been widely used in various computer vision tasks such as saliency detection [15], crowd counting [16] and facial expression recognition [38]. The operation can

* Corresponding author.

E-mail address: zhaojie.ju@port.ac.uk (Z. Ju).

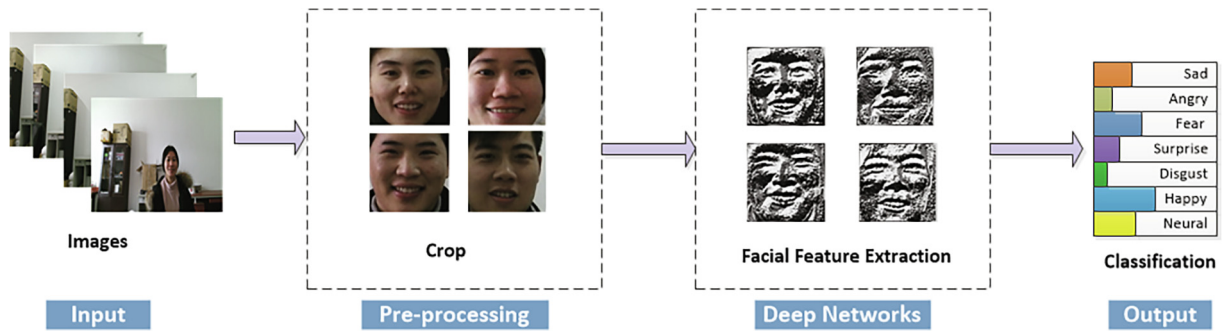


Fig. 1. The main steps of facial expression recognition.

select the most useful features for classification by learning an intermediate attention map and then applying element-wise product on attention maps and source feature maps to weight the importance of different features. For the task of facial expression recognition, the features that are useful for recognition are mainly in some key parts such as eyes, nose and mouth. The attention mechanism increases the weights of these key features and helps improve the expression recognition results.

In this paper, we design a novel Convolutional Neural Network with an attention model for recognizing facial expressions. In [43], it showed that using LBP features is better than using HOG and Gabor features because LBP can achieve rotation invariance and grey-scale invariance and thus is suitable for extracting texture features at different scales and can solve the imbalance of displacement, rotation angles and illumination conditions in facial images. In addition, LBP features can reflect fine facial changes in skin textures like wrinkles and furrows, which shows the changes of expressions. In [38], an end-to-end network with an attention model was presented for facial expression recognition. The attention module makes the network focus more on useful features which are vital for expression recognition by increasing the weights of these features. This makes the network recognize expressions more efficiently. Inspired by [38,43], we combine LBP features with an attention model for facial expression recognition. Embedding the attention model into the network allows the network to pay different attention and weight to different parts of the input data. This can make the neural network pay more attention to useful features, which is vital to expression recognition. Furthermore, we combine LBP features with convolution features to improve our recognition results. The proposed method has been tested on five facial expression datasets, which are CK+ [11], JAFFE [13], FER2013 [39], Oulu-CASIA [25], and our self-collected Nanchang University Facial Expression (NCUFE).

To verify the effectiveness of our algorithm, we collected a new facial expression dataset called NCUFE. The dataset consists of seven expressions (i.e., anger, disgust, fear, happiness, sadness, surprise and neutral). We collected these facial expression images

from 35 graduate students (6 females and 29 males) by a Microsoft Kinect sensor for acquiring both RGB images and depth images. The sample images are shown in Fig. 2. For each student, the size of these two types of images is 1280×1024 and 512×424 , respectively. For each image type, there are 245 image sequences, each of which contains 110 images, resulting in 26,950 images in total. During the process of image capture, the students sat on a chair in front of the Kinect and faced the camera. The distance between the face and the Kinect was about 100 cm. We asked each student to look the expression examples printed on some pieces of paper and then make seven expressions.

The main contributions of this paper are as follows:

- 1) We introduce a novel facial expression recognition method with attention mechanism. Not only raw images, but also LBP features are added to the attention layers of the network. LBP features contain texture information and can reflect fine facial changes in skin textures, which can help distinguish expressions with subtle difference.
- 2) We collected and labelled a new dataset named Nanchang University Facial Expression (NCUFE) for facial expression recognition. The dataset includes 490 image sequence collected from 35 subjects labeled with seven facial expressions (i.e., anger, disgust, fear, happiness, sadness, surprise and neutral). For each subject, we captured both RGB images and depth images.
- 3) We implement substantial experiments on five different datasets, as shown in Fig. 3. There are not only datasets collected under laboratory conditions, such as CK+, JAFFE, Oulu-CASIA and NCUFE, but also those collected in real world like FER2013. We also compare the model performance with some state-of-the-art expression recognition algorithms, and the results show that our model is superior.

The remainder of this paper is as follows. In Section 2, we introduce the related works in expression recognition and the existing algorithms. In Section 3, we describe the proposed method in



Fig. 2. Sample images in the NCUFE dataset. The left one is an RGB image, and the right one is its corresponding depth image.



Fig. 3. Examples of the datasets used in our experiments, from top to bottom is from CK+ [11], JAFFE [13], Oulu-CASIA [25], FER2013 [39], and our dataset NCUFE.

detail. We describe our experimental process and results on different datasets and compare them with the results of the state-of-the-art algorithms in Section 4. We give a conclusion of the whole paper in Section 5.

2. Related works

Traditional facial feature extraction algorithms can be separated into two categories: 1) geometric-based methods, such as Active Appearance Models (AAM) [17]; and 2) appearance-based methods, such as LBP [9] and Gabor Wavelet Representation [21]. After feature description, the features are fed into a classifier, such as SVM [22] and K-nearest Neighbors (KNN) [24], for recognizing different facial expressions. Therefore, the performance of the classifier depends to a large extent on the quality of the extracted features. In [43], various feature extraction techniques combined with different classification algorithms were presented to find the best combination that can be used for emotion intensity recognition. The results with LBP features are better than those using HOG and Gabor features.

