Assistant human facial recognition based on convolution feature-map

Jianlan Guo¹, Yuqiang Chen¹ *, Jiaming Li¹ and Jiabao Xie¹

¹Department of Computer Engineering, Dongguan Polytechnic, Dongguan, Guangdong, 523808, China

*Corresponding author’s e-mail: chenyuqiang@126.com

Abstract. Face recognition is a very important research direction in the field of image processing. In daily life, there are various problems that require face recognition everywhere, but how to accurately identify faces from images is a very important and urgent problem. To this end, we propose an assistant face recognition method that can effectively improve the recognition accuracy. The face recognition method based on deep learning will convolve the input image into five different convolution kernels to obtain the convolution feature-map of each layer. At the end, the corresponding description vector of the input image is generated by using the two layers fully connected layer and the pooling layer. Therefore, in the process of face recognition, it is only necessary to calculate the similarity between the description vectors of the two frames of images. Because of its recognition accuracy and room for improvement, we use the features of the fifth-level convolution feature-map, combined with the image description vector finally generated by the deep learning method, can significantly improve the recognition accuracy in the face recognition process. Through a large number of experiments, the method can be used to assist the image recognition problem based on the deep learning method, which can effectively improve the recognition accuracy.

1. Introduction

Face recognition and face authentication are very important and long-standing problems in the field of computer vision. The workers and enterprises in the field of vision related to the investment of a lot of energy and financial resources to improve the accuracy and efficiency of recognition. In 1991, m Turk and a Pentland and other scholars proposed the eigenface method (eigenface) [1], which uses PCA (principal component analysis) method to obtain the description vector of the input face image and uses the similarity between the two frames to calculate the similarity of the description vector to carry out face recognition. However, this method is greatly affected by the illumination of the image and its application range is very limited. In 1997, scholars such as PN belhumeur and JP hespanha proposed the classic face recognition algorithm Fisher face [2]. This method is based on linear discriminant analysis, which is similar to the principal component analysis used by M Turk and a Pentland scholar, so it has no Lu baton for illumination. In 2004, t Ahonen, a Hadid and other scholars proposed to use local binary patterns (LBP) [3] to extract local features as the basis for face recognition. But the advantage of LBP is that it is insensitive to light and cannot solve the problem of posture and expression. The intuitive image method cannot achieve good face recognition performance. Therefore, the feature descriptor sifts [4] generated by local information is introduced into the field of face recognition. In 2015, D Li and KM Lam proposed a feature descriptor pore sift [5] [6] [7]. Based on the skin pore transformation. After analyzing the features of face skin, the SIFT feature descriptor is modified, which
improves the number of key points detections. The nearest neighbour matching method is used to calculate the feature descriptor matching in two frame face images. The number of descriptors of the matching is used as the basis of face discrimination. However, the descriptor based on local information cannot fully excavate the information of face image, but the local information cannot fully excavate the information of face image and the calculation time is long. In recent years, with the increase of data and the upgrading of hardware computing capability, a large number of deep learning methods have been proposed, Geoff in 2006, Hinton proposed that the multi hidden layer neural network has more excellent characteristic learning ability [8]. Its complexity in the training process can be alleviated by layer-by-layer initialization, which solves the problem that deep learning cannot be applied to real life for a long time and promotes the rapid development of deep learning. Based on the work of 2006, Geoff Hinton proposed a convolutional neural network alexnet [9] in 2012, which won the championship in Imagenet competition, and defeated many recognition algorithms based on traditional methods. But it has not been invariant to face deformation and rotation, so it has not achieved good performance in face recognition. In 2014, K simonyan and a zisserman and other scholars improved alexnet and put forward a deeper deep learning network VggNet [10], but there are still the same defects as alexnet, and they have not achieved better performance in the field of face recognition. In 2015, park hi o m and a zisserman of Oxford University proposed VggFace [11], which is a convolutional network that can transform the input image into feature description vector. It uses this network to do face recognition, which has good performance in the open face data set LFW and YTF. Then, Chang Xing Ding and other scholars proposed a multi-directional face recognition algorithm [13] and a video sequence-based face recognition algorithm [14], while L. best Rowden and other scholars have made interesting and in-depth research on the vertical application of automatic face recognition.

