IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 30, 2021, accepted May 5, 2021, date of publication May 7, 2021, date of current version May 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3078258

Facial Expression Recognition Using Pose-Guided Face Alignment and Discriminative Features Based on Deep Learning

JUN LIU[1], YANJUN FENG[2], AND HONGXIA WANG[2]
[1]School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110159, China
[2]School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China

Corresponding author: Yanjun Feng (braverfyj@126.com)

This work was supported by the Liaoning Provincial Natural Fund Guidance Program Projects under Grant 2019-ZD-0251 and Grant 2019-ZD-0249.

ABSTRACT Face expression recognition is a key technology of robot vision, which can help the robotic understand human emotions. However, interference from the real-world, such as light changes, face occlusion, and pose variation, reduces the recognition rate of the model. To solve above problems, in this paper, a novel deep model is proposed to improve the classification accuracy of facial expressions. The proposed model has the following merits: 1) A pose-guided face alignment method is proposed to reduce the intra-class difference, which can overcome the impact of environmental noise; 2) A hybrid feature representation method is proposed to obtain high-level discriminative facial features that achieves better results in classification networks; 3) A lightweight fusion backbone is designed, which combines the VGG-16 and the ResNet to achieve low-data and low-calculation training. Finally, to evaluate the proposed model, we conduct a series of experiments on four benchmark datasets, including the CK+, the JAFFE, the Oulu-CASIA, and the AR. The results show that the proposed model achieves state-of-the-art recognition rates, that is, 98.9%, 96.8%, 94.5%, and 98.7%, respectively. Comparing with the traditional methods and other advanced deep learning methods, the proposed model can comparable performance in a variety of tasks.

INDEX TERMS Face expression recognition, deep learning, hybrid features, VGG, ResNet.

I. INTRODUCTION

Facial expression recognition is a key technology of robotic vision and aims to accurately classify various expressions [1]. This technology is wildly used in many actual applications, such as abnormal behavior detection, game entertainment, and film and television creation [2], [3]. In most works, researchers focus on six kinds of facial expressions, including ''Anger'', ''Disgust'', ''Fear'', ''Happiness'', ''Sadness'', and ''Surprise''.

The typical face expression recognition system (as shown in Figure 1) consists of four aspects, that is, face detection, face alignment, facial feature extraction, and classification [4], [5]. In this process, both face alignment and feature extraction are the key steps for classification performance. Especially, the recognition of the expressions collected in a complex environment achieves a lower accuracy because of pose variation and perspective transformation. Hand-crafted features (used in traditional methods) cannot achieve satisfactory results in most tasks with severe interference [6], [7]. Motivated by the success of Convolutional Neural Networks (CNNs), many state-of-the-art deep learning methods are proposed for face expression recognition and the recognition effect is better than traditional methods. CNN-based models can automatically model and learn the abstract features of the facial object because of the neural network structure, so these models can achieve better results in the real-world applications [8], [9]. More typical CNN-based face expression recognition models are described in details in the related work section, and some novel face recognition related technologies are shown as follows. For example, a deep learning strategy is proposed for partially occluded face recognition, such as wearing a mask [10]. Note that due to the COVID-19 pandemic, there are more scenes of people wearing masks, so this is a recent hot research area [11]. Similar works include the PAD model [12], the MFDD/RMFRD datasets [13], and

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Da Lin.

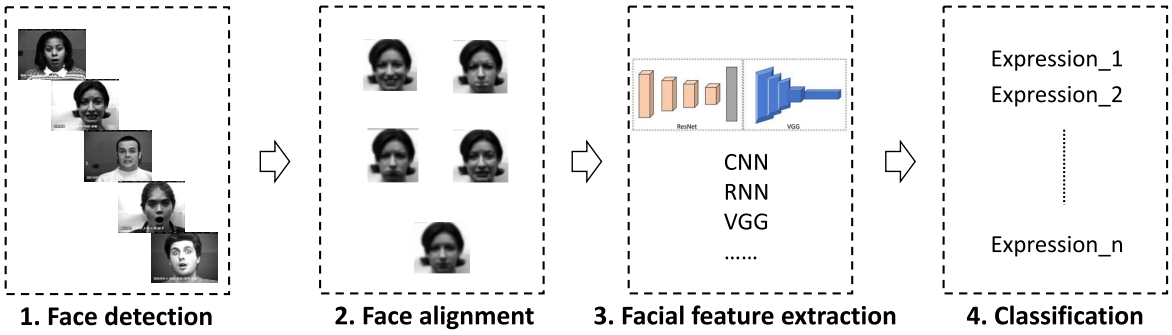


FIGURE 1. Face expression recognition system.

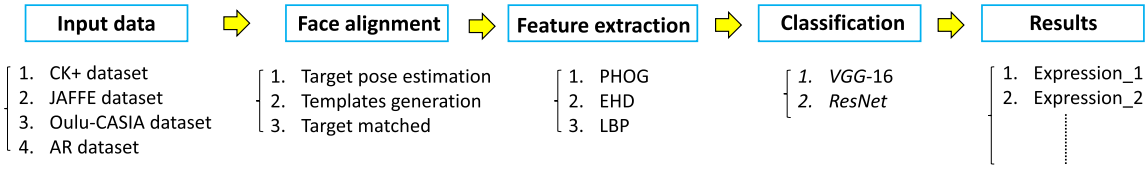


FIGURE 2. Structure of the proposed model.

a hybrid model [14]. A CNN-based low power processor is proposed for mobile-based face recognition [15]. The novel co-mining strategy is proposed for the noisy labels training, which can significantly improve the training and modeling capabilities of deep models [16].