The CK+ dataset [11] contains emotion annotations as well as action unit annotations. In the classification stage, the dataset was evaluated by AAM and SVM. AAM tracks the face and extracts facial features, and then SVM classifies the facial expressions. For each expression, the architecture achieves more than 65% accuracy, and the best recognition accuracy is 100% for the happy emotion.

In recent years, deep neural network has become popular in recognizing facial expressions and other computer vision tasks. In [36], LeNet-5, which is the earliest convolutional neural network, is presented to recognize handwriting. There are only seven layers in the network, including three convolutional layers, two sub-sampling layers and two fully connected layers. Then, many variants of networks based on this basic design are prevalent in deep learning tasks. VGG Net [32] used very small convolution filters (3×3) to increase the architecture depth, where the small-size filter can make the decision function more discriminative and decrease the number of parameters. The network often stacked several convolutional layers and then followed one pooling layers. When there are 16 or 19 wt layers in the network, the architecture can achieve significant improvement on the prior-art configurations. GoogLeNet [3] is a 22-layer deep network. Not only the width, but also the depth of the network is increased compared with previous networks. The main structure of the network is “Inception” layers, which contain several parallel convolution branches. “Inception” layers have difference sizes of convolution filters, and the input images can convolve at different scales of feature maps. In ResNet [4], skip connections are added between the input layer and output layer in the network. This structure not only increases the training speed and improves the training effect of the model, but also avoids gradient disappearance and network degradation.

For the facial expression recognition task, most works were inspired by the above-mentioned deep network architectures.

When classifying static images, a facial expression recognition network [2] inspired by GoogleNet was proposed. The network includes two convolutional layers, each of which is followed by a max-pooling layer, and then four inception layers are followed. Sun et al. [12] proposed a facial expression recognition network with visual attention. In this network, the deep convolution features are extracted from the face, and thus the regions of interest are detected and used to classify expressions. In [5], binarized Auto-encoders and Stacked Binarized Auto-encoders were used to learn a type of domain knowledge from unlabeled facial expression datasets. Fernandez et al. [38] proposed an end-to-end network with an attention model for facial expression recognition. The results showed that the attention module improves the classification performance. In [1], a GAN-based face frontalization method was presented. The input face images are frontalized by the generator and the identity and expression characteristics are preserved at the same time. Then, the discriminator makes a distinction between the real images and the generated face images.

While classifying video images, in [7], a DNN-based architecture combined with conditional Random Field was proposed to solve the expression recognition problem. Here, the Inception-ResNet networks [6] is used in the networks for facial expression recognition, the experimental results show good performance on CK+, MMI and FER2013. In [33], VGG-based convolutional neural network was first used to learn facial features, which are then linked to a Long Short Term Memory (LSTM) to exploit the temporal relation between video frames. An accuracy of 97.2% was obtained on CK+. A network termed Spatio-Temporal Convolutional features with Nested LSTM (STC-NLSTM) [8] was proposed to learn multi-level facial expression features and temporal dynamics of facial expressions in a joint way.

In this paper, a novel convolutional neural network with attention mechanism is presented to classify different facial expressions. We combine LBP features and convolution features in an attention module. With the help of LBP features which provide texture information and can reflect fine changes on the face, the abil-

ity of the attention module can be improved so as to improve the recognition accuracy of the network. In addition, in order to avoid overfitting, data augmentation is applied in the datasets used in our experiments. We also use the batch normalization after each layer to speed up the convergence of the network.

3. The proposed method

3.1. Network architecture

In this section, we introduce our newly proposed convolutional neural network with attention mechanism for automatically recognizing facial expressions. The new network consists of four parts, i.e., the feature extraction module, the attention module, the reconstruction module and the classification module. The architecture starts from the feature extraction module composed of two separate CNN processing streams: one is for raw images and the other is for LBP feature maps. Our model uses pure convolutional layers as the backbone to extract features. To prevent the network from being too complex, small-size convolution filters (3×3) are used in all layers. Because VGG-16 Net has strong ability of transfer learning and a flexible architecture, it can easily concatenate the backend to extract deeper features for classification. To this end, the first 13 layers of VGG-16 are used as the front-end of our model to extract initial features of the raw images. For LBP feature images, we also use the first 13 layers from VGG-16 to extract deeper features and then reduce the dimensionality to the same as that of the raw images.

As shown in Fig. 5, our dimensionality reduction-based CNN and feature extraction-based CNN have the same architecture. In order to reduce the computational complexity, we add three $1 \times 1 \times 64$ convolutions into the original VGG-16 net to reduce the channels. Unlike the traditional two-stream CNN networks such as Light-CNN [30,44] which extract two different features and then simply fuse them to classify, our feature extraction module is to obtain initial features for future processing in the

