IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 16, 2020, accepted March 3, 2020, date of publication March 6, 2020, date of current version March 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2978938

Similar Face Recognition Using the IE-CNN Model

AN-PING SONG[1], QIAN HU[1], XUE-HAI DING[1], XIN-YI DI[1], AND ZI-HENG SONG[2]
[1]School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
[2]College of Science, Purdue University, West Lafayette, IN 47907, USA

Corresponding author: An-Ping Song (apsong@shu.edu.cn)

ABSTRACT In the field of face recognition, similar face recognition is difficult due to the high degree of similarity of the face structure. The following two factors are needed to make progress in this field: (i) the availability of large scale similar face training datasets, and (ii) a fine-grained face recognition method. With the above factors fulfilled, we make two contributions. First, we show how a large scale similar face dataset (SFD) can be assembled by a combination of automation and human in the loop, and divide the dataset into five grades according to different degrees of similarity. Second, a new fine-grained face feature extraction method is proposed to solve this problem using the attention mechanism which combines the Internal Features and External Features. The Labeled Faces in the Wild (LFW) database, CASIA-WebFace and similar face dataset (SFD) were selected for experiments. It turns out that the true positive rate is improved by 1.94 - 5.66% and the recognition accuracy rate improved by 2.08 – 5.8% for the LFW and CASIA-WebFace database, respectively. Meanwhile for SFD, the recognition accuracy rate improved by 18.80 – 35.84%.

INDEX TERMS Face recognition, image databases, computer vision, machine learning, artificial neural networks.

I. INTRODUCTION

Recent studies have shown that deep neural networks perform well in face detection [1], [2], face alignment [3], and face verification [4], [5]. One of the most important ingredients to the success of such methods is the availability of large quantities of training data. For example, the most recent face recognition method of Google [6] was trained using 200 million images and 8 million unique identities. However, there are still some problems with the above method in similar face recognition. Table 1 shows the current popular public and non-public face recognition datasets. Obviously due to the lack of public similar face dataset in academia, it has become very difficult to research. Needless to say, building a dataset this large is beyond the capabilities of most international research groups, particularly in academia.

This paper has two contributions. The first one is to proposes a procedure to create a reasonably large similar face dataset (doi.org/10.6084/m9.figshare.11611071) while requiring only a limited amount of person-power for annotation. This procedure has been used to collect more than 30,000 pairs of similar faces which belong to different identities. Second, this paper introduces a model (IE-CNN) that enhances the internal and external features of the face, which can effectively improve the precision of face matching. The IE-CNN can effectively improve the true positive rate in the recognition task of similar face images. Meanwhile we propose a step-by-step training method to train the model, which reduces the time and hardware costs of model training under limited datasets.

TABLE 1. Dataset comparisons. Facebook and Google are not open to other scholars.

Dataset	Identities	Images
LFW [7]	5,749	13,233
WDRef [8]	2,995	99,773
CelebFaces [9]	10,177	202,599
CASIA-WebFace [10]	10,575	494,414
FaceBook [4]	4,030	4.4M
Google [6]	8M	200M

II. RELATED WORK

Influenced by the excellent performance of deep learning in the field of image classification, researchers often improve the model representation ability by repeatedly stacking convolution blocks when designing face recognition models [6]. Further study of these model structures reveals that

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

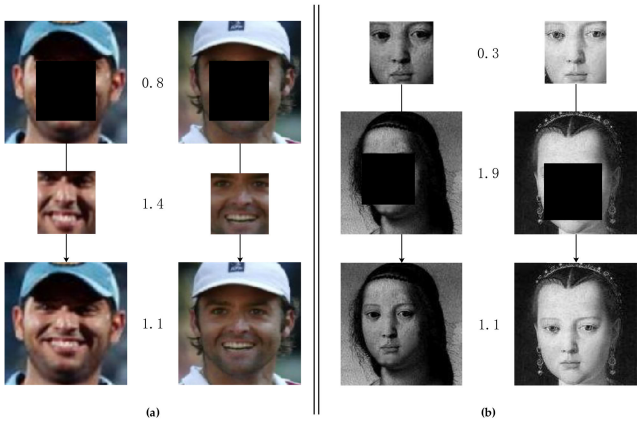


FIGURE 1. (a) Shows the importance of internal features and (b) shows the importance of external features in face recognition in different situations.

researchers design face recognition models in the same way as they design the general object classification models, which completely treat face images as general object images. Such methods would treat two similar face images from different person as the same category. In fact, human face features are composed of Internal Features (composed by eyes, nose and mouth) and External Features (located at head, chin, and ears) [11]. Unlike other objects such as cars or airplanes which have obvious differences in shape, color, volume and other characteristics, different face images of different people still have high structural similarities. The current face recognition algorithms show some quite interesting phenomena, which are manifested in the fact that the external features are unconsciously abandoned by the researchers in the face recognition algorithms [12], [13]. However, in the human brain's recognition of face images, internal and external features are treated as a whole in the relevant areas of the cerebral cortex(Fusiform Face Area, FFA; Occipital Face Area, OFA; Superior Temporal Sulcus, STS), rather than discarded or separated [14]. [15], [16] also states that the global processing of facial features generally has higher recognition efficiency compared to local processing. Fig 1 shows the importance of internal and external features in face recognition in different situations. The number represent the Euclidean distance between the two pictures. A distance of 0.0 means the faces are identical.

