IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 5, 2021, accepted January 8, 2021, date of publication January 13, 2021, date of current version January 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3051403

An End-to-End Deep Model With Discriminative Facial Features for Facial Expression Recognition

JUN LIU[1], HONGXIA WANG[2], AND YANJUN FENG[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: Hongxia Wang (sunny58258@sina.com)

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

ABSTRACT Due to the complex challenges of the environment and emotion expressions, most facial expression recognition systems cannot achieve a high recognition rate. More discriminative features can describe facial expressions more accurately, so facial feature extraction is the key technology for facial expression recognition. In this article, an effective end-to-end deep model is proposed to improve the accuracy of face recognition. Considering the importance of data pre-processing (very few studies have focused on this process), first, a data enhancement method is proposed to locate the range of the face target and enhance the image contrast. Next, to obtain further discriminative features, a hybrid feature representation method is proposed, in which four typical feature extraction method are combined. After that, an effective deep model is designed to train and test the samples which can obtain the optimal parameters with less computation cost. Ablation study results show that the proposed hybrid feature representation method can help improve recognition accuracy. Finally, to comprehensively evaluate the performance of the proposed model, a series of experiments are conducted on three benchmark datasets. The recognition rate is achieved 94.5%, 98.6%, and 97.2% for FER2013, AR dataset, and CK+ dataset, respectively.

INDEX TERMS Face recognition, feature extraction, data enhancement, CNN.

I. INTRODUCTION

In recent years, face recognition has been successful applied in many areas, such as, video security, video surveillance, mobile phone unlocking, game entertainment, so it has become a hot research topic [1], [2]. Driven by the success of Convolution Neural Networks (CNNs), most deep learning-based methods achieve promising results in the research area of face recognition [3]. The pipeline of face recognition consists of three stages, that is, face object detection, facial feature extraction, and classification process [4]. The main challenge is the reduced recognition effect caused by environmental changes and facial expression changes [5]–[7]. In this article, to solve this challenge, the research of feature extraction and classification is conducted.

Typical deep models are described as follows: 1) The DeepID2 deep model is proposed to reduce intra-personal variations [8]; 2) A novel deep model is proposed to solve the problem caused by missing mutation patterns and image degradation [9]; 3) A nine-layer deep model is designed to improve the alignment and the representation [10]. Recently, due to the COVID-19 pandemic, there are more scenes of facial occlusion (mask-wearing required), so facial expression recognition under facial occlusion has received widespread attention [11]. For example, the PDSN is proposed to specifically solve the problem of mask-wearing face recognition, and the results show that its performance is better than CNN-based models [12]. For mobile face unlocking payment scenarios, the PAD system is proposed based on the remote Photoplethysmography biomedical technique, which can achieve fast and accurate recognition under the masked situation [13]. Deep learning-based face recognition models rely on high-quality datasets. The RMFRD, MFDD, and SMFRD datasets are proposed that are the real-world masked face benchmark datasets [14].

The main advantage of the deep learning-based methods is that a large of samples can be used for training, so as to learn a face representation that is robust to changes in the training data [15]. This method does not need to design specific features that are robust to different types of intra-class differences, such as lighting, posture, facial expression, age, but can learn them from training data [16]. The main shortcoming of deep learning methods is that they need to use very

The associate editor coordinating the review of this manuscript and approving it for publication was Long Wang.

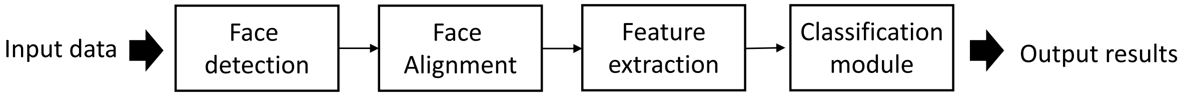


FIGURE 1. Face recognition system.

large datasets for training, and these datasets need to contain enough changes so that they can be generalized to unseen samples [17]. In addition to learning discriminative features, neural networks can also reduce dimensionality, and can be trained into classifiers or use metric learning methods [18]. CNN is considered to be an end-to-end trainable system and does not need to be combined with any other specific methods [19].

In summary, some challenges and limitations remain: 1) Most models only have a better recognition effect on datasets with fewer participants, and a single environment; 2) When the samples have complex light changes and occlusion, most models cannot determine the satisfactory recognition rates; 3) Most deep models rely on a large amount of data to achieve a high recognition rate, but they do not work well for datasets with fewer samples. Our ideas to solve the above problems are shown as follows: 1) Design an effective data pre-processing module to enhance the generalization of the proposed model; 2) Design a module that can extract highly discriminative features to alleviate the impact of environmental changes; 3) Design a deep model that efficiently obtains the optimal parameters to get high-level semantic features with less training samples.