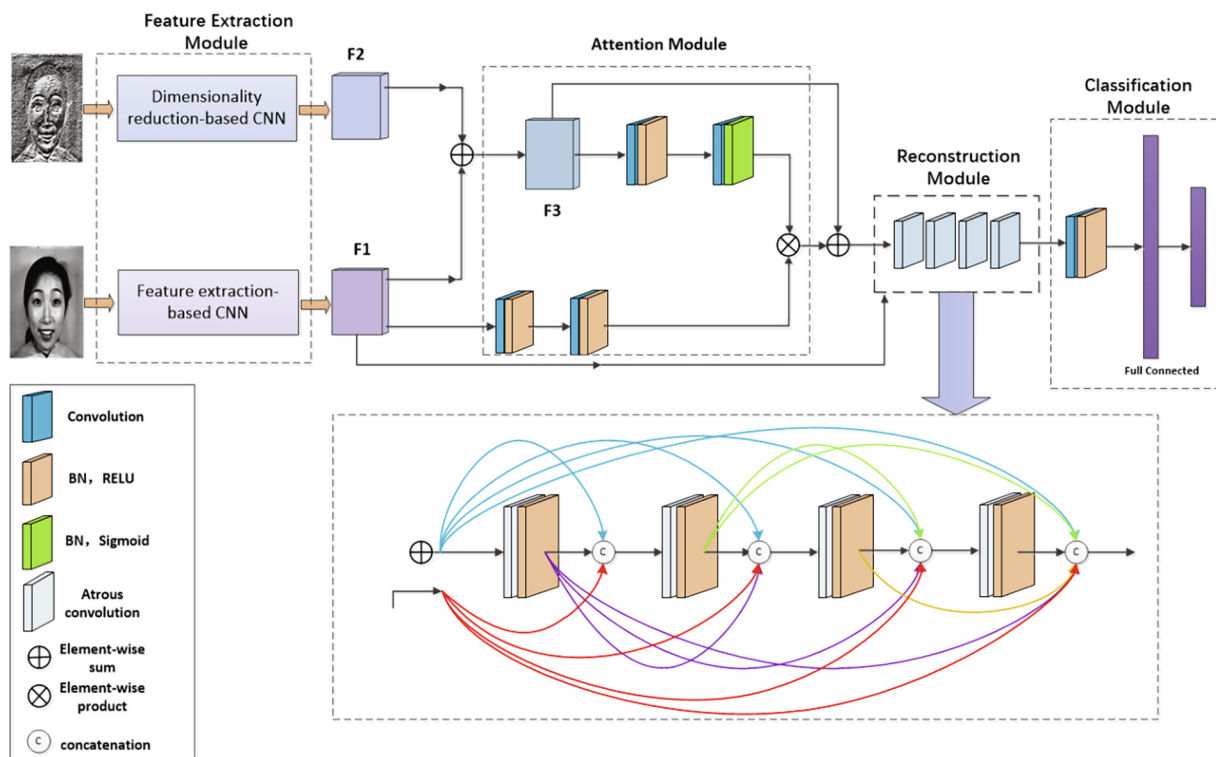


Fig. 4. The architecture of our network.

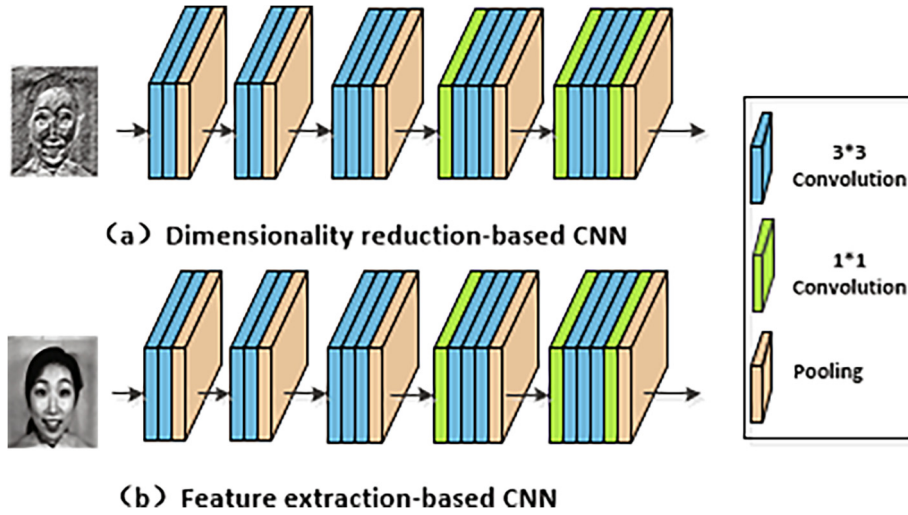


Fig. 5. The architecture of dimensionality reduction-based CNN and feature extraction-based CNN.

following modules. Afterwards, we fuse the features F_1 extracted from the raw images with the features F_2 extracted from the LBP feature images, and then add the fused features F_3 to an attention module.

The attention module works by increasing the weights of useful features and makes the network focus more on these features that are vital for expression recognition. In this way, the network can recognize different expressions more efficiently.

Followed the attention module, we use a dense connection convolution layers as the reconstruction module to adjust the attention map in order to create an enhanced feature map for the classification module. Many works like DenseNet [14] and ResNet [4] have shown that convolution networks with skip connections between different layers can be substantially deeper, more accurate, and more efficient to train. Motivated by DenseNet [14] and ResNet [4], in this work, we use dense atrous convolution for reconstruction. Atrous kernel can be dilated in varied rates by inserting zeros into appropriate positions in the kernel mask. A large dilation rate means a large receptive field, vice versa. Our dense atrous convolution module consists of four 3×3 atrous convolution, the dilation rates from lower layers to higher layers are 2, 3, 4 and 5, respectively. Compared to the traditional convolution operator, atrous convolution is able to achieve a larger receptive field size without increasing the number of kernel parameters. Feature maps from the attention module contain important information for expression recognition. Atrous convolution can increase the receptive field while keeping the resolution of feature maps unchanged. We can further extract features with atrous convolution without information loss and have larger a receptive field. For each layer, the feature-maps of all preceding layers and features F_1 are concatenated as inputs, and its own feature-maps are used as inputs into all subsequent layers. The fused feature maps F_3 are made element-wise sum operation with the output of the attention module. This architecture not only can extract deeper features but also can help alleviate the problem of vanishing-gradient and reuse useful features. At last, fully connected layers with softmax are used for classification. We use the batch normalization after each layer to speed up the convergence of the network and avoid overfitting. The detail of each module in the schema can be seen in Fig. 4.

3.2. Attention mechanism

For a classification task, we extract image features and classify the images into different categories by the differences among these

features. However, only useful features are helpful for classification and different features contribute different significance. The attention mechanism has been proved to be useful in pixel-wise computer vision tasks and is used to measure how much attention to pay to the features in different regions. In this paper, our attention module contains two branches, one is the trunk branch to obtain feature F_p , and the other is the mask branch which integrates LBP features to obtain attention maps F_m . Then, the element-wise product is applied on attention maps F_m and the feature maps F_p to generate refined feature maps F_m as:

$$F_{refine} = F_p F_m \quad (1)$$

Suppose the input of the last layer in the mask branch as f_m , the attention maps F_m are generated as:

$$F_m = \text{Sigmoid}(Wf_m + b) \quad (2)$$

where w and b are the weights and bias of the convolution layer, respectively; \otimes denotes the convolution operation and Sigmoid denotes the sigmoid function. The Sigmoid activation function gives out (0, 1) probability scores to make network discriminate the importance of different features. Fig. 6 provides two example images in FER2013 and their corresponding attention heatmaps generated by our method. As we can see, the heatmaps show the attention areas clearly. For the upper side 'angry' face, the attention of our network is mainly on the eyes; for the lower side 'happy' face, the attention of our network is mainly on the mouth. This indicates that the eyes area contains the most useful features for recognizing the angry expression; while the mouth area is the most suitable for recognizing the happy expression.