The above facts allow us to focus our research on the internal and external features of the human face. Reference [17] pointed out that the face features extracted in deep CNN are more robust. Similarly, this paper proposes a face feature enhancement model (IE-CNN) based on deep CNN. When humans recognize face images, they always focus on the various local features of the face image. The more difficult it is to distinguish face images, the more you should focus on the face feature details. Drawing on the process of human brain recognition, the soft attention mechanisms [18] are used to implement IE-CNN. [19] states that collecting large amounts of labelled face images is expensive. Generative adversarial networks(GANs) [20], [21] are a feasible solution to solve this problem. Such methods can be used to generate diverse images of the same subject. However, the purpose of this paper is to generate similar face images that do not belong to the same people.

III. DATASET COLLECTION

In this section we propose a multi-stage strategy to effectively collect the similar face dataset (SFD). Fig 2 shows some sample images of SFD and Table 2 shows the details of the SFD.

FIGURE 2. Sample images from the similar face dataset (SFD).

Stage 1. Collecting the Suitable Data Source: The first stage of establishing SFD is to obtain an available data source. Based on this data source, the prepared algorithm will be used to find pairs of similar faces images. In order to facilitate the follow-up work, we intend to perform this step on the existing public dataset (LFW and CASIA-WebFace). The LFW is a standard face recognition dataset, in which all face images contain positive faces without complex interference factors. The LFW dataset contains 13,233 images of human faces collected from the network. These face images belong to 5,749 people, of which 4,069 people have only one photo, which is very disadvantageous to training. Therefore, we introduce another public dataset CASIA-WebFace, which collects face images on the network through a semi-automatic approach. This dataset has a total of 10,575 people, 494,414 face images. Each person has at least two face images. It should be noted that there are many error labels and error images in above datasets since they are collected through network. All the images are $250 \times 250 \times 3$ and taken with different levels of light intensity, posture, and expression. After we manually filter some error label images such as that Andrew_Caldecott and Andrew_Gilligan are actually the same person, we merge them into a single dataset, in which all faces images are labeled.

Stage 2. Determining the Similarity Between Two Faces Images: Since the squared L2 distances in the embedding space directly correspond to face similarity, faces of the same person have small distances and faces of different person have large distances [6]. Therefore, the L2 distances is used to

measure the similarity between different face images. The embedding file of images collected by stage 1 is obtained through the pre-trained model [22].

Stage 3. Generating the Similar Face Dataset (SFD): Then for any vector in the embedding file, we find the other vector closest to it. If the face images represented by the two nearest vectors do not belong to the same person, we record the distance between them. If the distance is greater than 0 and less than or equal to 0.4, the face images represented by these two vectors are classified into Grade I. Different distances are divided into different grades, and the detailed division rules are shown in Table 2. Based on this, we collect a similar face dataset (SFD) which is divided into five grades.

TABLE 2. The details of the similar face dataset (SFD).

Grade	Pairs	Distances
I	760	$0 < d <= 0.4$
II	1,076	$0.4 < d <= 0.5$
III	1,781	$0.5 < d <= 0.6$
IV	3,198	$0.6 < d <= 0.7$
V	14,714	$0.7 < d <= 0.8$

The grades in Fig 2 from I to V indicate the similarity of the two face images from high to low. In SFD, each pair of pictures is similar but comes from different people. Similar faces tend to have similar facial features, which makes the model perform poorly on face-related tasks. By studying the face feature extraction problem of similar face images, the robustness of face recognition method can be effectively improved. Therefore, we believe that the proposed dataset (SFD) is a contribution to the research in the field of face recognition.

IV. NETWORK ARCHITECTURE AND TRAINING

A. IE-CNN MOUDLE

In the existing face matching methods [22], the feature extraction of the face image does not separate the internal and external features separately. These methods usually rely on more complex network structures to improve matching accuracy [23]–[25], rather than considering the internal and external features unique to the human face, which leads to a sharp drop in the accuracy of these methods under similar face images. Therefore, this paper proposes IE-CNN model which contains two modules to extract the unique internal features and external features on the face images to improve the effectiveness of face feature extraction. The schematic diagram of the entire algorithm is shown in Fig 3.

The algorithm is divided into two branches, one called IE-CNN branch and the other called trunk branch. Given the trunk branch output $T(x)$ with input x, the IE-CNN branch is to learn the same size mask $F(x)$ that softly weight output features $T(x)$. The output of algorithm $H(x)$ is:

$$H_{i,c}(x) = (1 + F_{i,c}(x)) \times T_{i,c}(x) \tag{1}$$

where i ranges over all spatial positions and $c \in \{1, \ldots, C\}$ is the index of the channel.

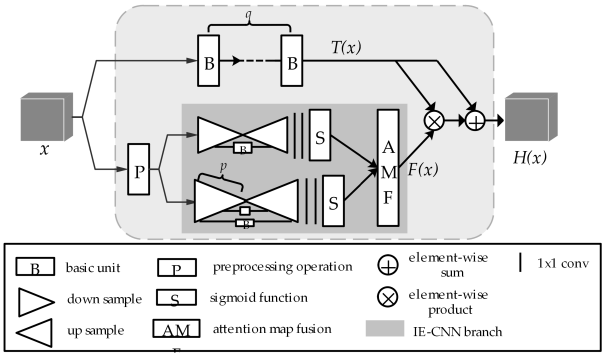


FIGURE 3. The entire algorithm contains two branches: the above is the trunk branch; the below is the IE-CNN branch.