To solve the above problems, in this article, an end-to-end face recognition deep model is proposed. Our research is motivated by the recent studies, especially the methods of feature extraction and classification, such as [20], [21], [22], [23], [24], [25], and [26]. The proposed model can reduce the intra-class difference to improve the recognition accuracy. The core of this model is to obtain more discriminative features that cannot be affected by environmental changes, such as lighting and occlusion. To obtain these facial features, both an effective data pre-processing method and a new hybrid feature representation method are proposed that are effective improvements of existing methods. Few studies focus on the data pre-processing but it is key for face recognition performance. Next, considering the importance of parameters in deep learning, a deep model is designed to obtain the optimal parameters without higher computation cost.

The main contributions of this work are shown as follows:

1) An effective data enhancement method is proposed to enhance the contrast of the data. The new designed transformation functions are utilized to map the intensity value of the original input image, and then output the improved output data.
2) A new hybrid facial feature representation method is proposed to obtain more discriminative features that can help the classification system learn facial features better. Considering different challenges presented in the real-world, three state-of-the-art feature representation methods are fused, that is, enhancing texture and shape information, eliminating useless noise information.
3) A new deep model is designed to train and test the proposed method which combines the VGG and the ResNet. Compared with the traditional single structure network, this deep model can obtain more discriminative semantic information with a lower calculation cost.
4) A series of experiments are conducted, including the ablation studies and comparation analysis. To evaluate the performance of the model in practical applications and analysis the advantages of advanced models, both the competition benchmark and the datasets collected from the actual scene are utilized to test the proposed model.

The remainder of this article is organized as follows. Section II reviews the related works. Section III describes the proposed model in details, including the structure, the data pre-processing, the hybrid feature extraction, the classification system, and the training process. In section IV, a series of experiments are conducted, and then the results are compared with other state-of-the-art works. Section V concludes this article and gives the further research direction.

II. RELATED WORK

A. FACE RECOGNITION SYSTEM

In recent years, face recognition technology has been greatly advanced. Traditional methods rely on the combination of artificially designed features, such as edge and texture descriptions [8]. Machine learning techniques rely on the rich data, such as principal component analysis, linear discriminant analysis, or support vector machines [27]. It is very difficult for human-based design to adapt to different changes in an unconstrained environment, such as lighting and occlusion. hence, most researchers focus on specific methods for each type of change, such as methods that can cope with different ages, methods to cope with different postures, methods to cope with different lighting conditions [28], [29]. Recently, traditional face recognition methods have been replaced by deep learning-based methods, such as CNNs [30].

Normally, the face recognition system consists of the following sub-modules, as shown in Fig. 1. 1) Face Detection: the face detector is used to find the position of the face in the image. If there is a face, it returns the coordinates of the bounding box containing each face. 2) Face alignment: the core of face alignment is to use a set of reference points at a fixed position in the image to scale and crop the face

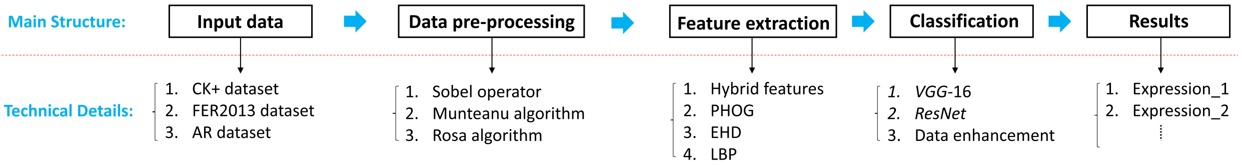


FIGURE 2. The structure of the proposed deep model for face recognition that consists of three mian contributions, namely, data enhancement method, more discriminative facial features extraction method, and classification network.

image. This process usually requires the use of a feature point detector to find a set of facial feature points. For the 2D alignment situation, it is to find the best affine transformation that is most suitable for the reference point. More complex 3D alignment algorithms can also realize the frontal face, that is, adjust the posture of the face to the front. 3) Facial feature extraction: the pixel values of the face image can be converted into compact and discriminable feature vectors, which are also called templates. All faces of the same subject should be mapped to similar feature vectors. 4) Classification system: two templates are compared to obtain a similarity score, which gives the possibility that the two belong to the same subject. Note that the above summary refers to the [28], [31], [32], and [33].

B. DEEP LEARNING-BASED FACE RECOGNITION

Motivated by the deep learning technology, face recognition has achieved rapid development. The latest research results of face recognition show that the facial feature expression obtained by deep learning has important characteristics that manual feature expression does not achieve [34]. For example, it is moderately sparse and has strong selectivity for face identity and face attributes, and achieves good robustness to partial occlusion [35]. These deep features are naturally obtained through big data training, and no explicit constraints or post-processing are introduced into the model [36]. This is the main reason that deep learning can be successfully applied to face recognition.