3.3. Local binary patterns

Local binary patterns (LBPs) reflect the basic information that is helpful to recognize facial expressions. More specifically, subtle changes can be reflected by the features extracted with LBPs. In addition, LBPs can achieve rotation invariance and grey-scale invariance and thus are suitable for extracting texture features at different scales and can solve the displacement imbalance, rotation angles and illumination conditions in facial images. In Fig. 7, we provide some examples of LBP images extracted from five datasets used in this paper.

In [9], the original LBP defined in a 3×3 window was first introduced. The center pixel of the window is taken as the threshold and then is compared with the gray values of the adjacent 8 pixels. If



Fig. 6. Results of attention heatmap.

the value of the surrounding pixel is greater than the value of the center pixel, the position of the pixel is marked as 1; otherwise, it is 0. In this way, the comparison of the 8 points in the 3×3 neighborhood can generate 8-bit binary numbers, which are usually converted to decimal numbers which is namely LBP code and have a total of 256 kinds. Through the above steps, the LBP value of the pixel in the center of the window is obtained, and the texture information of the region is reflected by the value. It can be defined as follows:

$$LBP(x_c, y_c) = \sum_{p=0}^7 S(i_p - i_c) 2^p \quad (1)$$

where p is the number of pixels, (x_c, y_c) is the coordinate of the center pixel, i_c is the pixel value of the center pixel and i_p is the pixel value of the neighborhood pixel, s is the sign function and can be defined as follows:

$$S(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (2)$$

The biggest drawback of the original LBP is that only a small region is covered by it within a fixed radius, and this is not suitable when the sizes and frequency of textures are different. Circle LBP was proposed to make it suitable for different size and frequency textures features and meet the needs of gray scale and rotation invariance. It is not only 3×3 neighborhood but can be any neighborhood, and the circular neighborhood is also replaced with a square neighborhood. Circle LBP allows arbitrary multiple pixel points in the circular neighborhood with radius R . The LBP operator with P sampling points in the circular region with radius R is obtained. It can be defined as follows:

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} S(i_p - i_c) 2^p \quad (3)$$

where p is the number of sampling points, R is the radius of the circle neighborhood, (x_c, y_c) is the coordinate of center pixel, i_c is the pixel value of the center pixel and i_p is the pixel value of P sampling points, and s is the sign function which is same as in Eq. (2).

Further extension of LBP such as uniform patterns [10] was used to solve the problems of too many binary eigenvalue encoding modes and improve statistical performance. When the bitwise pattern is circularity, uniform LBP includes at most two bitwise transitions from 0 to 1 or 1 to 0. It can be defined as follows:

$$LBP_{P,R}^{rius2} = \begin{cases} \sum_{p=0}^{P-1} s(i_p - i_c) & \text{if } U(LBP_{P,R}) \leq 2 \\ P + 1 & \text{otherwise} \end{cases} \quad (4)$$

where U is the uniformity measure and the 'rius2' represents rotation-invariant uniform pattern.

3.4. Data augmentation

In order to ensure that the network can achieve a good generalization ability, enough training data are required. However, most publicly-available datasets such as JAFFE do not have enough number of images for training. The amount of data is small, which can also lead to the overfitting problem. Therefore, data augmentation is important for facial expression recognition.

Label-preserving transformation is one of the most common methods to enlarge the number of images in datasets. In this paper, four methods are used to enlarge the image number of the original data and the example images after augmentation are shown in Fig. 8. We take random rotation, flipping, shifting and scaling on the images. The rotation range is $0-20^\circ$ and the shifting ranges of width and height are set to $0-15\%$. Both the shear range and the zoom range are $0-0.15$. Before data augmentation, the number of images in FER2013, CK+, JAFFE, Oulu-CASIA and NCUFE are 35,795, 981, 213, 1440 and 735, respectively. After data augmentation, the number of images in FER2013, CK+, Oulu-CASIA and NCUFE are increased to 45,017, 12,000, 11,843, 13,800 and 13,827, respectively.

4. Experimental results

In this work, we design a novel deep Convolutional Neural Network with an attention model to automatically recognizing facial expressions. Except for the famous facial expression datasets such as CK+ [11], JAFFE [13], Oulu-CASIA [25] and FER2013 [39], we also evaluate our proposed method on our self-collected dataset NCUFE. Because CK+, JAFFE, Oulu-CASIA and NCUFE do not provide specified training and testing sets, we employ 5-fold cross-validation protocol in these four datasets. The proposed facial expression recognition framework is implemented with Keras. The framework is trained with an initialized learning rate 0.000001, and the batch size is 40. The size of input data is 256×256 because using large images as input to the network can make the network as deep as possible and help extract more useful features. We add batch normalization after each convolutional layer. Our network is trained on one TITAN RTX GPU. In this section, we first conduct an ablation study on FER2013 [39] to analyze the configuration of our framework (as shown in Table 1). Then, we evaluate our proposed method in all these five datasets and compare the result to some state-of-the-art methods.