However, the pervious methods have several limitations. First, the effect of the face recognition model in an unconstrained environment is not ideal. Generally, in this environment, the facial expression of the same person looks different due to environmental factors, such as occlusion, lighting changes, and posture changes. Second, misclassification of similar expressions caused by in-class similarity and data noise. Third, the training process of the deep model relies on a large number of samples and computational cost. Additionally, some neutral expressions or differentiated expressions (every participant has his own emotional expression) bring challenges, such as "contempt" and "disgust". Even manual classification of such expressions can also cause misclassification. To solve above problems, our ideas are described as follows:

- Design a face alignment method to reduce the intra-class difference and correct the background noise, which can overcome the influence of light changes and occlusion.
- Extract highly discriminative facial features to increase the discrimination of similar expressions, which can reduce misclassification.
- Design a lightweight backbone to make training easier and reduce the cost of calculations.

In the paper, we propose a new facial expression recognition model based on deep learning to reduce the intra-class difference and improve recognition accuracy, as shown in Figure 2. The main idea is to propose a novel face alignment method to overcome challenges from the real world, such as lighting variation, background noise, and target occlusion. Next, considering that intra-class similarity can cause misclassification, we propose an effective hybrid feature representation method to obtain more discriminative facial features. The proposed model has the following notable properties: 1) The pose-guided face alignment method can adaptively finish the template generation and matching, which is better than the methods of hand-crafted designing templates; 2) The proposed feature extraction method can obtain high-level facial representations and retain more distinguishing features; 3) The proposed deep model cannot require a lot of training samples to get the optimal number of templates; 4) The proposed model is easier to converge and reduces training time.

The main contributions of the paper are summarized as follows:

- We propose a new pose-guided face alignment method to adaptively conduct target template matching process, in which three key steps are included. This method can reconstruct the 3D face structure and reduce the intra-class difference to eliminate the influence of noise.
- We propose an effective feature representation method to generate high-level semantic features, which can dig deep into the intra-class similarity to improve the recognition accuracy of facial expression classification.
- We re-design a deep model by introducing the Resnet to replace the part of the VGG, which can improve the

modeling effect of the proposed method on high-level facial features and reduce the computational cost.
- We conduct a series of experiments on four benchmark datasets, including ablation studies, parameter selecting, and comparison analysis. The results show that the proposed model achieves state-of-the-art performance and outperform the traditional methods and deep learning models.

The remainder of the paper is organized as follows. In Section 2, the related work is described in details, including the face alignment method and the feature extraction method. Section 3, the proposed model is introduced in details. In Section 4, a series of experiments are conducted and the results are discussed. Section 5 concludes the work and gives further study direction.

II. RELATED WORK

A. FACE ALIGNMENT

Face alignment, normally, consists of two main research directions, that is, template-based methods and STN-based methods. In the previous works, face alignment related models are proposed based on the active appearance technology, such as the ASM and the AMM [17], [18]. Motived by the success of the cascaded pose regression, many state-of-the-art models are proposed to learn vector regressors to infer landmarks from input face images. Typical models are introduced as follows: 1) A CNN-based model is proposed to effective model the whole facial expression without the requisition of accurate initialization and face detection, in which the convolutional aggregation of local evidence is the main contribution [19]; 2) A novel facial recognition system is proposed by introducing the robust cascaded pose regression that can achieve better performance in a complex environment [20]; 3) A unified model is designed for face detection, pose estimation and landmark estimation for chaotic images in the real-world, which is better than the Google Picasa [21]. These models have poor generalization ability for invisible images and low training efficiency, and cannot achieve satisfactory performance in challenging real environments.

With the development of deep learning, many deep models are proposed for face alignment and achieve state-of-the-art performance. The release of a large number of annotated face tagging data sets has greatly promoted the development of face alignment technology. Most face alignment methods based on deep learning complete face target alignment during training and testing. Typical deep models are described as follows: 1) An end-to-end deep model is proposed, and both prior knowledge-based facial landmarks and artificial-based geometric transformations are not required in the modeling process [22]; 2) To make reasonable use of the intra-class similarity, an angular Softmax-based model is proposed that has strong robustness and generalization, and it is introduced by many models [23]; 3) The ArcFace is proposed for large-scale face recognition based on Deep-CNN, and it is evaluated on ten benchmark datasets and achieves state-of-the-art results [24]; 4) The LMCL method is proposed and applied in the area of the center of face recognition, which is an extension of the center loss, large margin Softmax, and angular Softmax [25]. Due to easy implementation and strong performance, in this paper, the template-based method is utilized for the design of the proposed face alignment module. Moreover, pose-invariant is also a challenge for facial analysis, and the details can be seen in work [4]. To solve this problem, recently, many state-of-the-art works are proposed, such as [26]–[29], and [30].

B. FEATURE EXTRACTION

Feature extraction is a key technology in the study of facial expression recognition, the focus is on extracting key facial expression features from facial images and inputting them into the classifier to recognize different facial expressions. In the previous works, many studies focus on the hand-crafted features, including appearance, geometry, and movement features. Many typical works are proposed to represent the whole facial expression by extracting global and local information, such as the HOG, the pixel intensity, and LBP [31]–[33]. However, these features are more representative of global features, ignoring local features where expression changes are highly correlated, such as eyes, nose, and mouth. Next, many state-of-the-art methods are proposed by introducing the hybrid features. For example, a MTSL framework is designed to distinguish similar expressions, a multi-modal learning scheme is proposed to extract the texture and landmark features, a LCRF model is proposed to automatically recognize complex facial expressions [34]–[36]. However, hand-crafted-based facial recognition methods cannot achieve ideal results in complex and changing actual scenes. Recently, transfer learning technology is utilized to help model training. This enables the facial recognition system to obtain better performance with low training costs, such as in [37].

In recent years, the deep learning technique is widely used in the area of pattern recognition, computer vision, and image processing. Hence, many deep learning methods are proposed for facial expression recognition and achieve state-of-the-art results. Many deep networks are very effective for key feature extraction, such as the EmotiW, the DBNs, the AUDNs, the E3D-LSTM, 3D-MM, and the D-ConvLSTM [38]–[43]. Deep learning-based models focus on the following research aspects: 1) Improve the deep model by re-designing the network structure and random weight initialization; 2) Mining more discriminative facial features; 3) Reduce the dependence of the deep model on a large number of data samples; 4) Publish datasets that have high-quality labels and are collected under actual world.