There are many typical applications of deep learning in face recognition, including face recognition method based on CNNs, deep non-linear face shape extraction method, robust modeling of face pose based on deep learning, fully automatic face recognition in a constrained environment, face recognition under video surveillance based on deep learning, low-resolution face recognition based on deep learning, and other face-related information recognition based on deep learning [18], [37]–[39]. Among them, CNN is the better learning algorithm that successfully trains a multilayer network structure. The face recognition method based on CNNs can mine the local features of data and extract global training features and classification [40]. Its weight sharing structure network makes it more similar to biological neural network, and it has been successfully applied in various fields of pattern recognition [41]. CNN makes full use of the features of the locality contained in the data itself by combining the local perception area of the face image space, shared weights, and spatial or temporal down-sampling to optimize the model structure and ensure a certain displacement invariance [42], [43].

C. CHALLENGES AND MOTIVATIONS

Facial expression recognition aims to analyze many types of human expressions that are defined by human experience in the real world. Normally, most of the works focus on achieving high recognition rates in complex environments. An effective facial expression recognition system should solve the following challenges that are also the motivation of this work.

In real applications, first, the model is run in a particularly complex environment, including complex lighting, facial occlusion, and changing poses. This is the main challenge for facial expression recognition. Although many datasets are collected in real scenarios and set challenges manually, they do not represent all actual situations. It is a large gap between real action and the training samples. To solve this challenge, more discriminative features should be extracted.

Next, most facial expression recognition methods focus on improving the accuracy of a particular dataset instead of improving its generalization and robustness. Because of this reason, the performance of many state-of-the-art models in different types of data sets varies greatly. To solve this challenge, the model should verify the design ideas on multiple types of data sets.

As is well known, third, deep learning-based models rely on a large amount of high-quality data, this leads to complex training networks and high computational resource consumption. This is also a challenge that limits the development of deep models. To solve this challenge, lightweight training networks and efficient training schemes should be proposed.

III. PROPOSED MODEL

A. STRUCTURE

To tackle the above-mentioned consideration, three challenges must be solved, that is, ineffective data processing strategies, fewer representative features, and low classification accuracies. In this article, a full end-to-end face recognition deep model is proposed that consists of a new designed data enhancement method, a new facial representation extraction method, and an effective feature classification method, as shown in Fig. 2.

B. DATA PRE-PROCESSING

For data pre-processing, our main purpose is to obtain the optimal model parameters that can express the effectively mapping process. On the one hand, to enhance the gray value of the input facial data, we design a transformation method based on the Munteanu and Rosa function, as shown in (1). Where $i(x, y)$ denotes the input grey value, $i^E(x, y)$ denotes the enhanced grey value, $m(x, y)$ denotes the grey mean function, $\delta(x, y)$ denotes the standard deviation function, and M denotes the global mean function, respectively; k, k_m, and k_δ denote the enhancement weights, and b denotes the deviation coefficient.

$$i^E(x, y) = k(\frac{M(x, y)((i(x, y) - k_m m(x, y))}{\delta(x, y) + k_\delta}) + b \qquad (1)$$

On the other hand, motived by the classical image enhancement technology, a new evaluation function is designed to automatically locate the face object range. The idea of our design is to effectively calculate four key parameters, including the number of pixels in the edge (N_e), the number of pixels in the foreground (N_f), the value of signal-to-noise ratio (r), and the value of entropic measure (Δ). First, the Sobel operator is utilized as the edge detector because of immunity to noise and high contrast segmentation, as shown in (2). Where m and n denote vertical dimension and horizontal dimension, and $E(x, y)$ denotes the edge intensity of the input data. Next, N_f can also be shown in (3), where $P(x, y)$ denotes the enhanced pixel in the foreground. After that, Δ is shown in (4), where f_n denotes the frequency of grey pixels. Finally, r is shown in (5), where $l_{\max}$ denotes the value of the maximum pixel intensity.

$$N_e = \sum_{i=1}^{m} \sum_{j=1}^{n} E(x, y) \qquad (2)$$

$$N_f = \sum_{i=1}^{m} \sum_{j=1}^{n} P(x, y) \qquad (3)$$

$$\Delta = -\sum_{n}^{256} f_n \log(f_n) \qquad (4)$$

$$r = \lg[\frac{(1 - l_{\max})}{\sqrt{m \cdot n}[i(x, y) - i^E(x, y)]}] \qquad (5)$$

C. DISCRIMINATIVE FEATURE EXTRACTION

The discriminative features are important for face recognition because these features are directly used for classification and then input the final results. our main purpose is to simplify a large number of useless features that come from the pre-processed facial data, and then obtain unique and effective features for classification.