The hyper-parameter p denotes the number of Basic Units in the IE-CNN branch. q denotes the number of Basic Units in trunk branch. In our experiments, we use the following hyper-parameters setting:$p = 3$, $q = 2$. The numbers of channels in the IE-CNN and corresponding trunk branches are the same. In this work, the pre-activation Residual Unit [26] is used as basic unit to construct the trunk branch. In IE-CNN branch, on the contrary, a custom large-size pre-activation Residual Unit is adopted as a basic unit. In order to make our model achieve better performance, Stochastic depth [27], Batch Normalization [28] and Dropout [29] exploit regularization for convergence and avoid overfitting and degradation. Fig 4 shows the structure of these two basic units.

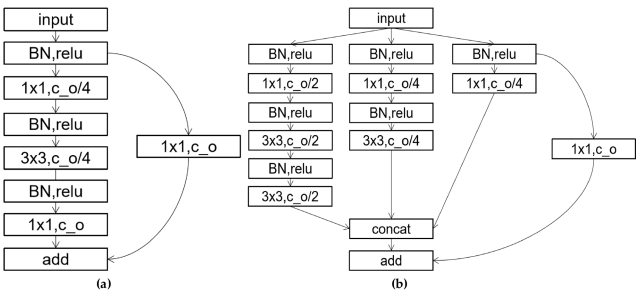


FIGURE 4. (a) Shows the structure of pre-activation residual unit and (b) shows the large-size one.

1) PRE-ACTIVATION RESIDUAL UNIT

The pre-activation Residual Unit is activated before each convolution in the residual branch. The matrix element addition is combined to satisfy the activation requirement and to eliminate the need for additional activation outside the branch. This type of method has a certain regularization effect and is easier to converge than the original residual unit.

In all experiments, the IE-CNN is trained by using Stochastic Gradient Descent (SGD) with standard backprop [30], [31] and AdaGrad [32]. As the model becomes more complex, the difficulty of training is getting higher, and the result is more likely to cause gradient dispersion, gradient explosion or other issues. In order to solve the above problem, as shown

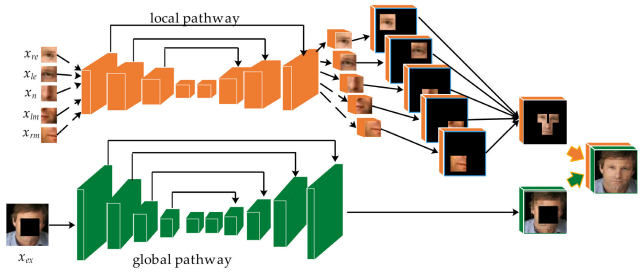


FIGURE 5. Visualization of internal details of IE-CNN branch.

in Fig 4, the batch normalization (BN) layer which is a normalization process (normalized to a mean of 0, a variance of 1) is added before each layer of convolution. In the pre-activation Residual Unit, the ReLU [33] is used as the activation function.

In pre-activation Residual Unit, both the BN and the activation function are operated before the convolution operation, which can increase the speed of training and alleviate the problem of internal covariate shift. In addition, in order to make the model deeper and the amount of parameters not be increased too much, several $1 \times 1 \times d$ convolution layers are added in model inspired by the work of [34]. As shown in Fig 4, this can ensure the model representation ability while reducing the amount of parameters, which makes the model more complex to design and improve generalization ability.

2) ARCHITECTURE

The purpose of IE-CNN branch is to extract the features of Internal Features (composed by eyes,nose and mouth) and External Features (located at head, chin, and ears). The bottom-up top-down structure [35]–[38] is used for the internal design of IE-CNN. The bottom-up top-down structure mimics the fast feedforward and feedback attention process. Due to the particularity of each local area, a five-way parallel structure is adopted, and each branch parameter is not shared. For the five local feature maps, a position aggregation strategy is used for feature fusion.

The details of IE-CNN branch are visualized in the Fig 5. The local pathway is used to extract the local information of internal features and to improve the performance of fine-grained recognition. The global pathway is used to capture the global information of the external features, ensuring the general face matching accuracy.

It is fact that making attention change adaptively with features without additional constraint leads to better performance. And the attention provided by local pathway should change adaptively with global pathway features. Therefore, the mixed attention f_1 which uses simple sigmoid for each channel and spatial position without additional restriction is adopted. And its mathematical representation is as follows:

$$f_1(x_{i,c}) = \frac{1}{1 + exp(-x_{i,c})} \tag{2}$$

where i ranges over all spatial positions and c ranges over all channels, x_i denotes the feature vector at the ith spatial position.

TABLE 3. Structure of the local pathway.