III. PROPOSED METHOD

A. FACE ALIGNMENT

For face alignment, in this sub-section, three key steps are included, that is, target pose estimation, templates generation, and target matched. Motived by the study of face features

FIGURE 3. Schematic process of template generation.

positing, like in [44], [45], and [46], the novel face alignment method is proposed.

Because of the three-dimensional structural characteristics of the face target in space, first, the compensate module is designed for yaw variations. We mainly focus on three parts of the face, that is, the center of eyes, the corners of mouth and the tip of nose. A cascaded CNN is utilized to map 2D features into 3D space to restore the key representation of the face in the real scene. The above key 2D points are denoted as $L_{2D} = (x_k, y_k)$, the corresponding 3D points are denoted as $L_{3D} = (x_k, y_k, z_k)$, and the transformation process between the 2D and the 3D is shown as in (1). Where ω is the weight, $tr(A)$ is the mapping matrix, and $R_{3\times3}$ is the face pose matrix

$$L_{3D} = \omega \cdot tr(A) \cdot R_{3\times3} \cdot L_{2D} \tag{1}$$

Next, for the template generation process, the key challenge is to mitigate the impact of the intra-class similarity, which is caused by the inappropriate number of templates. Hence, in this work, an effective template generation module is proposed to obtain the optimal number of templates that can eliminate the intra-class similarity and retain enough useful information. Note that too many templates can lead to misclassification due to the intra-class similarity, while too few templates can lead to the inability to obtain discriminative features. Typically, the single-template-based methods are utilized in the previous works. Considering the improvement of the adaptive ability of the proposed model, in this work, the pose-guided-based method is designed to output the final templates, in which a large number of samples can be used to cover as many face samples as possible. We also eliminate redundant samples based on angular symmetry to improve the training speed of the proposed model.

Our main idea is to adaptive obtain the optimal number of templates (N) to balance clustering and data partitioning. Divide a large number of samples into several categories based on the intra-class similarity, and then select the most representative templates in each category. The schematic diagram is shown in Figure 3 and the process is shown in (2). Where D is the set of all the clustering centers, P_i is the pose obtained from the pose estimation, $K - means$ is the common clustering function. The best value of N can be obtained by the training process. Note that other traditional methods are utilized to obtain the final clustering results, but the $K - means$ method can achieve the best results with a large number of samples.

$$D = \sum_{i=1}^{N} K - means(P_i) \tag{2}$$

After the above process, we assign the best alignment template for each sample. First, the facial key points proposed above are used as the main matching feature. Then, the pose estimation results are used to locate the face target pose. Finally, the pose-guided method is used to select the optimal templates. The best template selection process is shown in (3). Where T denotes the final selected template, x denotes the face target, and t_i denotes the recommended template. After determining the best template, the input face sample can be aligned to the template by the transformation matrix ($tr(T)$), as shown in (4). Where θ_t is the pose rotation matrix, it is determined by the target rotation angle and the translation vector.

$$T = \sum_{i}^{N} \arg\min (x - t_i)^2 \tag{3}$$

$$tr\,(T) = \sum_{t}^{N} \arg\min \|\theta_t - T_t\|^2 \tag{4}$$

B. FEATURE EXTRACTION

After the face alignment process, more discriminative features should be obtained for effective classification. In this sub-section, a novel hybrid feature representation is proposed to obtain more useful information, in which the pyramid

TABLE 1. Description of the used datasets.

Name	Subjects	Samples	Expressions	Specialties
CK+	123	593	8	Illumination variation, Pose variation
JAFFE	10	213	7	Illumination variation, Pose variation
Oulu-CASIA	80	10800	6	Illumination variation, Pose variation
AR	126	4000	13	Illumination variation, Pose variation, face occlusion

histogram orientation gradient method (PHOG), the edge histogram descriptor (EHD), and the local binary pattern (LBP) are included. Our main idea is to connects the above three algorithms in series to enhance the discrimination of facial features. First, the Canny detector is utilized to detect the edge of the image, and then to divide into pyramid level spatial grid. To estimate the edge contour, a Sobel mask is also utilized, in which the gradients are merged at each pyramid level. Next, the EHD is used to calculate the edge feature descriptor, where each image is converted into the corresponding gray format. It is divided into 4×4 blocks to get the percentage of pixels corresponding to the edge histogram, and then both the local and global histogram values are stored in the feature vector of each image. Finally, the LBP is utilized to convert each input image into an integer label array describing the small-scale appearance of each target. The LBP can describe the texture associated with each image, in which each pixel has the corresponding label. The LBP is also used to mark the center of the key points, and then the differences are calculated. In this process, when the difference value is greater than 0, it is denoted as 1, otherwise, it is 0.

C. CLASSIFICATION

After obtaining more discriminative features, we design a deep network for feature classification and output the final recognition results. To balance the speed and the accuracy, a fusion network is proposed, in which both the VGG-16 and the ResNet are included.

The structure of the classification network is shown as follows: 1) In the Conv1, the input size is $256 \times 256 \times 1$ and the kernel size is $[7 \times 7, 64]$; 2) In the Conv2_X, the input size is $64 \times 64 \times 64$ and the kernel size is $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$; 3) In the Conv3_X, the input size is $64 \times 64 \times 64$ and the kernel size is $\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$; 4) In the Conv4_X, the input size is $32 \times 32 \times 128$ and the kernel size is $\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$; 5) In the Conv5, the input size is $16 \times 16 \times 256$ and the kernel size is $[3 \times 3, 512]$; 6) In the Conv6, the input size is $8 \times 8 \times 512$ and the kernel size is $[3 \times 3, 1024]$; 7) In the Conv7, the input size is $4 \times 4 \times 1024$ and the kernel size is $[3 \times 3, 2048]$.

Note that the samples in the CK+ dataset are used as the input, and the input size is 256×256 and the output expression category is 8. The designed network has the following metrics: 1) Better ability to adapt to high-dimensional functions; 2) Obtain more advanced semantic features; 3) Fewer training parameters.