4.1. Ablations on FER2013

We perform an ablation study to analyze the effects of three kinds of LBPs (original LBP, circle LBP and uniform LBP) in our method on the FER2013 dataset [39], which is a large-scale and unconstrained dataset collected from Internet. It is challenging to classify these images because of varied perspectives of face, wrong labels and some other noises. The recognition results are shown in

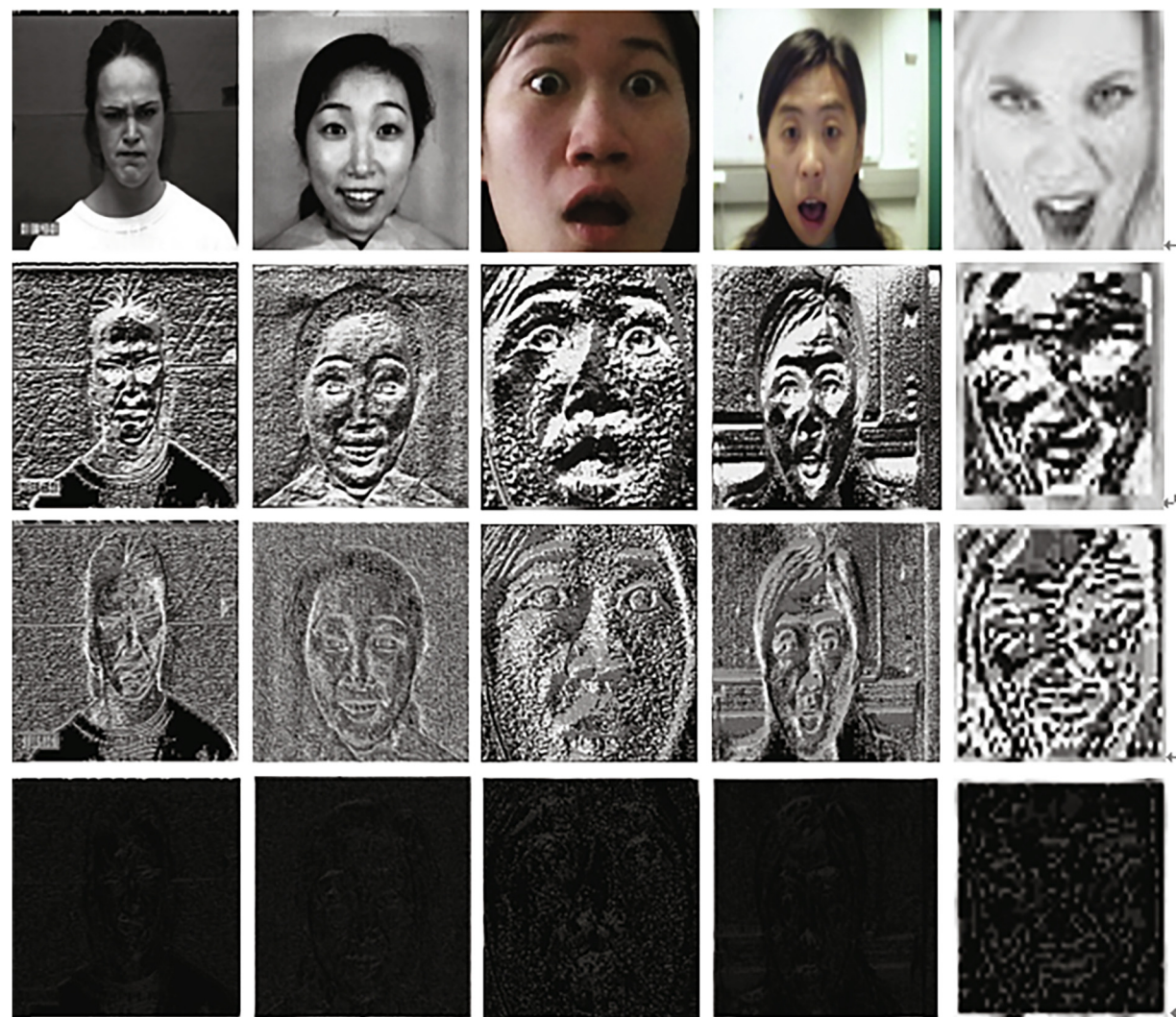


Fig. 7. Examples of LBP images extracted from CK+, JAFFE, NCUFE, Oulu-CASIA and FER2013. From top to bottom is the original image, circle LBP image, original LBP image, and uniform LBP image.

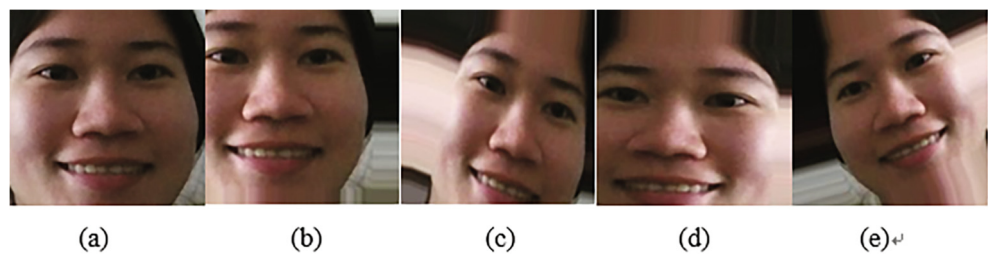


Fig. 8. Examples for data augmentation in the NCUFE dataset, (a) is the image from the original dataset, (b), (c), (d) and (e) are the images after augmentation.

Table 1
Comparison of different LBPs on FER2013 dataset.

LBP	Recognition Rates (%)
Original LBP	72.56
Uniform LBP	70.32
Circle LBP	75.82
Without LBP	67.73

Table 1, where circle LBP achieves the best accuracy. Therefore, we use circle LBP in our model for the following experiments.

Our model includes four parts, which are the feature extraction module, the attention module, the reconstruction module and the classification module. In order to figure out how each module affects the performance of our network we perform an ablation study on the networks. Because the classification module is necessary to classify in our networks, we retain the classification module and remove the feature extraction module, the attention module, and the reconstruction module separately and then conduct differ-

Table 2

Comparison of different architectures on the FER2013 dataset.