Layer	Input	Filter Size	Output Size
conv1	x_{in}	$3 \times 3/1$	$w \times h \times 16$
basic unit1	-	$\begin{pmatrix} 1\times1, 8 \\ 3\times3, 8 \\ 1\times1, 32 \end{pmatrix} /2$	$w/2 \times h/2 \times 32$
basic unit2	-	$\begin{pmatrix} 1\times1, 16 \\ 3\times3, 16 \\ 1\times1, 64 \end{pmatrix} /2$	$w/4 \times h/4 \times 64$
basic unit3	-	$\begin{pmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{pmatrix} /2$	$w/8 \times h/8 \times 256$
up sample1	basic unit3	-	$w/4 \times h/4 \times 64$
up sample2	basic unit2	-	$w/2 \times h/2 \times 32$
up sample3	basic unit1	-	$w \times h \times 16$
conv2	conv1	$3 \times 3/1$	$w \times h \times 3$

TABLE 4. Structure of the global pathway.

Layer	Input	Filter Size	Output Size
conv1	x_{ex}	$3 \times 3/1$	$160 \times 160 \times 16$
basic unit1	-	-	$80 \times 80 \times 32$
basic unit2	-	-	$40 \times 40 \times 64$
basic unit3	-	-	$20 \times 20 \times 256$
basic unit4	-	-	$10 \times 10 \times 512$
up sample1	basic unit4	-	$20 \times 20 \times 256$
up sample2	basic unit3	-	$40 \times 40 \times 64$
up sample3	basic unit2	-	$80 \times 80 \times 32$
up sample4	basic unit1		$160 \times 160 \times 16$
conv2	conv1	$3 \times 3/1$	$160 \times 160 \times 16$
conv3	-	$3 \times 3/1$	$160 \times 160 \times 3$

To effectively integrate the information from the global and local pathways, an intuitive method is adopted for attention map fusion. As shown in Fig 5, the output attention tensors (multiple attention maps) of five local pathways are fused to one single attention tensor that is of the same spatial resolution as the global attention tensor. Specifically, each feature tensor is put at a ''template landmark location'', and then a max-out fusing strategy is introduced to reduce the stitching artifacts on the overlapping areas. Then, the attention tensor from each pathway is simply concatenated to produce a fused attention tensor and its channel count is twice the input.

In IE-CNN branch, the pre-activation Residual Unit is used as the basic unit in the down sample operation, which can better extract the information contained in the internal feature image of the face. The bilinear interpolation algorithm is used to implement the up sample operation.

Table 3 shows the structures of the local pathway and Table 4 shows the structures of the global pathway. The w and h denote the width and the height of the cropped patch. For the patches of the two eyes, the w and h are set as 40; for the patch of the nose, the w is set as 32 and the h is set as 40; for the patch of the mouth, the w and h are set as 32 and 40 respectively.

B. FRAMEWORK OF THE PROPOSED TRAINING METHOD

Most of the previous face recognition algorithms use a classification layer [17] to achieve face matching after extracting

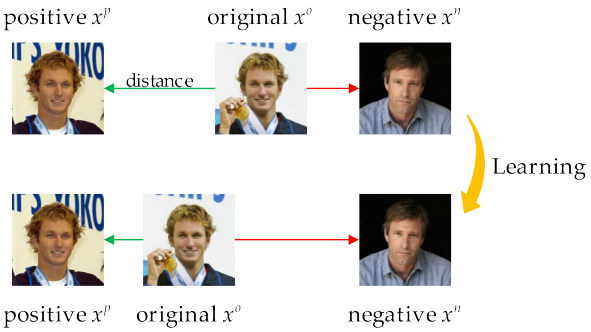


FIGURE 6. The Triplet Loss minimizes the distance between an original and a positive, both of which have the same identity, and maximizes the distance between the original and a negative of a different identity.

face features from the convolutional layer. The biggest disadvantage of this approach is that the generalization ability of the model is very weak. Inspired by [6], our method maps a face image one-to-one to a Euclidean embedding by combining IE-CNN with a pre-trained convolutional layer. The squared L2 distances in the embedding space directly correspond to face similarity. This means faces of the same person have small distances and faces of different people have large distances. Therefore, our training method uses a triplet based loss function based on LMNN [39]. Choosing the suitable triplets turns out to be very important for achieving good performance. There is a picture of the original face x^o (original), other pictures of the same person x^p (positive), any pictures of other different people x^n (negative). Our target is to make the face image x^o of a specific person closer to all other images x^p of the same person than it is to any image x^n of any other person. Its schematic diagram is shown in Fig 6.

The aim is:

$$\left\|f(x^o)-f(x^p)\right\|_2^2+\theta<\left\|f(x^o)-f(x^n)\right\|_2^2, \quad \forall(f(x^o),f(x^p),f(x^n))\in T. \tag{3}$$

where θ is a margin that is enforced between positive and negative pairs. T is the set of all possible triplets in the training set and has cardinality N.