D. TRAINING

For the proposed model training, we focus on two parts, that is, the template generation (denoted as L_1) and the target matched (denoted as L_2). First, both the intra-class difference (denoted as $l_{1.1}(i)$) and the artifact of the face target (denoted as $l_{1.2}(i)$) are the key factors for the performance of the template generation. The loss function is shown in (5). Second, the optimal number of the templates is the key factor for the proposed model. the Elbow function is utilized as the loss function, as shown in (6). Where $l_{2.1}$ denotes the loss function of the cluster center point, $l_{2.2}$ denotes the loss function of the mean of all input samples. Additionally, the mini-batches are 256 in the stochastic gradient descent process, the initial learning rate is set as 0.1, and the range of the fine-tuning learning rate is 0.003 to 0.001.

$$L_1 = \arg\min \sum_{i=1}^{N} (l_{1.1}(i) + l_{1.2}(i)) \tag{5}$$

$$L_2 = \sum_{i=1}^{N} \sum_{j=1}^{P} \| l_{2.1} - l_{2.1} \|^2 \tag{6}$$

IV. EXPERIMENT AND DISCUSSION

A. DATASETS AND EXPERIMENT SETTINGS

1) DATASET

Following the state-of-the-art works, we evaluate the propose model on four benchmark dataset (as shown in Table 1), that is, the CK+ dataset, the JAFFE dataset, the Oulu-CASIA dataset, and the AR dataset [47], [48]. The details are described as follows: 1) 593 sequence samples are included in the CK+ dataset that consists of eight types of face expressions, and 123 participants completed these samples; 2) The JAFFE dataset consists of 213 samples finished by 10 females, and it includes seven types of facial expressions; 3) 10800 samples (six types of facial expressions) are collected in the Oulu-CASIA dataset, and these samples are finished by 80 participants; 4) The AR dataset consists of 4000 samples completed by 126 participants, and more challenges are included in the dataset, such as different lighting, different background, and face occlusion. Note that the data augmentation method is utilized in the training process to increase the training samples which is the common operation in most works, and it aims to let the model "see" more

TABLE 2. Results of the ablation study on the CK+ dataset.

Component	The proposed model (CK+)					
	1	2	3	4	5	6
Face alignment	✓		✓	✓	✓	✓
PHOG	✓	✓	✓	✓		
EHD	✓	✓	✓		✓	
LBP	✓	✓	✓			✓
VGG-16			✓			
Fusion network	✓	✓		✓	✓	✓
Accuracy(%)	98.9	91.5	95.4	91.1	87.6	88.2

TABLE 3. Results of the ablation study on the JAFFE dataset.

Component	The proposed model (JAFFE)					
	1	2	3	4	5	6
Face alignment	✓		✓	✓	✓	✓
PHOG	✓	✓	✓	✓		
EHD	✓	✓	✓		✓	
LBP	✓	✓	✓			✓
VGG-16			✓			
Fusion network	✓	✓		✓	✓	✓
Accuracy(%)	96.8	90.7	94.1	89.5	87.3	84.7

TABLE 4. Results of the ablation study on the Oulu-CASIA dataset.

Component	The proposed model (Oulu-CASIA)					
	1	2	3	4	5	6
Face alignment	✓		✓	✓	✓	✓
PHOG	✓	✓	✓	✓		
EHD	✓	✓	✓		✓	
LBP	✓	✓	✓			✓
VGG-16			✓			
Fusion network	✓	✓		✓	✓	✓
Accuracy(%)	94.5	91.3	91.7	88.6	85.3	85.8

challenging samples because more samples are rotated and tilted.

2) EXPERIMENT SETTINGS

The proposed model is conducted under Python 3.6, TensorFlow-GPU 1.11.0, Keras framework, NVIDIA GeForce RTX-2060 GPU (8 GB), 16GB memory, and Ubuntu 16.04 OS system.

B. ABLATION STUDY

In the sub-section, we evaluate the different designs of the proposed model by conducting a series of ablation experiments on four datasets. Note that the main contributions of this paper are the face alignment method and the hybrid feature representation method, so the following experiments and discussion are structured around two main contributions.

On the one hand, the different designs of the proposed model are shown as follows: 1)The PHOG is utilized for feature extraction, named PHOG; 2) The EHD is applied to extract features, named EHD; 3) Use the LBP for feature extraction, named LBP; 4) Apply the VGG-16 to model the features, named VGG-16; 5) Use the proposed fusion deep model to learn features, named Fusion network; 6) The model is designed by introducing the face alignment module, named Face alignment. Four groups of experimental results are shown in Table 2,3,4, and 5, and "✓" denotes the model includes the corresponding module. As shown in the tables, comparing Group 1 and 2, it can be seen that the model with the face alignment module can significantly improve the recognition rate. Especially for the CK+ dataset and the AR dataset, the recognition rate has increased by 7.4% and 10%, respectively. Comparing Group 1 and 3, the results show that the proposed fusion deep model can also improve the classification rate, and the recognition rates in the four datasets have increased by nearly 3%. Comparing Group 1, 4, 5, and 6, it can be concluded that the proposed hybrid feature representation method can obtain more discriminative features that can improve the performance of the model, and the recognition accuracies in the four data sets have improved by nearly 10%.

TABLE 5. Results of the ablation study on the AR dataset.

Component	The proposed model (AR)					
	1	2	3	4	5	6
Face alignment	✓		✓	✓	✓	✓
PHOG	✓	✓	✓	✓		
EHD	✓	✓	✓		✓	
LBP	✓	✓	✓			✓
VGG-16			✓			
Fusion network	✓	✓		✓	✓	✓
Accuracy(%)	98.7	88.7	95.2	87.3	86.4	85.2

To further show the effect of the main contributions, next, the comparison results of various designs are shown in Figure 4. Where the red histogram denotes the results of the proposed model and the green histogram denotes the results of the model without the face alignment module. It can be clearly seen that the introduction of the face alignment module can greatly improve the recognition rate of the model on the four datasets, and the design of the hybrid feature representation is also better than other feature extraction methods.