2. Face recognition algorithm based on feature matching of VggFace convolutional neural network

Based on SIFT [4] and other descriptors using local information cannot fully mine all the information of the face image, while the deep learning method based on convolutional neural network can use the information of the whole face image to generate the image descriptor and do feature matching based on the descriptor to realize face recognition. However, the recognition accuracy of the current mainstream convolutional neural network VggFace [11] in the field of face recognition still has a large space to improve. We use the convolutional feature map information to carry out secondary feature matching, which plays an auxiliary recognition function in the face recognition process.

2.1. VggFace Model

The description vector used in this system and the convolution vector graph used in auxiliary recognition are generated by convolution neural network VggFace [11]. In order to clarify the discussion, this paper first introduces VggFace.

VggFace is improved from alexnet [9] and VggNet [10]. However, alexnet mainly consists of five convolution layers, three fully connected layers and three Max pooling layers, with a total of 11 layers, as shown in Figure 1.

![Figure 1. The specific process of alexnet convolution neural network.](image-url)
VggNet [10] modified the convolution layer [9] of the above Alex net, added multiple convolution layers of $3 \times 3$ after the second and third convolution layers of Alex net, and set all convolution cores to be $3 \times 3$, unlike Alex net, each convolution core is different. The unified convolution kernel size improves the training convergence speed. In addition, the convolution kernel of $3 \times 3$ is the smallest convolution kernel that can capture the concepts of left, right, top, bottom and center. Therefore, the $3 \times 3$ convolution kernel enables VggNet to discover more information of the image than Alex net. VggNet improves the recognition performance by increasing the number of convolution layers. The specific configuration is shown in Figure 2.

| ConvNet Configuration |
|-----------------------|
| A | A-LRN | B | C | D | E |
| 11 weight layers | 11 weight layers | 13 weight layers | 16 weight layers | 16 weight layers | 19 weight layers |
| input (224 x 224 RGB image) | | | | | |
| conv3-64 | conv3-64 | conv3-64 | conv3-64 | conv3-64 | conv3-64 |
| LRN | conv3-64 | conv3-64 | conv3-64 | conv3-64 | conv3-64 |
| maxpool | | | | | |
| conv3-128 | conv3-128 | conv3-128 | conv3-128 | conv3-128 | conv3-128 |
| conv3-128 | conv3-128 | conv3-128 | conv3-128 | conv3-128 | conv3-128 |
| maxpool | | | | | |
| conv3-256 | conv3-256 | conv3-256 | conv3-256 | conv3-256 | conv3-256 |
| conv3-256 | conv3-256 | conv3-256 | conv3-256 | conv3-256 | conv3-256 |
| maxpool | | | | | |
| conv3-512 | conv3-512 | conv3-512 | conv3-512 | conv3-512 | conv3-512 |
| conv3-512 | conv3-512 | conv3-512 | conv3-512 | conv3-512 | conv3-512 |
| maxpool | | | | | |
| conv3-512 | conv3-512 | conv3-512 | conv3-512 | conv3-512 | conv3-512 |
| conv3-512 | conv3-512 | conv3-512 | conv3-512 | conv3-512 | conv3-512 |
| maxpool | | | | | |
| FC-4096 | FC-4096 | FC-1000 | | | |
| soft-max | | | | | |

**Figure 2.** Experimental configuration of convolution layers used by VggNet.

Literature [10] points out that when the convolution layer is increased to 19 layers, the recognition performance of VggNet can only be similar to that of 16 layers, but it will consume more computing resources and time. Therefore, the literature recommends 16 layers of VggNet, and this paper also uses VggNet 16 to generate the description vector of face image. VggFace [11] is based on VggNet [10], changing the full connection layer of the last layer from the original 1000 to 2622. The specific configuration is shown in Figure 3.

**Figure 3.** VggFace convolution neural network structure.
The input layer of VggFace requires the image size to be $224 \times 224 \times 3$, so the input image needs to be scaled and cropped to meet the input requirements. After the operation of convolution layer and full connection layer, the $1 \times 1 \times 2622$-dimensional description vector is finally output, as shown in Figure 4.