The loss that is being minimized is then

$$L=\sum_i^N\left[\left\|f(x_i^o)-f(x_i^p)\right\|_2^2-\left\|f(x_i^o)-f(x_i^n)\right\|_2^2+\theta\right]_+ \tag{4}$$

The more ineffective triplets passed through the network, the lower efficiency this training will be. Therefore, a small mini-batch training strategy is used to prevent excessive selection of triplets. Meanwhile, the model would prefer using small mini-batches as these tend to improve convergence during Stochastic Gradient Descent (SGD) [40]. In our experiments, to ensure that the selected positive distance is valid, the number of images per person is set to 30 in each mini-batch. Additionally, randomly sampled negative faces are added to each mini-batch. But this only solves the problem of the sample size. Our next focus is on how to select the most effective triplets that violate the triplet constraint in Eq. (3). This means that for each x^o, the farthest sample x^p and the nearest sample x^n are needed:

$$\begin{aligned} &argmax_{x^p}\left\|f(x^o)-f(x^p)\right\|_2^2, \\ &argmin_{x^n}\left\|f(x^o)-f(x^n)\right\|_2^2 \end{aligned} \tag{5}$$

instead of picking the farthest positive, all original-positive pairs in a mini-batch are used while still selecting the nearest negatives. Because this method is more stable and converges slightly faster at the beginning of training. Selecting the nearest negatives can in practice lead to bad local minima early on in training, specifically it can result in a collapsed model. According to Eq. (3), as long as the negatives lie inside the margin θ and satisfy the following formula:

$$\left\|f(x^o)-f(x^p)\right\|_2^2<\left\|f(x^o)-f(x^n)\right\|_2^2 \tag{6}$$

those negatives are further away from the anchor than the positive exemplar, but still effective because the squared distance is close to the original-positive distance. And those negatives can help mitigate the above problem. As shown in the Fig 7, the optimization object selected is shown as the shaded area of the figure.

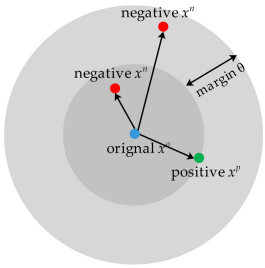


FIGURE 7. The data in the dark gray area of the figure is very suitable for optimization, but the amount of data in such area is obviously not enough, so the data of the outer light gray area are still selected.

According to the above method, the appropriate triplets can be selected for training and improving the training quality. In our experiment, the batch-size is set to 90, the number of people per batch is set to 45, and the number of images per person is 30. So for each batch of training, this paper uses around 1350 exemplars.

An original face image will be pre-processed by a pre-trained cascaded CNNs [41] in Stage I, then enhanced by internal and external features through Stage II which is IE-CNN, and finally mapped to high-dimensional space through Stage III. The whole training method is visualized in Fig 8.

In summary, for the research on similar face recognition, this paper first collected a large-scale similar face dataset (SFD), and then proposed a fine-grained face feature extraction method (IE-CNN). In next section, multiple sets of comparative experiments will be used to verify the effectiveness and accuracy.

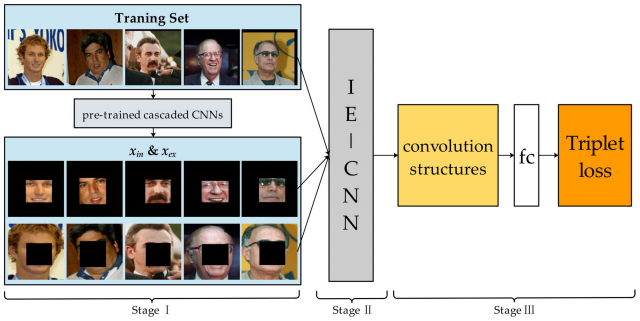


FIGURE 8. The schematic diagram of the whole training method.

TABLE 5. The test accuracy (%) on LFW and CASIA-WebFace of IE-CNN branch with different fusion strategy.

Fusion Strategy	Operate type	LFW	CASIA-WebFace
f_1	concatenate	**96.23**	**84.35**
f_2	element-wise sum	94.80	83.81
f_3	only local pathway	66.84	52.33
f_4	only global pathway	75.33	63.28

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we first experimented with the difference in the feature map fusion method in IE-CNN branch. Then we evaluated the effectiveness of the proposed method on a series of benchmark datasets including LFW, CASIA-WebFace. Meanwhile the experimental results are compared with other methods on Similar Face Dataset (SFD).

A. EXPERIMENT ON FEATURE MAP FUSION IN IE-CNN BRANCH

As introduced in section Architecture, given the input that comes from a pixel space of size $W \times H \times C$ with C color channel, the outputs size of local pathway and global pathway are all $W \times H \times C$. In IE-CNN branch, an output of the same size as input as the final attention map is needed. So for the output of these two pathway there are two methods for fusing feature maps.

The first method called f_1 is to directly concatenate the output of local pathway and global pathway. Then a bottleneck layer is used to adjust the number of channels, and finally get the same size tensor as the input $W \times H \times C$. The second method is to directly add the outputs of two pathway by element-wise sum, which is called f_2. We also explored the recognition accuracy in the case of local pathway only called f_3 or global pathway only called f_4. Table 5 shows the test accuracy (%) on LFW and CASIA-WebFace of IE-CNN branch with different feature maps fusion operation.

Based on the above experiments, f_1 performs best in two different datasets. When a face image is fully integrated with internal and external features which is called f_1, the increase in recognition accuracy is an inevitable trend. In this paper, the f_1 method is used to integrate internal and external features in IE-CNN branch as shown in Fig 5. However, compared with other methods, f_1 has a disadvantage that the parameters of the model will be increased, which will undoubtedly bring more time and hardware costs.