In this article, a new hybrid feature representation method is proposed to extract the enhanced data. Motived by the spatial pyramid network, first, we design a new pyramid histogram orientation gradient method (PHOG) to active extract features, in which the histogram of orientation gradient technology is introduced and the canny edge detector is also utilized. Specifically, the Canny edge detector is used to form a feature space corresponding to the pyramid grid, and the histogram is used to fuse and match features with feature space. Next, considering the importance of texture and shape information, the edge histogram descriptor (EHD) is used to describe the edge features of each input image, and the details are shown as follows: 1) Divide the image into 4×4 sub-images, and the purpose is to locate the edge direction in a certain area; 2) Each sub-image is divided into several image blocks, and the number of image blocks divided depends on the situation; 3) Each image block is further divided into 4 sub-blocks. After that, the local binary pattern (LBP) method is utilized to further describe the texture features, and the details are shown as follows: 1) Divide the detection window into 16×16 cells; 2) For a pixel in each cell, the gray value of the adjacent 8 pixels is compared with it, if the value of the surrounding pixel is greater than the value of the central pixel, the position of the pixel is denoted as 1, otherwise it is 0; 3) Calculate the histogram of each cell, that is, the frequency of occurrence of each number; 4) Normalize the histogram; 5) Connect the obtained statistical histogram of each cell into a feature vector, that is, the LBP texture feature vector of the entire image. Hence, the discriminative features that are useful for face recognition system are obtained by the above methods.

D. CLASSIFICATION

When obtaining useful features, the classification deep network focuses on modelling these features to output the final results. In the work a fusion deep network is designed, in which the first half of the VGG network is replaced by the ResNet network and the Softmax function is used as classification layer. As shown in Table 1, the details of the designed classification network is described.Note that the input is from the CK+ dataset, and the size is 256 × 256. The designed deep model is effective for data enhancement and new hybrid feature selection. This is because the short connections in the ResNet have better ability to fit high-dimensional functions than other ordinary connections, and more high-level semantic features can be obtained based on the deep structure of the VGG. Hence, the designed classification deep model is better than other traditional CNN-based networks for feature modelling.

E. TRAINING

In the training process, the Softmax function and the Arcface function are utilized as the loss functions, the details are shown in [44]–[46]. The mini-batches is set as 256 that is used for stochastic gradient descent processing, the initial learning rate is 0.1, the fine-tuning learning rate is from 0.003 to 0.001. Note that subtract the average value of each channel for each pixel, and then reduce the overfitting on the color image, use the conversion to monochrome enhancement with a probability of 20%.

TABLE 1. The proposed classification network.

Backbone	Layer name	Input size	Layers	Output size
ResNet	CONV1	$256 \times 256 \times 1$	$7 \times 7, 64, stride2$	$128 \times 128 \times 64$
	Max pool	$128 \times 128 \times 64$	$3 \times 3, stride2$	$64 \times 64 \times 64$
	CONV2_X	$64 \times 64 \times 64$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2, stride1$	$64 \times 64 \times 64$
	CONV3_X	$64 \times 64 \times 64$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2, stride2$	$32 \times 32 \times 128$
	CONV4_X	$32 \times 32 \times 128$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2, stride2$	$16 \times 16 \times 256$
VGG	CONV5	$16 \times 16 \times 256$	$3 \times 3, 512stride2$	$8 \times 8 \times 512$
	CONV6	$8 \times 8 \times 512$	$3 \times 3, 1024stride2$	$4 \times 4 \times 1024$
	CONV7	$4 \times 4 \times 1024$	$3 \times 3, 2048stride2$	$2 \times 2 \times 2048$
	Flatten	$2 \times 2 \times 2048$	\	8192
	FC	8192	\	8