Architecture	Recognition Rates (%)
Without the feature extraction module	57.86
Without the attention module	73.96
Without the reconstruction module	74.25
Complete network	75.82

ent experiments. The results are given in Table 2. As we can see, when one of these modules is removed, the recognition rate has a certain degree of decline compared with the complete network which achieves the recognition rate of 75.82%. Especially, when there is no feature extraction module, the recognition rate drops to 57.86%. In the feature extraction module, initial features are extracted from raw images and then sent to later modules for future processing. Generally, the obtained initial features are too coarse, but we need more refined features to send to the later attention module in order to make it work directly. Therefore, the feature extraction module is necessary in the network for extracting refined features. The attention module can make our network focus more on useful features and improve the recognition rate. When the attention module is removed, the recognition rate reduces to 73.96%, which proves the effectiveness of the attention module. The reconstruction module can improve the recognition rate by adjusting the attention map to create an enhanced feature map. When there is no reconstruction module in the network, the recognition rate drops to 74.25%. Based on the experimental results, we conclude that each module has a certain degree of improvement on the final results.

When designing the reconstruction module, we try several atrous kernels of different dilation rates. We perform experiments on five kinds of configurations and use A, B, C, D and E to refer to them separately. There are four atrous convolutions in the reconstruction module. In A, B, C and D, we use invariable dilation rates, the dilation rates are 1, 2, 3, and 4 separately. We use four atrous convolutions with four different dilation rates in E, which are 2, 3, 4 and 5 separately. The results are provided in Table 3. We can see that the reconstruction module E achieves the highest recognition rates. Therefore, we use variable dilation rates which are 2, 3, 4, 5 in the proposed method in this paper.

4.2. Evaluation and comparison

4.2.1. FER2013 dataset

FER2013 [39] is a dataset collected in real world, which includes 28,709 training images and 3589 test images. Unlike CK+, Oulu-CASIA and JAFFE, FER2013 contains pictures of different postures, unbalanced illumination, and occlusion. We evaluate and compare our method with other five recent works [18,30,40,42,44]. The recognition results are given in Table 4, which show that our method is superior to all of the five advanced algorithms. The recognition rates are reduced in this dataset when there is no

Table 3

Comparison of different dilation rates of the reconstruction module on the FER2013 dataset.

Dilation Rate	Recognition Rates (%)
A (1)	74.83
B (2)	73.91
C (3)	75.52
D (4)	74.31
E (2, 3, 4, 5)	75.82

Table 4

Comparison of different methods on the FER2013 dataset.

Method	Recognition Rates (%)
Wang [18]	71.1
Pramerdorfer et al. [42]	75.2
Kim et al. [41]	73.73
Guo et al. [40]	71.33
Shao et al. [30]	71.14
Ours	75.82
Ours (without LBP)	67.73
Ours (without attention module)	73.96

Table 5

Comparison of different methods on the CK+ dataset.

Method	Recognition Rates (%)
VGG Net + LSTM [33]	97.2
Yang et al. [35]	97.3
Zhang et al. [37]	98.9
Turan [20]	96.10
Shao et al. [30]	95.29
Ours	98.68
Ours (Without LBP)	97.53
Ours (Without the attention module)	97.10

LBP or attention module. The result is reduced to 67.73% without LBP and 73.96 without attention module.

4.2.2. CK+ dataset

The CK+ dataset [11] includes 593 image sequences collected from 123 subjects, and 327 of these sequences are labeled with six facial expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise). Each image sequence gradually reaches a peak expression from neutral face. 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. Our method can achieve competitive result by only using the facial expression data. When removing the LBP or the attention module, the recognition results are reduced from 98.68% to 97.53% and 97.10%, respectively. This shows that LBP and the attention module can improve the recognition accuracy.

4.2.3. Oulu-CASIA dataset

The Oulu-CASIA dataset [25] contains 2880 image sequences which are collected from 80 subjects and each sequence is labeled with six facial expression labels (i.e., anger, disgust, fear, happiness, sadness, and surprise). Each image sequence gradually reaches a peak expression from neutral face. The last three frames from each expression sequence are selected to be new data used in our experiment. We compare our approach with previous state-of-the-art

Table 6

Comparison of different methods on the Oulu-CASIA dataset.

Method	Recognition Rates (%)
Zhong et al. [28]	93.06
Yang et al. [26]	88.00
Zhang et al. [23]	86.95
Kuo et al. [19]	88.75
Ours	94.63
Ours (Without LBP)	93.52
Ours (without the attention module)	91.17

Table 7

Comparison of different methods on the JAFFE dataset.

Method	Recognition Rates (%)
Hamester et al. [31]	95.8
Liu et al. [27]	91.8
Turan [20]	91.8
Wang [18]	95.7
Ours	98.52
Ours (Without LBP)	96.53
Ours (Without the attention module)	95.12

Table 8

Comparison of different methods on NCUE dataset.

Method	Recognition Rates (%)
Li et al. [34]	91.58
Arriaga et al. [29]	81.3
Light-CNN. [30]	82.5
Ours	94.33
Ours (Without LBP)	93.91
Ours (Without attention module)	92.07
Ours (raw + depth)	90.35

methods and give the results in Table 6. As we can see, our method is superior to all the other methods and achieves the state-of-the-art performance which is 94.63%. This also demonstrates that LBP and the attention module can improve the accuracy in different degrees. When there is no LBP or attention module, the results are reduced to 93.52% and 91.17 respectively.

4.2.4. JAFFE dataset

The JAFFE dataset was collected from 10 Japanese females in a laboratory condition and includes 213 images of posed expressions. For each subject, there are three or four images belonging to one of the six expressions which are anger, disgust, fear, happiness, sadness, and surprise, and one image belongs to neutral face. We compare our method with five existing methods [18,20,27,31] and the comparison results are provided in Table 7, which shows that the result of our method achieves 98.52% and outperforms the others on the JAFFE dataset. When LBP or attention is not used, the recognition rates are reduced to 96.53% and 95.12%, respectively.