B. EXPERIMENTAL EVALUATION INDEX

There are two types of evaluation methods for our experimental results, one is judged by the currently widely used comparison method. The formulas are given below.

$$TPR = \frac{TP}{TP + FN}, \tag{7}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}. \tag{8}$$

The second custom evaluation method called *Top*1&*Top*5 *precision* was also used in our experiments. Traversing the N samples, if the total number of *Top*1 hits is N_1 and the total number of *Top*5 hits is N_5, *precision* is calculated as follows:

$$Precision_{top1} = \frac{N_1}{N}, \tag{9}$$

$$Precision_{top5} = \frac{N_5}{N}. \tag{10}$$

C. EXPERIMENTS ON LFW AND CASIA-WebFace

In this experiments, Inception-Resnet-v2 [22] was used as a comparative experiment. It is a very authoritative model in the field of face recognition which is called *v*2 in our experiments. In order to eliminate the influence of interference factors, the paper adopted the same training environment to retrain v2 model and our method by using the triplet loss constraint. The maximum number of iterations for training was set to 500, and the learning rate was attenuated from 0.1 to 0.001 using the piecewise constant attenuation method. Throughout the training process, we adopted the 10-fold cross validation method to realize the division of training sets and test sets. Fig 9 shows the differences in training between these two models.

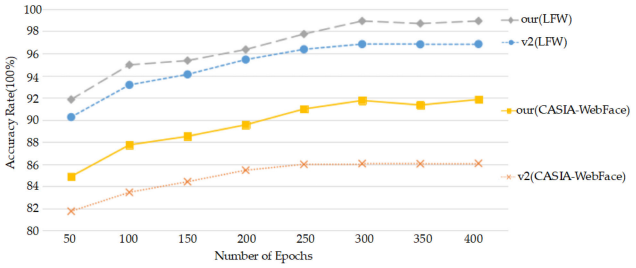


FIGURE 9. Comparison of accuracy for different training epochs.

As can be seen from the Fig 9, when the number of epochs is around 300, the model was already converged. In order to save training costs, in this paper, the training epoch of the model used was set to 300. The v2 and our model of the subsequent experiments in this paper were all from this training on CASIA-WebFace.

Meanwhile, we performed face matching experiments about the two methods on the two different datasets. In the LFW, we randomly took 6,000 pairs of samples, including

3,000 positives and 3,000 negatives. And in the CASIS-WebFace, we randomly took 3,0000 pairs samples, which contains half of the positive examples. The experimental results were averaged over 10 experiments. The face images applied to the two different methods are the same and they all passed the alignment pretreatment. The matching effects of two models for the two face databases are shown in Fig 10 (*acc* represents the recognition accuracy rate, and *tpr* is the true positive rate).

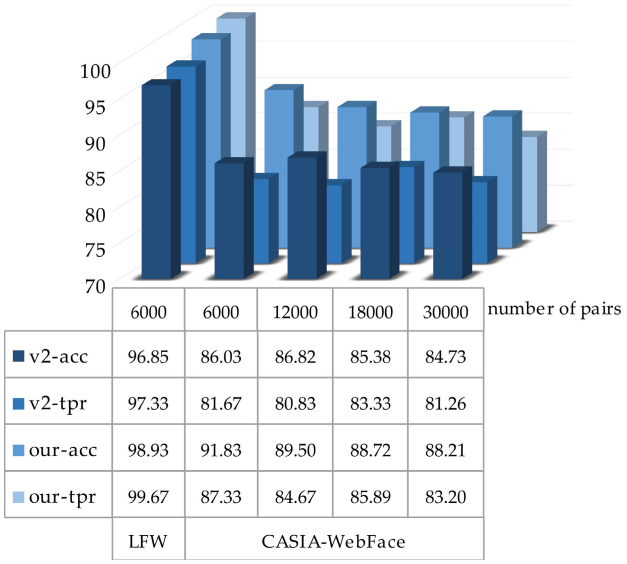


FIGURE 10. Comparison between v2 and our methods (LFW and CASIA-WebFace).

The matching effects of two methods for the LFW and CASIA-WebFace are shown in Fig 10. For our improved face matching method which combines the internal and external features of the face, the recognition accuracy rate improved by 2.08% and the true positive rate improved by 2.34% in LFW. In the CASIA-WebFace, the face matching accuracy in different scale of samples improved by 2.67 - 5.8% and the true positive rate improved by 1.94 - 5.66%.

1) *TOP*1&*TOP*5 *PRECISION* EXPERIMENTS UNDER DIFFERENT GLOBAL SCOPES

In this experiment, some interference data needs to be excluded in advance. In the datasets used above, many individuals have only one face images. These data obviously do not have *top*1 and *top*5 hit ratios, which are called interference data. In LFW, 1,680 individuals have more than 2 face images, the remaining 4,069 people only have one, and there are 9,164 valid face images in total. In CASIA-WebFace, everyone has at least two face images. Therefore, their effective number remains the same. Table 6 shows the *Top*1&*Top*5 *Precision* experiments results of the two methods.

Based on the above experiments, compared with the traditional method, the facial images matching method adding

TABLE 6. The *Top*1&*Top*5 *Precision* (%) on LFW and CASIA-WebFace between two methods.