On the other hand, a key parameter critical to model performance in the work, that is, the number of templates generated N. First, following the state-of-the-art works [44] and [49], we set the number of the templates is 1 to 9. Next, we conduct a series of tests on four datasets, as shown in Figure 5. Since the scale of the datasets is not very large, the process is not time-consuming. It can be seen that when N=5, the model has the best recognition rate on the four datasets. Normally, a small number of templates can result in the inability to obtain high-level semantic

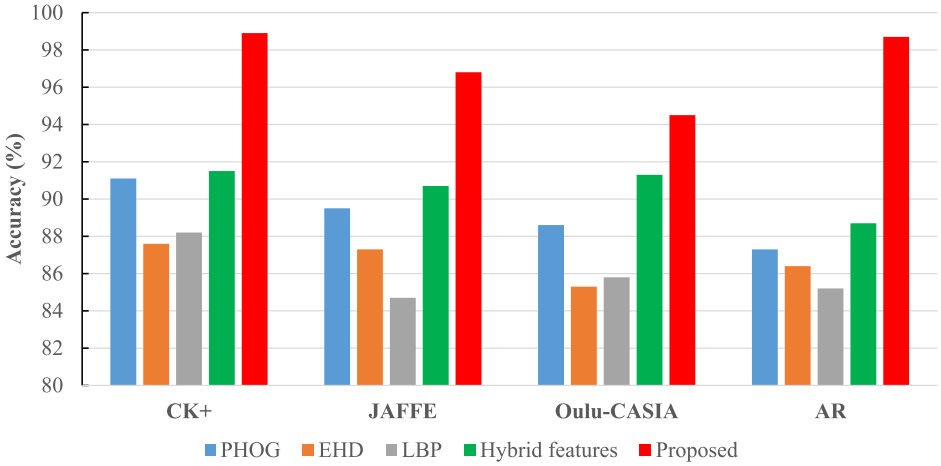


FIGURE 4. Accuracies of different designs on four datasets.

TABLE 6. AR dataset.

Basic expression\ Pose variation	Illumination\Pose variation	Face occlusion\ Pose variation
1.Neutral expression	5.Left light on	8.Wearing sun glasses
2.Smile	6.Right light on	9.Wearing sun glasses and left light on
3.Anger	7.All side lights on	10.Wearing sun glasses and right light on
4.Scream	9.Wearing sun glasses and left light on	11.Wearing scarf
	10.Wearing sun glasses and right light on	12.Wearing scarf and left light on
	12.Wwearing scarf and left light on	13.Wearing scarf and right light on
	13.Wearing scarf and right light on	

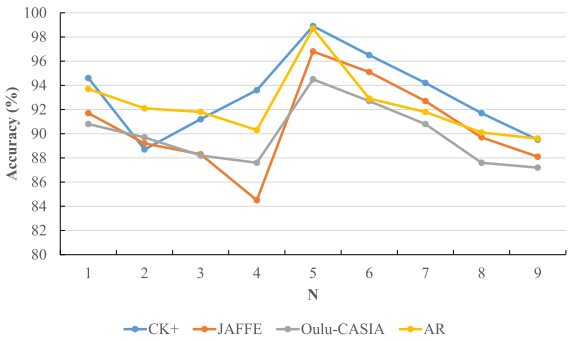


FIGURE 5. Classification performance varies with the *N* on four datasets.

features, and a large number of templates can lead to more misclassifications.