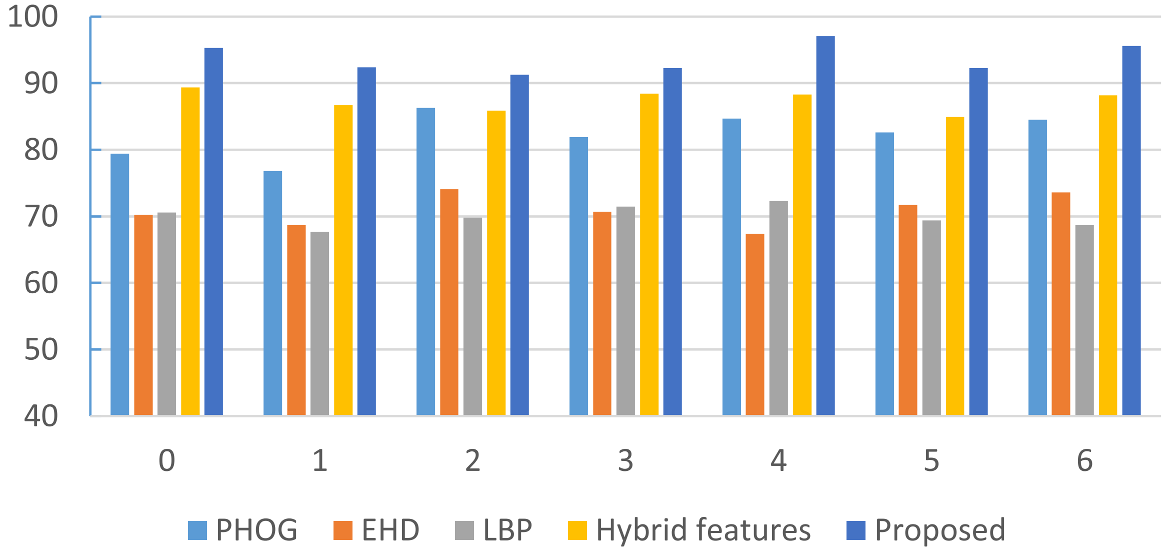


FIGURE 3. Comparison of recognition rates of different methods.

IV. EXPERIMENTS AND DISCUSSION

A. EXPERIMENTAL SETTINGS

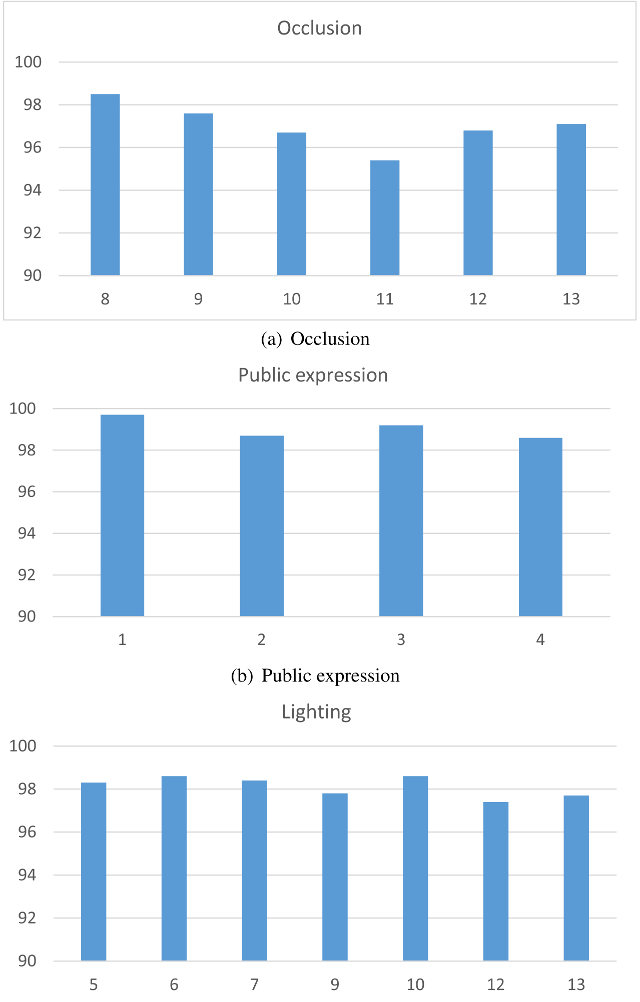
For fair comparison, in this work, the experiments are conducted on the public benchmark, namely, FER2013 database, which is used for international competition. It consists of 35886 facial expression images, in which 28708 test images is for training, 3589 public is for public verification, 3589 is for private verification, and each image is fixed in size to 48×48 grayscale image composition. A total of 7 expressions are included, corresponding to the number labels 0-6, the corresponding labels and Chinese and English of the specific expressions are as follows: 0 anger; 1 disgust; 2 fear; 3 happy; 4 sad; 5 surprised; 6 normal neutral. Note that the dataset does not directly give images, but saves expressions, picture data, and purpose data to a csv file.

On the other hand, to comprehensively analyze the performance of the proposed model, especially the application in real situations, both the AR face dataset and the CK+ dataset are also used. The AR dataset contains more than 4,000 color images, corresponding to the faces of 126 people (70 men and 56 women). The images have positive expressions, these faces have different facial expressions, lighting conditions and occlusion (sunglasses and scarves). Considering the application scenario in the real-world, the samples of the dataset are divided into three parts, that is, public expression, samples with changing lighting, and samples with occlusion. The CK+ dataset consists of eight types of facial expressions, in which both the "Neutral" and the "Contempt" are the challenge for facial expression recognition. The description of the above three datasets is shown in Table 2.