Method	LFW		CASIA-WebFace	
	$Top5-acc$	$Top1-acc$	$Top5-acc$	$Top1-acc$
v2	96.96	93.63	85.97	78.24
our	99.20	98.86	93.31	84.92

IE-CNN model is greatly improved. For the face images that were not recognized by our method in the above experiments, most of them have incorrect tags. For example, two photos which labeled Andrew_Caldecott and Andrew_Gilligan are actually the same person. Considering these inevitable factors, the recognition accuracy rate of *Top*5 improved by 2.24% and *Top*1 improved by 5.23% in LFW. For the CASIA-WebFace which contains about 37 times more face images than the previous datasets, the recognition accuracy rate of *Top*5 improved by 7.34% and *Top*1 improved by 6.68%. In the actual application of face recognition, sometimes it is necessary to implement a search for N to N such as the police arrest the suspect rather than giving you two faces images to tell me whether it is the same person. Therefore, the comprehensive performance of the model is also one of the directions of our research.

D. EXPERIMENTS ON SIMILAR FACE DATASET (SFD)

In SFD, the labels of the two images in each pair do not belong to the same person. So in this subsection experiment, we added some face images which are of the same identity in each grade. The similar face images of each grade were tested separately which could specifically explore the face recognition problem under different similarities rather than as a whole. In this section, we compared the DeepID2+ model [17], v2 and our method. The experimental results were averaged over 10 experiments. Table 7 shows the results of this experiment (*acc* represents the recognition accuracy rate, and *tpr* is the true positive rate).

TABLE 7. Comparison about v2, DeepID and our methods from grade I to V (SFD).

Method	Similar Face Dataset (SFD)				
	I	II	III	IV	V
[17]-acc	49.73	59.75	58.64	64.37	66.33
[17]-tpr	71.77	78.08	76.28	81.77	83.17
v2-acc	42.81	54.75	55.64	66.97	68.22
v2-tpr	63.33	70.00	65.33	71.67	68.08
our-acc	78.29	83.19	82.31	86.11	86.40
our-tpr	79.17	83.33	79.67	85.67	84.67

According to the experimental results, the IE-CNN method presented in this paper obtained the optimal face recognition performance compared with [17] and [22]. As the similarity level goes from high to low (from I to V), the accuracy was improved, despite the number of face images increased. The more similar the face image, the more difficult it is to identify. As we can see from the Table 7, in grade I 1200 pairs of

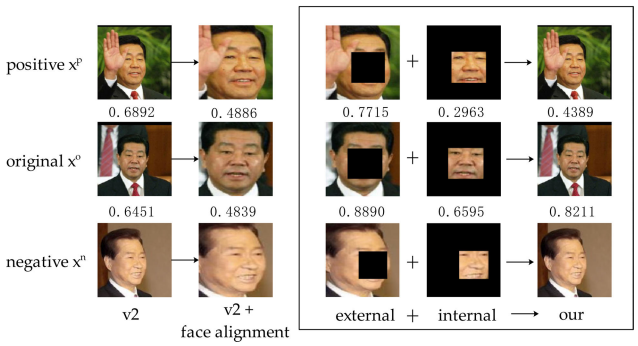


FIGURE 11. Comparison between v2, v2 + face alignment and our methods.

samples (600 pairs negatives were taken from SFD) were randomly sampled for testing, and the recognition accuracy of v2 was less than half. And in the grade II, III, IV and V, we randomly sampled 1800, 3000, 6000, 24000 pairs (half of them from SFD) respectively. The highest recognition accuracy of v2 is only 68.22% in grade V, which shows that the current very popular v2 does not perform well on SFD. As for DeepID2+ [17], it shows some robustness on SFD, and its tpr is higher than v2. But these two models both do not perform well on the SFD. Correspondingly, our method performed well with five grades, which proves that the proposed IE-CNN is very effective for similar face recognition. For the grade I, the recognition accuracy rate improved by 35.84% and the true positive rate improved 15.84%. Meanwhile as the picture similarity in the dataset decreases from II to V, the recognition accuracy rate improved by 18.80 – 28.44%. We fully considered the complementary information of internal features and external features which can solve the problems in similar face recognition that other methods cannot avoid rather than only original face image information. We fused complementary features, and the results of the experiment show that the fusion was a success.

E. SUPPLEMENTARY EXPERIMENT ON SFD

In this experiment, we verified the advantages of dividing the facial features into internal and external features in our method. One example is selected to display as shown in Fig 11.

The number in the middle of the two face images in the figure represents the squared L2 distances obtained by the corresponding method. The smaller the number is, the more likely it is that the method considers the two face images to belong to the same person.

According to the Fig 11, the v2 model did not distinguish the three face images very well. It would consider such three face images belongs to one person according to the results in the first column of the figure. Obviously, when trying to align the image by [41], the v2 model more likely thought that the three images belong to one person. In SFD, such an error phenomenon will be more obvious, which will make it more difficult to identify similar faces. But in our method, we use the external features of the face to distinguish those faces that are similar in internal features, and further improve the recognition accuracy by combining internal features to achieve fine-grained face recognition. The results of the last column in the figure show that our method can very well distinguish face images that do not belong to one person.