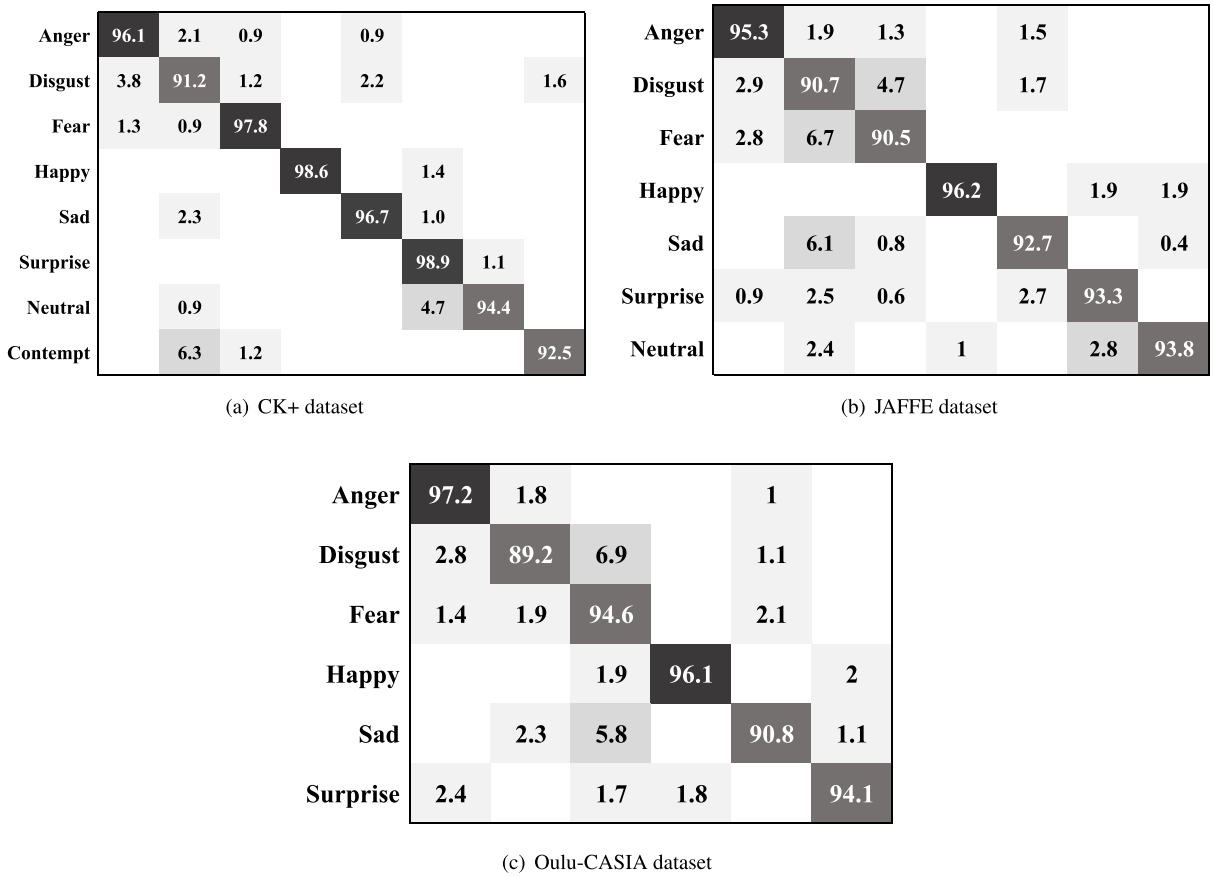
C. QUANTITATIVE ANALYSIS

In the sub-section, we quantitatively evaluate the performance of the proposed model by comparing the classification accuracy of each expression. First, the confusion matrixes are shown in Fig., including the results on the CK+ dataset (Figure 6(a)), the results on the JAFFE dataset (Figure 6(b)), and the results on the Oulu-CASIA dataset (Figure 6(c)). For the CK+ dataset, the proposed model achieves a recognition rate of more than 90% for all facial expressions, especially for ''Happy'' and ''Surprise'', the recognition rate is close to 99%. The recognition rate of the ''Disgust'' is not as high as others because it is similar with the ''anger'' and the ''sad''. The ''Contempt'' is an unstable emoticon, and there is a certain chance of misclassification during manual classification. For the JAFFE dataset, it has one less expression than the CK+ dataset, that is, the ''Contempt''. We also achieve satisfactory recognition rates for all facial expressions. The lower recognition rate shows in the ''Disgust'' and the ''Fear'', that is, 90.7% and 90.5%, respectively. This is because both express ''bad'' emotions and are prone to misclassification. The Oulu-CASIA dataset consists of six types of basic facial expressions, and the similarity between these expressions is higher than the expressions in other datasets. Misclassification of the ''Disgust'' is more likely to occur. Next, considering the challenge of the AR dataset, that is, illumination variation, pose variation, and face occlusion, we divide it into the following categories (as shown in Table 6), and the recognition rate of each expression is shown in Figure 7.

D. COMPARISON WITH STATE-OF-THE-ART WORKS

1) COMPARISON

In the sub-section, we compare the results of the proposed method with other state-of-the-art results. For a fair comparison, note that the experimental settings in this subsection follow the introduction listed in [50] and [51]. For the CK+ dataset and the JAFFE dataset, both traditional



(a) CK+ dataset

(b) JAFFE dataset

(c) Oulu-CASIA dataset

FIGURE 6. Confusion matrixes of three datasets.

TABLE 7. Comparison results on the CK+ dataset.

Methods	Types	Accuracy(%)
Hsieh et al. [52]	Traditional	94.7
Mlakar et al. [53]	Traditional	95.64
Happy et al. [54]	Traditional	97.09
Siddiqiet al. [55]	Traditional	96.83
Uccar et al. [56]	Traditional	95.17
Aly et al. [57]	Deep learning	88.14
Rivera et al. [58]	Deep learning	91.51
Lopes et al. [59]	Deep learning	93.68
Zhang et al. [35]	Deep learning	95.12
Yang et al. [50]	Deep learning	97.02
Saiyed et al. [60]	Deep learning	97.69
Proposed	Deep learning	**98.9**

TABLE 8. Comparison results on the JAFFE dataset.

Methods	Types	Accuracy(%)
Mlakar et al. [53]	Traditional	87.82
Happy et al. [54]	Traditional	91.79
Siddiqiet al. [55]	Traditional	96.33
Uccar et al. [56]	Traditional	94.65
Aly et al. [57]	Deep learning	87.32
Rivera et al. [58]	Deep learning	88.75
Lopes et al. [59]	Deep learning	88.73
Zhang et al. [35]	Deep learning	91.48
Yang et al. [50]	Deep learning	92.21
Proposed	Deep learning	**96.8**

models and deep learning-based models are considered for comparison, the results are shown in Table 7 and 8. The proposed model outperforms the traditional methods (hand-crafted feature) and the recognition rate is higher than most state-of-the-art deep learning-based models, especially the recognition rate has increased to 98.9% (CK+) and 96.8% (JAFFE). CK+: It is 1.21% higher than the traditional methods and 1.88% higher than the deep learning-methods. JAFFE: Compared with the traditional methods and deep learning-based methods, it has increased by 0.47% and 4.59%, respectively. Since the Oulu-CASIA and AR datasets are challenging, we only consider the deep learning methods for a comparison. In the [50] and [64], a lot of works are conducted and multiple sets of experimental results are obtained. It can be seen that the proposed model outperforms state-of-the-art models in terms of the recognition accuracy, that is, 94.5% (Oulu-CASIA) and 98.7% (AR), respectively.

TABLE 9. Comparison results on the Oulu-CASIA dataset.

Methods	Types	Accuracy(%)
Aly et al. [57]	Deep learning	84.21
Rivera et al. [58]	Deep learning	85.18
Lopes et al. [59]	Deep learning	86.42
Zhang et al. [35]	Deep learning	87.88
Yang et al. [50]	Deep learning	87.3
Yang et al. [50]	Deep learning	86.73
Yang et al. [50]	Deep learning	85.52
Yang et al. [50]	Deep learning	92.89
Proposed	Deep learning	**94.5**

TABLE 10. Comparison results on the AR dataset.