The experimental environment consists of Python 3.6, TensorFlow-GPU 1.11.0, NVIDIA GeForce RTX-2060 GPU, 16GB memory, and Ubuntu 16.04 OS system.

TABLE 2. Description of the used datasets.

Name	Samples	Expressions	Specialties
CK+	593	8	Illumination variation, Pose variation
FER2013	35886	7	Illumination variation, Pose variation
AR	4000	13	Illumination variation, Pose variation, face occlusion



(a) Occlusion

(b) Public expression

(c) Lighting

FIGURE 4. Recognition rate on AR face dataset.

TABLE 3. Results of the ablation study on the FER2013 dataset.

Component	The proposed model (FER2013)					
	1	2	3	4	5	6
PHOG	✓	✓	✓	✓		
EHD	✓	✓	✓		✓	
LBP	✓	✓	✓			✓
VGG-16		✓				
Data enhancement	✓	✓		✓	✓	✓
Fusion network	✓		✓	✓	✓	✓
Accuracy(%)	94.5	89.7	92.6	89.3	87.4	87.5

TABLE 4. Results of the ablation study on the AR dataset.

Component	The proposed model (AR)					
	1	2	3	4	5	6
PHOG	✓	✓	✓	✓		
EHD	✓	✓	✓		✓	
LBP	✓	✓	✓			✓
VGG-16		✓				
Data enhancement	✓	✓		✓	✓	✓
Fusion network	✓		✓	✓	✓	✓
Accuracy(%)	98.6	93.7	97.1	89.3	85.4	87.5

TABLE 5. Results of the ablation study on the CK+ dataset.

Component	The proposed model (CK+)					
	1	2	3	4	5	6
PHOG	✓	✓	✓	✓		
EHD	✓	✓	✓		✓	
LBP	✓	✓	✓			✓
VGG-16		✓				
Data enhancement	✓	✓		✓	✓	✓
Fusion network	✓		✓	✓	✓	✓
Accuracy(%)	97.2	92.8	95.8	90.2	88.5	87.4

TABLE 6. Results on the FER2013 database.

Method	Accuracy (%)
Alexnet	74.2
Inception	81.3
ResNet	85.2
DenseNet	83.3
Proposed	**94.5**

B. ABLATION STUDY

In this sub-section, the ideas of our design are evaluated on three datasets. In our work, two main contributions are given, that is, a new hybrid feature representation method, and an effective classfication network. Hence, different feature representation methods and backbones are tested and the results are discussed. Specifically, results are shown in Table 3,4,and 5. Where using the PHOG for feature extraction, named PHOG; the EHD is used to extract features, named EHD; selecting the LBP for feature extraction, named LBP; modeling the features by the VGG-16, named VGG-16; training the model with the data enhancement opreation, named Data enhancement; using the fusion network for classification, named Fusion network. Comparing Group 1 and 2, it is can be seen that the designed fusion network is better that the VGG-16 in terms of classification. Comparing Group 1 and 3, the results show that the introduction of the data enhancement can improve the recognition rate. Comparing Group 1, 4, 5, and 6, it can be concluded that the

TABLE 7. AR face dataset.

(No.) Public expression	(No.)Lighting	(No.) Occlusion
1.Neutral expression 2.Smile 3.Anger 4.Scream	5.Left light on 6.Right light on 7.All side lights on 9.Wearing sun glasses and left light on 10.Wearing sun glasses and right light on 12.wearing scarf and left light on 13.wearing scarf and right light on	8.wearing sun glasses 9.wearing sun glasses and left light on 10.wearing sun glasses and right light on 11.wearing scarf 12.wearing scarf and left light on 13.wearing scarf and right light on

proposed hybrid feature representation method can significantly improve model classification performance.

The comparison of the recognition rates of different expressions in the FER2013 dataset by each method is shown in Fig. 3. It can be concluded that the proposed method achieves the best recognition accuracies on all the types of data.

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

In the sub-section, the results of the proposed model are compared with other state-of-the-art methods. Next, considering the comprehensive performance of the proposed model, three datasets are selected for comparison, that is, a benchmark dataset for the international competition, and two challenge datasets collected in the various real scenarios.

On the one hand, Table 6 shows the results obtained on the FER2013 database. Four typical deep models are used for comparison, that is, Alexnet [47], Inception [48], ResNet [49], and DenseNet [50]. It can be seen that the proposed method achieves a recognition rate of 94.5%, which is about 10% higher than other typical deep models. This is because the proposed model can obtain more discriminative features than other deep models and the data pre-processing method can also enhance the contrast of the data.