![Figure 4. Process of VggFace converting input image into description vector.](image)

2.2. VggFace describes vector matching

Feature descriptor matching is a traditional research problem. Using feature descriptor for face recognition is to encode all the information of face image into feature descriptor. When two frames of face image are matched, only the similarity between two feature descriptors needs to be calculated. In this paper, the nearest neighbour ratio method is adopted. Since the description vector output by VggFace is an $1 \times 1 \times 2622$-dimensional vector, the cosine value of the angle $\theta$ between the description vectors of two frames of images is calculated as the matching score, as shown in equation (1).

$$v_{gg\_scores} = \cos \theta = \frac{v_{gg1} \cdot v_{gg2}}{|v_{gg1} \cdot v_{gg2}|}$$  \hspace{1cm} (1)

And compared with the pre-set threshold $\varepsilon_{vgg}$. If $v_{gg\_scores} > \varepsilon_{vgg}$, the two images are considered to be matched, otherwise it is not.

3. Feature matching based on semantic information

The description vector generated based on VggFace [11] cannot achieve good recognition effect in face recognition. Therefore, this paper hopes to make full use of the convolution feature map of the middle layer of convolution network to assist matching, so as to improve the recognition rate. In the experiments on different layers, we finally select the convolution feature map of the fifth layer as the matching feature map for auxiliary recognition, which adds further constraints and verification to the VggFace recognition process and improves the recognition performance of VggFace. In this section, we will introduce the auxiliary matching based on VggFace convolution feature graph in detail.

3.1. Auxiliary recognition of VggFace convolution feature map

This paper is an auxiliary recognition method based on the recognition process of VggFace. Therefore, $v_{gg\_scores}$ is calculated for each image of the current input image $I_1$ and the standard database $database = \{I_1^{train}, I_2^{train}, \ldots, I_m^{train}\}$, and all $v_{gg\_scores}$ are sorted, where $m$ is the number of pictures in the standard database. We take the image of the first $K$ frames with the highest $v_{gg\_scores}$ score and match the features of the convolution feature map between each candidate frame and the current input image $I_1$, the specific implementation will be described in detail later. Select the frame with the largest optimal matching number $N_{best}$ among all candidate frames and compare it with the pre-set matching number threshold $N_{thresh}$, as shown in equation (2).

$$matching\_result = \max(N_1^{cand}, N_2^{cand}, \ldots, N_K^{cand}) > N_{thresh}$$  \hspace{1cm} (2)

It is the optimal match $matching\_result$. Otherwise, the optimal match is not found, and the match fails. Since this method is further verified based on the first $K$ frame images with the highest filtering score after $v_{gg\_scores}$ sorting, it can be used to assist VggFace to further improve the recognition accuracy.
3.2. Feature matching of VggFace convolution feature graph

This paper maintains all the operations of VggFace on the input image and extracts the convolution feature map of the fifth convolution layer for research. The fifth layer convolution graph rolls the input image of $224 \times 224 \times 3$ into $14 \times 14 \times 512$ dimensions, it will be the input image is divided into $14 \times 14$ copies, in which the area block in the original graph is represented as a 512 dimensional feature vector in the layer convolution feature graph.

The convolution characteristic graph of the input image $I_1$ can be expressed as $F_1 = \{f_1^1, f_2^1, \ldots, f_{96}^1\}$, where $f_i^1 \in R^{512}, (i = 1, 2, \ldots, 196)$. The convolution feature map matching method still adopts the nearest neighbour ratio method. For each feature $f_i^1, (i = 1, 2, \ldots, 196)$ in the input image $I_1$, the optimal matching is found in the convolution feature graph $F_2 = \{f_1^2, f_2^2, \ldots, f_{96}^2\}$ in the standard dataset image $I_2$. The cosine value of the vector angle $\delta$ between the feature $f_i^1, (i = 1, 2, \ldots, 196)$ in the image $I_1$ and each feature $F_2 = \{f_1^2, f_2^2, \ldots, f_{96}^2\}$ in the image $I_2$ is calculated as the matching score, as shown in equation (3).

$$feat\_scores = \cos \delta = \frac{f_i^1 \cdot f_j^2}{|f_i^1| |f_j^2|}, (i = 1, 2, \ldots, 196, j = 1, 2, \ldots, 196)$$ (3)

The feature $feat\_scores$ with the largest $f_j^2$ is taken as the optimal matching of feature $f_i^1$. Obviously, finally, the feature matching relationship of the convolution feature map of image $I_1$ and image $I_2$ and the corresponding matching score $feat\_scores$ can be established.

A round of screening is carried out for each match, and the optimal match is determined by the matching threshold score $\epsilon_{feat}$. The matching pair of $feat\_scores > \epsilon_{feat}$ is regarded as the optimal matching. The statistical optimal matching number is $N_{best}$.