Methods	Types	Accuracy(%)
Dezfoulian et al. [61]	Deep learning	92.4
Chang et al. [62]	Deep learning	94.8
Munteanu et al. [63]	Deep learning	94.9
Munteanu et al. [63]	Deep learning	94.5
OLOYEDE et al. [64]	Deep learning	89.4
OLOYEDE et al. [64]	Deep learning	91.8
OLOYEDE et al. [64]	Deep learning	92.6
OLOYEDE et al. [64]	Deep learning	98.4
Proposed	Deep learning	**98.7**

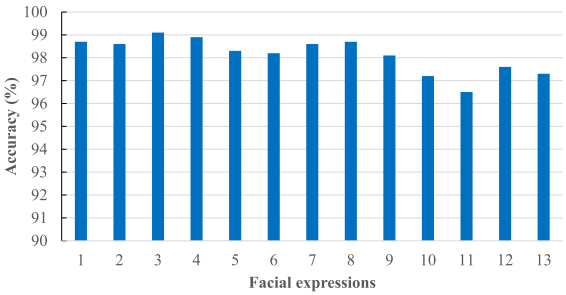


FIGURE 7. Recognition rate on the AR dataset.

2) DISCUSSION

The comparison results show that the proposed model can significantly improve the accuracy of facial expression classification. The main advantage of the proposed model is fully using the face alignment method to mitigating the impact of environmental noise, including illumination change, posture change, and target occlusion. Next, the proposed hybrid feature representation method can obtain more discriminative features. Hence, the proposed model can effectively distinguish similar expressions, such as the "Disgust" and the "Fear", due to the rational use of the within-class similarity. For expressions that cannot be accurately distinguished by human prior experiences, such as the "Neutral" and the "Contempt", the proposed model can also effectively classify, due to the highly discriminative facial features and effective classification network.

V. CONCLUSION

In this work, we propose a deep model for face expression recognition based on deep learning. First, we proposed a novel face alignment method that consists of target pose estimation, templates generation, and target matched. This method aims to reduce the intra-class difference to improve the recognition performance. Next, we propose an effective feature extraction module to obtain key semantic information to reduce the intra-class similarity. After that, a lightweight backbone is designed to train model with low computational and low data.

In future works, we will focus on real-time face expression recognition in videos and design a more concise network. Furthermore, we will research a model for occlusion robust face recognition, especially for mask face detection and recognition and collect a face expression dataset in a more challenging real-world scenario.

REFERENCES

[1] M. O. Oloyede and G. P. Hancke, "Unimodal and multimodal biometric sensing systems: A review," *IEEE Access*, vol. 4, pp. 7532–7555, 2016.

[2] N. Dagnes, E. Vezzetti, F. Marcolin, and S. Tornincasa, "Occlusion detection and restoration techniques for 3D face recognition: A literature review," *Mach. Vis. Appl.*, vol. 29, no. 5, pp. 789–813, Jul. 2018.

[3] A. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," *Appl. Sci.*, vol. 7, no. 1, p. 110, Jan. 2017.

[4] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 3, pp. 1–42, Apr. 2016.

[5] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 471–478.

[6] W. Deng, J. Hu, and J. Guo, "Compressive binary patterns: Designing a robust binary face descriptor with random-field eigenfilters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 758–767, Mar. 2019.

[7] J. Lu, V. E. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1979–1993, Aug. 2018.

[8] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel, "Multi-layer temporal graphical model for head pose estimation in real-world videos," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 3392–3396.

[9] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1, no. 2. Cambridge, MA, USA: MIT Press, 2016.

[10] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 773–782.

[11] S. Jia, G. Guo, and Z. Xu, "A survey on 3D mask presentation attack detection and countermeasures," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107032.

[12] S.-Q. Liu, X. Lan, and P. C. Yuen, "Temporal similarity analysis of remote photoplethysmography for fast 3D mask face presentation attack detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2608–2616.

[13] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, H. Chen, Y. Miao, Z. Huang, and J. Liang, "Masked face recognition dataset and application," 2020, *arXiv:2003.09093*. [Online]. Available: http://arxiv.org/abs/2003.09093

[14] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," *Measurement*, vol. 167, Jan. 2021, Art. no. 108288.

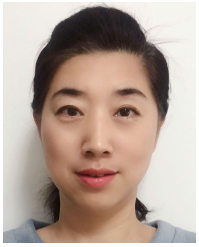
[15] S. Kim, J. Lee, S. Kang, J. Lee, and H.-J. Yoo, "A power-efficient CNN accelerator with similar feature skipping for face recognition in mobile devices," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 4, pp. 1181–1193, Apr. 2020.