TABLE 8. Results on the AR face dataset.

Method	Accuracy (%)
HE	92.4
CLAHE	94.8
MP	94.9
MG	94.5
Proposed	**98.6**

On the other hand, for analyzing the performance in the real-world, three parts samples are used to evaluate the proposed model, the details are shown in Table 7. The recognition rates of three different types of data are shown in Fig. 4. The comparison result between the proposed model and other state-of-the-art methods is shown in Table 8. The following methods are considered: 1) HE: A model is proposed to output final results without parameter setting [51]; 2) CLAHE: A method is proposed to enhance image contrast [52]; 3) MG/MP: A new data enhancement method is proposed based on the evolutionary optimization technology [20]. It can be seen that the proposed method achieves the highest recognition rate in terms of changing lighting and occlusion because the data preprocessing method enhances the contrast, and the hybrid feature method enhances the discriminability of the facial feature maps. Next, we also evaluate the proposed model on the CK+ dataset. As shown in Table 9, the results show that the proposed model is better than state-of-the-art models, including the traditional methods and deep learning-based methods. Note that this comparison contains more works,because the CK+ contains more high-quality data and many methods use it to evaluate performance.

TABLE 9. Results on the CK+ dataset.

Methods	**Types**	**Accuracy(%)**
Hsieh et al. [53]	Traditional	94.7
Mlakar et al. [54]	Traditional	95.64
Happy et al. [55]	Traditional	97.09
Siddiqiet al. [56]	Traditional	96.83
Uccar et al. [57]	Traditional	95.17
Aly et al. [58]	Deep learning	88.14
Rivera et al. [59]	Deep learning	91.51
Lopes et al. [60]	Deep learning	93.68
Zhang et al. [61]	Deep learning	95.12
Yang et al. [62]	Deep learning	97.02
Proposed	Deep learning	**97.2**

V. CONCLUSION

In this work, an end-to-end deep model is proposed to improve the face recognition rate. The proposed model consists of three stages, that is, data pre-processing, feature extraction, and classification. The data enhancement method is proposed to enhance the contrast of the raw input data. A hybrid feature representation method is proposed to model the enhanced data to obtain more discriminative features. A fusion deep model is designed as the classification system. In the experimental stage, the comparison between the model and other state-of-the-art models and its performance in practical applications are considered. Three benchmark datasets are utilized to evaluate model performance, including the AR face dataset, the FER2013 database,and the CK+ dataset. We achieve state-of-the-art recognition rates, that is, 98.6%,94.5%, and 97.2%, respectively. In future work, we will focus on the face alignment technology and propose a new deep model to improve the recognition rate under large-area occlusion of the face.

REFERENCES

[1] 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.

[2] 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.

[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] F. D. Guillen-Gamez, I. Garcia-Magarino, J. Bravo-Agapito, R. Lacuesta, and J. Lloret, "A proposal to improve the authentication process in m-health environments," *IEEE Access*, vol. 5, pp. 22530–22544, 2017.

[5] S. Soltanpour, B. Boufama, and Q. M. Jonathan Wu, "A survey of local feature methods for 3D face recognition," *Pattern Recognit.*, vol. 72, pp. 391–406, Dec. 2017.

[6] M. Chihaoui, A. Elkefi, W. Bellil, and C. Ben Amar, "A survey of 2D face recognition techniques," *Computers*, vol. 5, no. 4, p. 21, Sep. 2016.

[7] R. Raghavendra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *Comput. Surv.*, vol. 50, no. 1, pp. 1–37, Mar. 2017.

[8] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.

[9] G. Hu, F. Yan, C.-H. Chan, W. Deng, W. Christmas, J. Kittler, and N. M. Robertson, "Face recognition using a unified 3D morphable model," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 73–89.

[10] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[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] 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.

[13] 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.

[14] 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

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

[16] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 980–993, Mar. 2015.

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

[18] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," *IEEE Access*, vol. 7, pp. 23012–23026, 2019.

[19] 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.

[20] 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.

[21] Z. Ye, M. Wang, Z. Hu, and W. Liu, "An adaptive image enhancement technique by combining cuckoo search and particle swarm optimization algorithm," *Comput. Intell. Neurosci.*, vol. 2015, pp. 1–12, Jan. 2015.

[22] U. R. Acharya, Y. Hagiwara, J. E. W. Koh, J. H. Tan, S. V. Bhandary, A. K. Rao, and U. Raghavendra, "Automated screening tool for dry and wet age-related macular degeneration (ARMD) using pyramid of histogram of oriented gradients (PHOG) and nonlinear features," *J. Comput. Sci.*, vol. 20, pp. 41–51, May 2017.