VI. CONCLUSION AND FUTURE WORK

This paper collected a new similar face datasets (SFD) and introduced a robust face recognition method combining internal features and external features of face. By combining the external feature and the internal feature, our method not only ensures that face matching accuracy is improved, but also achieves fine-grained recognition effect. Our method improves the face features extraction effectiveness of the traditional face recognition model while maintaining the advantages of the original deep CNN model, which tackles the recognition accuracy problem caused by the traditional methods when there are very similar face images here. The higher the similarity between the two face images is, the lower the recognition accuracy is, which is inevitable and will be further studied in the future.

The biggest drawback after adding internal and external features to assist the fine-grained face recognition is that it will bring additional parameters to the model and increase the difficulty of training. Therefore, the method of training model proposed in this paper is an end-to-end training mode, which greatly reduces the difficulty of model training. However, there are still many problems here. How to improve the efficiency of model training will also be one of the focuses of follow-up research.

REFERENCES

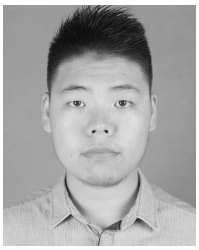
[1] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp. 42–50, Jul. 2018.
[2] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu, "Scale-aware face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6186–6195.
[3] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 88–97.
[4] 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.
[5] J. Li, T. Qiu, C. Wen, K. Xie, and F.-Q. Wen, "Robust face recognition using the deep C2D-CNN model based on decision-level fusion," *Sensors*, vol. 18, no. 7, p. 2080, Jun. 2018.
[6] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[7] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, 2008.
[8] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 566–579.
[9] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1891–1898.

[10] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: http://arxiv.org/abs/1411.7923
[11] H. D. Ellis, J. W. Shepherd, and G. M. Davies, "Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition," *Perception*, vol. 8, no. 4, pp. 431–439, Jun. 2016.
[12] T. F. Cootes and J. J. Gareth, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.
[13] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3539–3545.
[14] J. Davies-Thompson, A. Kingstone, A. W. Young, and T. J. Andrews, "Internal and external features of the face are represented holistically in face-selective regions of visual cortex," *J. Vis.*, vol. 10, no. 7, p. 674, Aug. 2010.
[15] A. W. Young, D. Hellawell, and D. C. Hay, "Configurational information in face perception," *Perception*, vol. 42, no. 11, pp. 1166–1178, Jan. 2013.
[16] P. Sinha and T. Poggio, "I think I know that face ldots," *Nature*, vol. 384, no. 6608, p. 404, 1996.
[17] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2892–2900.
[18] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[19] D. S. 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
[20] H. Ding, K. Sricharan, and R. Chellappa, "ExprGAN: Facial expression editing with controllable expression intensity," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.
[21] C. Donahue, Z. C. Lipton, A. Balsubramani, and J. McAuley, "Semantically decomposing the latent spaces of generative adversarial networks," 2017, *arXiv:1705.07904*. [Online]. Available: http://arxiv.org/abs/1705.07904
[22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[24] 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.
[25] R. K. Srivastava and K. G. J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 630–645.
[27] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 646–661.
[28] 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
[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[30] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.
[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
[32] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Feb. 2011.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[34] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400
[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[36] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.
[37] R. Mesbah, B. McCane, and S. Mills, "Deep convolutional encoder-decoder for myelin and axon segmentation," in *Proc. Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, Nov. 2016, pp. 1–6.
[38] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.
[39] K. Q. Weinberger and K. L. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.* vol. 10, pp. 207–244, Feb. 2009.
[40] D. R. Wilson and T. R. Martinez, "The general inefficiency of batch training for gradient descent learning," *Neural Netw.*, vol. 16, no. 10, pp. 1429–1451, Dec. 2003.
[41] 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.

AN-PING SONG was born in Shanghai, China, in 1977. He received the B.S., M.S., and Ph.D. degrees in computer application from Shanghai University, Shanghai, in 2009. Since 2009, he has been a Professor with the Computer Engineering and Science Department, Shanghai University. He is the author of one book and more than 20 articles. His research interests include medical image processing and algorithm, bioinformatics, database application, brain-computer interface, and parallel computing.

QIAN HU was born in Hubei, China, in 1995. He received the B.S. degree in computer science from Shanghai University, in 2016, where he is currently pursuing the M.S. degree in computer science. He is mainly engaged in the research of computer vision and image processing.

XUE-HAI DING was born in Ningbo, China, in 1981. He received the B.S. degree in information management and system from Ningbo University, Zhejiang, China, in 2004. He is currently pursuing the M.S. degree in computer science and technology with Shanghai University, Shanghai, China, in 2015. Since 2015, he has been an Experimentalist with the Computer Engineering and Science Department, Shanghai University. His research interests include data mining, brain-computer interface, deep learning, and data visualization analysis.



XIN-YI DI was born in Shanghai, China, in 1995. He received the B.S. degree in computer science from Shanghai University, in 2017, where he is currently pursuing the M.S. degree in computer science. His academic interests mainly include the intersection of computer vision, and machine learning; this includes but is not limited to graphics, 3D vision, and computational geometry.

ZI-HENG SONG was born in Shanghai, China, in 1998. He is currently pursuing the B.S. degree in computer science with Purdue University. His main research fields include image processing, data mining, and machine learning.