[16] X. Wang, S. Wang, H. Shi, J. Wang, and T. Mei, "Co-mining: Deep face recognition with noisy labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9358–9367.
[17] X. Liu, "Discriminative face alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1941–1954, Nov. 2009.
[18] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 200–215, Jan. 2011.
[19] A. Bulat and Y. Tzimiropoulos, "Convolutional aggregation of local evidence for large pose face alignment," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–13.
[20] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1513–1520.
[21] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.
[22] Y. Zhong, J. Chen, and B. Huang, "Toward end-to-end face recognition through alignment learning," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1213–1217, Aug. 2017.
[23] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 212–220.
[24] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
[25] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.
[26] C. Ding and D. Tao, "Pose-invariant face recognition with homography-based normalization," *Pattern Recognit.*, vol. 66, pp. 144–152, Jun. 2017.
[27] X. Zeng, J. Huang, and C. Ding, "Soft-ranking label encoding for robust facial age estimation," *IEEE Access*, vol. 8, pp. 134209–134218, 2020.
[28] T. Zhou, C. Ding, S. Lin, X. Wang, and D. Tao, "Learning oracle attention for high-fidelity face completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7680–7689.
[29] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 1002–1014, Apr. 2018.
[30] C. Ding, J. Choi, D. Tao, and L. S. Davis, "Multi-directional multi-level dual-cross patterns for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 518–531, Mar. 2016.
[31] M. R. Mohammadi, E. Fatemizadeh, and M. H. Mahoor, "PCA-based dictionary building for accurate facial expression recognition via sparse representation," *J. Vis. Commun. Image Represent.*, vol. 25, no. 5, pp. 1082–1092, Jul. 2014.
[32] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.
[33] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 151–160, Apr. 2013.
[34] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.
[35] W. Zhang, Y. Zhang, L. Ma, J. Guan, and S. Gong, "Multimodal learning for facial expression recognition," *Pattern Recognit.*, vol. 48, no. 10, pp. 3191–3202, Oct. 2015.
[36] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–8.
[37] L. Anzalone, P. Barra, S. Barra, F. Narducci, and M. Nappi, "Transfer learning for facial attributes prediction and clustering," in *Smart City and Informatization*, G. Wang, A. El Saddik, X. Lai, G. M. Perez, and K.-K. R. Choo, Eds. Singapore: Springer, 2019, pp. 105–117, doi: 10.1007/978-981-15-1301-5_9.
[38] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *Proc. 16th Int. Conf. Multimodal Interact.*, 2014, pp. 461–466.
[39] A. Fathallah, L. Abdi, and A. Douik, "Facial expression recognition via deep learning," in *Proc. IEEE/ACS 14th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Oct. 2017, pp. 303–308.
[40] M. Liu, S. Li, S. Shan, and X. Chen, "AU-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, Jul. 2015.
[41] A. Jourabloo and X. Liu, "Large-pose face alignment via CNN-based dense 3D model fitting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4188–4196.
[42] Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, "Eidetic 3D LSTM: A model for video prediction and beyond," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–14.
[43] J. Yu, H. Gao, W. Yang, Y. Jiang, W. Chin, N. Kubota, and Z. Ju, "A discriminative deep model with feature fusion and temporal attention for human action recognition," *IEEE Access*, vol. 8, pp. 43243–43255, 2020.
[44] Z. An, W. Deng, J. Hu, Y. Zhong, and Y. Zhao, "APA: Adaptive pose alignment for pose-invariant face recognition," *IEEE Access*, vol. 7, pp. 14653–14670, 2019.
[45] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4838–4846.
[46] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[47] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition-Workshops*, Jun. 2010, pp. 94–101.
[48] M. J. Lyons, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets (IVC special Issue)," 2020, *arXiv:2009.05938*. [Online]. Available: http://arxiv.org/abs/2009.05938
[49] F. Ahmad, L.-M. Cheng, and A. Khan, "Lightweight and privacy-preserving template generation for palm-vein-based human recognition," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 184–194, 2020.
[50] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630–4640, 2018.
[51] N. B. Kar, K. S. Babu, A. K. Sangaiah, and S. Bakshi, "Face expression recognition system based on ripplet transform type II and least square SVM," *Multimedia Tools Appl.*, vol. 78, no. 4, pp. 4789–4812, Feb. 2019.
[52] C.-C. Hsieh, M.-H. Hsih, M.-K. Jiang, Y.-M. Cheng, and E.-H. Liang, "Effective semantic features for facial expressions recognition using SVM," *Multimedia Tools Appl.*, vol. 75, no. 11, pp. 6663–6682, Jun. 2016.
[53] U. Mlakar and B. Potočnik, "Automated facial expression recognition based on histograms of oriented gradient feature vector differences," *Signal, Image Video Process.*, vol. 9, no. S1, pp. 245–253, Dec. 2015.
[54] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2015.
[55] M. H. Siddiqi, R. Ali, A. M. Khan, Y.-T. Park, and S. Lee, "Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1386–1398, Apr. 2015.
[56] A. Uçar, Y. Demir, and C. Güzeliş, "A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering," *Neural Comput. Appl.*, vol. 27, no. 1, pp. 131–142, Jan. 2016.
[57] S. Aly, A. L. Abbott, and M. Torki, "A multi-modal feature fusion framework for kinect-based facial expression recognition using dual kernel discriminant analysis (DKDA)," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.
[58] A. R. Rivera, J. R. Castillo, and O. O. Chae, "Local directional number pattern for face analysis: Face and expression recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1740–1752, May 2013.
[59] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.
[60] S. Umer, R. K. Rout, C. Pero, and M. Nappi, "Facial expression recognition with trade-offs between data augmentation and deep learning features," *J. Ambient Intell. Humanized Comput.*, early access, pp. 1–15, Jan. 2021, doi: 10.1007/s12652-020-02845-8.
[61] M. Shakeri, M. H. Dezfoulian, H. Khotanlou, A. H. Barati, and Y. Masoumi, "Image contrast enhancement using fuzzy clustering with adaptive cluster parameter and sub-histogram equalization," *Digit. Signal Process.*, vol. 62, pp. 224–237, Mar. 2017.

[62] Y. Chang, C. Jung, P. Ke, H. Song, and J. Hwang, "Automatic contrast-limited adaptive histogram equalization with dual gamma correction," *IEEE Access*, vol. 6, pp. 11782–11792, 2018.
[63] C. Munteanu and A. Rosa, "Gray-scale image enhancement as an automatic process driven by evolution," *IEEE Trans. Syst., Man Cybern. B, Cybern.*, vol. 34, no. 2, pp. 1292–1298, Apr. 2004.
[64] M. O. Oloyede, G. P. Hancke, and H. C. Myburgh, "Improving face recognition systems using a new image enhancement technique, hybrid features and the convolutional neural network," *IEEE Access*, vol. 6, pp. 75181–75191, 2018.

JUN LIU received the B.E. and M.E. degrees from the Shenyang University of Technology, in 1995 and 2000, respectively, and the dual Ph.D. degree from the Graduate University of Chinese Academy of Sciences, and the Shenyang Institute of Automation, Chinese Academy of Sciences, in 2010. He is currently an Associate Professor with Shenyang Ligong University. His research interests include intelligent sensors and detection technology, image and signal processing, and intelligent robots.

YANJUN FENG received the B.E. degree from the Liaoning University of Technology, in 1997, and the M.E. degree from the Shenyang University of Technology, in 2000. She is currently a Lecturer with Shenyang Ligong University. Her research interests include the IoT technology and intelligent information processing.



HONGXIA WANG received the B.E. and M.E. degrees from Shenyang Ligong University, in 1999 and 2005, respectively, and the Ph.D. degree from the Nanjing University of Technology, in 2011. She is currently a Professor with Shenyang Ligong University. Her research interests include network computing and artificial intelligence.