[23] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *Proc. Face Gesture*, Mar. 2011, pp. 878–883.

[24] C. Turan and K.-M. Lam, "Histogram-based local descriptors for facial expression recognition (FER): A comprehensive study," *J. Vis. Commun. Image Represent.*, vol. 55, pp. 331–341, Aug. 2018.

[25] L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Local binary features for texture classification: Taxonomy and experimental study," *Pattern Recognit.*, vol. 62, pp. 135–160, Feb. 2017.

[26] J. Chen, V. M. Patel, L. Liu, V. Kellokumpu, G. Zhao, M. Pietikäinen, and R. Chellappa, "Robust local features for remote face recognition," *Image Vis. Comput.*, vol. 64, pp. 34–46, Aug. 2017.

[27] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-CELEB-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 87–102.

[28] D. Sáez Trigueros, L. Meng, and M. Hartnett, "Face recognition: From traditional to deep learning methods," 2018, *arXiv:1811.00116*. [Online]. Available: http://arxiv.org/abs/1811.00116

[29] U. Park, Y. Tong, and A. K. Jain, "Age-invariant face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 947–954, May 2010.

[30] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2691–2698.

[31] X. Xu, H. A. Le, P. Dou, Y. Wu, and I. A. Kakadiaris, "Evaluation of a 3D-aided pose invariant 2D face recognition system," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 446–455.

[32] M. Z. Al-Dabagh, M. Alhabib, and F. Al-Mukhtar, "Face recognition system based on kernel discriminant analysis, K-nearest neighbor and support vector machine," *Int. J. Res. Eng.*, vol. 5, no. 2, pp. 335–338, Mar. 2018.

[33] P. M. Kumar, U. Gandhi, R. Varatharajan, G. Manogaran, R. Jidhesh, and T. Vadivel, "Intelligent face recognition and navigation system using neural learning for smart security in Internet of Things," *Cluster Comput.*, vol. 22, no. S4, pp. 7733–7744, Jul. 2019.

[34] S. Karahan, M. Kilinc Yildirum, K. Kirtac, F. S. Rende, G. Butun, and H. K. Ekenel, "How image degradations affect deep CNN-based face recognition?" in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2016, pp. 1–5.

[35] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein CNN: Learning invariant features for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1761–1773, Jul. 2019.

[36] M. Coskun, A. Ucar, O. Yildirim, and Y. Demir, "Face recognition based on convolutional neural network," in *Proc. Int. Conf. Modern Electr. Energy Syst. (MEES)*, Nov. 2017, pp. 376–379.

[37] S. Banerjee and S. Das, "Mutual variation of information on transfer-CNN for face recognition with degraded probe samples," *Neurocomputing*, vol. 310, pp. 299–315, Oct. 2018.

[38] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Comput. Vis. Image Understand.*, vol. 189, Dec. 2019, Art. no. 102805.

[39] A. Sepas-Moghaddam, F. M. Pereira, and P. L. Correia, "Face recognition: A novel multi-level taxonomy based survey," *IET Biometrics*, vol. 9, no. 2, pp. 58–67, Mar. 2020.

[40] Y.-X. Yang, C. Wen, K. Xie, F.-Q. Wen, G.-Q. Sheng, and X.-G. Tang, "Face recognition using the SR-CNN model," *Sensors*, vol. 18, no. 12, p. 4237, Dec. 2018.

[41] S. Bhattacharjee, A. Mohammadi, and S. Marcel, "Spoofing deep face recognition with custom silicone masks," in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2018, pp. 1–7.

[42] H. Zhang, Z. Qu, L. Yuan, and G. Li, "A face recognition method based on LBP feature for CNN," in *Proc. IEEE 2nd Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Mar. 2017, pp. 544–547.

[43] M. Xi, L. Chen, D. Polajnar, and W. Tong, "Local binary pattern network: A deep learning approach for face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3224–3228.

[44] 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.

[45] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, and R. Chellappa, "An end-to-end system for unconstrained face verification with deep convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 118–126.

[46] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2016, pp. 1–8.

[47] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *J. Med. Imag.*, vol. 3, no. 3, Aug. 2016, Art. no. 034501.

[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[49] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual QA," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 361–369.

[50] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] A. Ramirez Rivera, J. Rojas Castillo, and O. Oksam 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.

[60] 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.

[61] 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.

[62] 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.

JUN LIU received the B.E. and M.E. degrees from the Shenyang University of Technology, in 1995 and 2000, respectively, and the 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 main research interests include intelligent sensors and detection technology, image and signal processing, and intelligent robots.

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 main research interests include network computing and artificial intelligence.

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 main research interests include the IoT technology and intelligent information processing.