4.2.5. NCUE dataset

We collected a new facial expression dataset called Nanchang University Facial Expression (NCUE) from 35 graduate students (6 females and 29 males) by the Microsoft Kinect sensor. For each student, we captured both RGB images and depth images, and the sizes of these two kinds of images are 1280*1024 and 512*424, respectively. The NCUE dataset was annotated with seven expressions (i.e., anger, disgust, fear, happiness, sadness, surprise, and neutral). For each image type, there are 245 image sequences, each of which contains 110 images, resulting in 26,950 images in total. Because none of the other four datasets has depth images, we only use RGB images as our datasets in this work. To better distinguish different kinds of expressions, for every expression sequence, we select three images that are close to the peak expression as our new dataset. We evaluate our method on this dataset and achieve 94.33% accuracy. When we remove LBP or the attention module, the recognition rates are reduced to 93.91% and 92.07%, respectively. In addition, when we replace LBP with depth images, the recognition result is lower than that using only raw images or LBP images. It indicates that depth images are not helpful in our networks. The reason is that in our network we fuse these two kinds of features by element-wise sum, as a kind of auxiliary fea-

ture, depth features have a negative effect on the raw image features when fusing these two kinds of features. In our future work, we will apply depth images to other methods, for example, we can concatenate the feature vectors of these two kinds of features. We also compare our method with some other works. From Table 8, we can find that our method outperforms the others on our dataset. The recognition results of [33,30,34] on NCUE are 91.58%, 81.3% and 82.5%, respectively.

5. Conclusions and future work

This paper presents a novel convolutional neural network with attention mechanism for facial expression recognition. The method fuses LBP features and convolution features, and then is combined with attention mechanism to improve the performance of the network. In order to prevent overfitting and ensure the generalization ability of the network, we apply data augmentation in the datasets we used in the experiments. In addition, we collected a new dataset called Nanchang University Facial Expression (NCUE) and the NCUE dataset was annotated with seven expressions, which are anger, disgust, fear, happy, sad, surprise, and neutral. For each image type, there are 245 image sequences, each of which contains 110 images, resulting in 26,950 images in total. The presented method is evaluated on NCUE and four famous facial expression datasets, i.e., Oulu-CASIA, JAFFE, CK+ and FER2013. The experimental results show that our method is superior to many existing methods on these datasets.

However, our method is only suitable for 2D images. In the future, we will improve our architecture to make it suitable for video data, 3D face datasets and our depth images data, and explore better machine learning methods to enhance the network.

CRedit authorship contribution statement

Jing Li: Conceptualization, Methodology, Writing - original draft. **Kan Jin:** . **Dalin Zhou:** . **Naoyuki Kubota:** Writing - review & editing. **Zhaojie Ju:** Conceptualization, Methodology, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant 61963027, 61703198, and 51575412, Natural Science Foundation for Distinguished Young Scholars of Jiangxi Province under Grant 2018ACB21014.

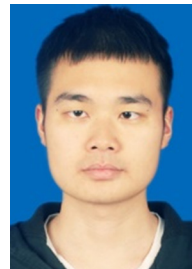
References

- [1] Y.-H. Lai, S.-H. Lai, Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition, in: Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on. IEEE, 2018, pp. 263–270.
- [2] A. Mollahosseini, D. Chan, M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, (2016) 1–10.
- [3] C. Szegedy, W. Liu, Y. Jia, et al., Going deeper with convolutions, IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015, pp. 1–9.
- [4] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, (2015) 770–778.
- [5] Sun, H. Jin, Z. Zhao, An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks, *Neurocomputing* 267 (2017) 385–395.

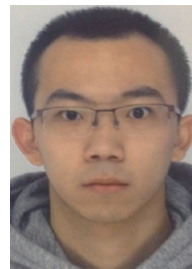
- [6] C. Szegedy, S. Ioffe, V. Vanhoucke, et al., Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, 2016
- [7] B. Hasani, M.H. Mahoor, Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields, (2017) 790–795
- [8] Z. Yu, G. Liu, Q. Liu, et al., Spatio-temporal convolutional features with nested LSTM for facial expression recognition[J], Neurocomputing 317 (2018) 50–57.
- [9] T. Ojala, I. Harwood, A comparative study of texture measures with classification based on feature distributions, Pattern Recogn. 29 (1) (1996) 51–59.
- [10] T. Ojala, M. Pietikinen, T. Menp, Multiresolution grayscale and rotation invariant texture classification with local binary patterns, IEEE PAMI, vol. 24, no. 7, 2002
- [11] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE, 2010, pp. 94–101.
- [12] W. Sun, H. Zhao, Z. Jin, A visual attention based ROI detection method for facial expression recognition, Neurocomputing 296 (2018) 12–22.
- [13] M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with Gabor wavelets, in: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 200–205
- [14] G. Huang, Z. Liu, K. Q. Weinberger, L. van der Maaten, Densely connected convolutional networks. arXiv preprint arXiv:1608.06993, 2016
- [15] T. Zhao, X. Wu, Pyramid feature attention network for saliency detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, 3085–3094
- [16] R.R. Varior, B. Shuai, J. Tighe, et al., Scale-Aware Attention Network for Crowd Counting. arXiv preprint arXiv:1901.06026, 2019.
- [17] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2011) 6815
- [18] W. Wang, Q. Sun, T. Chen, et al., A Fine-Grained Facial Expression Database for End-to-End Multi-Pose Facial Expression Recognition, arXiv preprint arXiv:1907.10838, 2019.
- [19] C.M. Kuo, S.H. Lai, M. Sarkis, A compact deep learning model for robust facial expression recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2121–2129
- [20] C. Turan, K.M. Lam, X. He, Soft Locality Preserving Map (SLPM) for Facial Expression Recognition. arXiv preprint arXiv:1801.03754, 2018.
- [21] Z. Zhang, M. Lyons, M. Schuster, S. Akamatsu, Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron, in Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. Proceedings, Nara 1998, pp. 454–9.
- [22] C. Cortes, V.N. Vapnik, Support vector networks, Mach. Learn. 20 (1995) 273–297.
- [23] K. Zhang, Y. Huang, Y. Du, et al., Facial expression recognition based on deep evolutionary spatial-temporal networks, IEEE Trans. Image Process. 26 (9) (2017) 4193–4203.
- [24] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theor. 13 (1) (1967) pp. 21–7
- [25] G. Zhao, X. Huang, M. Taini, et al., Facial expression recognition from near-infrared videos, Image Vision Comput. 29 (9) (2011) 607–619.
- [26] Huiyuan Yang, Umur Ciftci, Lijun Yin, Facial expression recognition by de-expression residue learning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018
- [27] P. Liu, S. Han, Z. Meng Y. Tong, Facial expression recognition via a boosted deep belief network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1805–1812
- [28] Zhong, Lei, et al., A graph-structured representation with BRNN for static-based facial expression recognition, 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019.
- [29] Arriaga, Octavio, Matias Valdenegro-Toro, Paul Plöger, Real-time convolutional neural networks for emotion and gender classification. arXiv preprint arXiv:1710.07557 (2017).
- [30] J. Shao, Y. Qian, Three convolutional neural network models for facial expression recognition in the wild, Neurocomputing 355 (2019) 82–92.
- [31] D. Hamester, P. Barros, S. Wermter, Face expression recognition with a 2-channel convolutional neural network, in: Neural Networks (IJCNN), 2015 International Joint Conference on, IEEE, 2015, pp. 1–8
- [32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Sci. (2014).
- [33] P. Rodriguez, G. Cucurull, J. Gonzalez, et al., Deep pain: exploiting long short-term memory networks for facial expression classification, IEEE Trans. Cybern. PP (99) (2017) 1–11
- [34] J. Li, Y. Mi, J. Yu, et al., A novel convolutional neural network for facial expression recognition, International Conference on Cognitive Systems and Signal Processing, Springer, Singapore, 2018, pp. 310–320
- [35] H. Yang, U. Ciftci, L. Yin, Facial expression recognition by de-expression residue learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2168–2177
- [36] Y. Lecun, L. Bottou, Y. Bengio, et al., Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324
- [37] Z. Zhang, P. Luo, C.L. Chen, X. Tang, From facial expression recognition to interpersonal relation prediction, Int. J. Comput. Vis. 126 (5) (2018) 1–20.
- [38] P.D.M. Fernandez, F.A.G. Peña, T.I. Ren, et al., FERAtt: facial expression recognition with attention net. arXiv preprint arXiv:1902.03284, 2019.
- [39] <https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data>
- [40] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, D. Tao, Deep neural networks with relativity learning for facial expression recognition, in: Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on, IEEE, 2016, pp. 1–6.
- [41] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, S.-Y. Lee, Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 48–57
- [42] C. Pramerdorfer, M. Kampel, Facial expression recognition using convolutional neural networks: State of the art, arXiv preprint arXiv:1612.02903, 2016.
- [43] Dhvani Mehta, Mohammad Faridul Haque Siddiqui, Ahmad Y. Javaid, Recognition of emotion intensities using machine learning algorithms: a comparative study, Sensors 19 (8) (2019) 1897
- [44] T. Connie, M. Al-Shabi, W.P. Cheah, et al., Facial expression recognition using a hybrid CNN-SIFT aggregator, International Workshop on Multi-disciplinary Trends in Artificial Intelligence, Springer, Cham, 2017, pp. 139–149



Jing Li received the B.E. degree in Electronic Information Engineering from Nanchang University, China, in 2005, and obtained her PhD degree in Electronic and Electrical Engineering from the University of Sheffield, UK, in 2011. From 2011 to 2012, she was a Research Associate with the Department of Computer Science at the University of Sheffield. Currently, she is an Associate Professor with the School of Information Engineering at Nanchang University. Her research interests include visual tracking, behavior recognition, and crowd scene analysis. She has authored or coauthored in various journals, such as IEEE Transactions on Industrial Informatics, IEEE Transactions on Image Processing, Information Sciences (Elsevier), etc.



Kan Jin received the B.E. degree in Mechatronic Engineering from Hubei Polytechnic University, China, in 2017. He is currently pursuing the M.S. degree in the School of Information Engineering at Nanchang University, China. His research interests include facial expression recognition and crowd counting.



Dalin Zhou received the B.S. degree in automation from the University of Science and Technology of China, Hefei, China, in 2012, the Ph.D. degree in computing from the University of Portsmouth, UK, in 2018 and is currently a Lecturer in Computer Science with the School of Computing, the University of Portsmouth, UK. His current research interests include signal processing, machine learning, computational intelligence, biomedical robotics and multimodal sensor fusion. His current research contributes to the monitoring and rehabilitation of limb motor function improving the daily life activity and working capability for both the disadvantaged group of limb-impaired patients and the aging community.



Naoyuki Kubota, Faculty of Systems Design, Tokyo Metropolitan University, Tokyo, Japan Naoyuki Kubota received the B.Sc. degree from Osaka Kyoiku University, Kashiwara, Japan, in 1992, the M.Eng. degree from Hokkaido University, Hokkaido, Japan, in 1994, and the D.E. degree from Nagoya University, Nagoya, Japan, in 1997. He joined the Osaka Institute of Technology, Osaka, Japan, in 1997. He joined the Department of Human and Artificial Intelligence Systems, University of Fukui, Fukui, Japan, as an Associate Professor in 2000. He joined the Department of Mechanical Engineering, Tokyo Metropolitan University, Tokyo, Japan, in 2004.

He was an Associate Professor from 2005 to 2012, and has been a Professor since 2012 at the Department of System Design, Tokyo Metropolitan University.



Zhaojie Ju received the B.S. in automatic control and the M.S. in intelligent robotics both from Huazhong University of Science and Technology, China, and the Ph. D. degree in intelligent robotics at the University of Portsmouth, UK. He held a research appointment at the University College London, London, U.K., before he started his independent academic position at the University of Portsmouth, U.K., in 2012. His research interests include machine intelligence, pattern recognition, and their applications on human motion analysis, multi-fingered robotic hand control, human-robot interaction and collaboration, and robot skill learning.

He has authored or co-authored over 180 publications in journals, book chapters, and conference proceedings and received four best paper awards and one Best AE Award in ICRA2018.