The second round of verification is performed on the matching results, and the pre-set threshold $N_{thresh}$ is used to determine whether the face image $I_1$ and the face image $I_2$ are the same person. If $N_{best} > N_{thresh}$, it means that the face of the two images is the same person; otherwise, it is not.

4. Experimental comparison and analysis

In this regard, we have done a lot of experiments on a public high-definition face dataset (Bosphorus) [12]. Bosphorus face database is an image with 105 people's posture and expression changes. The face images in this database are generally $1000 \times 1000$ high-definition images. We have carried out experiments on this database and achieved excellent results. In the experiment, 105 people's facial expressionless is taken as the standard database, and face occlusion, face pitch, face rotation of 10 degrees, face rotation of 20 degrees, face rotation of 30 degrees, face rotation of 45 degrees are taken as the test set, and the accuracy of face recognition is taken as the measure of the algorithm.

In this paper, VggFace [11] is used to calculate the description vector $I_1, vgg$ for the test set image $I_1$, and the matching score $vgg\_scores$ is calculated with each $train\_vgg_i, (i = 1, 2, \ldots, m)$ in the standard database database = {$I_1^{train}, I_2^{train}, \ldots, I_m^{train}$}. Set the matching score threshold $\epsilon_{vgg}$, if the matching score is $vgg\_scores > \epsilon_{vgg}$, take the maximum value in all sets greater than the matching score threshold $\epsilon_{vgg}$ as the optimal matching, as shown in equation (4).

$$vgg\_matching\_result = \max(vgg\_scores_1, vgg\_scores_2, \ldots, vgg\_scores_k), k < m$$ (4)

Although the matching based on VggFace [11] description vector can achieve better recognition performance, there is still much room for improvement. Therefore, after VggFace completes the matching score calculation, this paper selects the first $K$ frame images as candidate frames, uses its fifth layer convolution feature map for feature matching, selects better matching points through the score threshold $\epsilon_{feat}$, determines the optimal matching through the pre-set matching threshold $N_{thresh}$, and then obtains the recognition results.
Figures 5-7 show the recognition effect of some experiments. Each frame of test image $I_1$ is matched with 105 frames of standard data set one by one, and the matching score is calculated to obtain the dimensional matching score matrix, in which the diagonal is the brightest, and the other areas are all dark, which means that the matching effect is the best.

**Figure 5.** Under different occlusion conditions, left: VggFace matching score and right: auxiliary recognition matching score: (a) eye occlusion, (b) mouth occlusion.

**Figure 6.** Under different pitching conditions, left: VggFace matching score and right: auxiliary recognition matching score: (a) head pitching up, (b) head pitching down.
The matching score between the feature vectors generated by VggFace [11] and the method based on convolution feature map are all tested in this paper. The visualization of experimental results is shown in Figure 5-7. The recognition accuracy is shown in Table 1 and table 2.

Table 1. Recognition accuracy of VggFace and our method on occlusion and pitch angle transformation.

| method  | Mouth occlusion | Eye's occlusion | Head up   | Head down |
|---------|-----------------|-----------------|-----------|-----------|
| VggFace | 80.0000%        | 84.7619%        | 66.0377%  | 58.0592%  |
| ours    | 97.1429%        | 97.1429%        | 87.7359%  | 83.8095%  |

From the experiment, we can see that although VggFace has mined the global information of face image, its recognition rate still has a lot of room to improve in the recognition of occlusion and head changes. Through the auxiliary recognition method proposed in this paper, the recognition accuracy of VggFace is significantly improved.

Table 2. Recognition accuracy of rotation angle transformation by VggFace and our method.

| method  | R10       | R20       | R30       | R45       |
|---------|-----------|-----------|-----------|-----------|
| VggFace | 96.1905%  | 92.3810%  | 79.0476%  | 69.5238%  |
| ours    | 99.0476%  | 97.1429%  | 96.1905%  | 86.6667%  |

From the experiments, we can see that the auxiliary recognition method proposed in this paper is more rotation invariant than single VggFace and has better recognition performance than VggFace for face images with rotation angle less than 45 degrees.
5. Conclusion
We propose an auxiliary face recognition method based on convolution feature graph, which is a very reliable auxiliary face recognition method. This method uses VggFace to extract the description vector and the data in the standard database for matching score calculation, selects the first $K$ candidate frames with the highest matching score, uses the convolution vector map of VggFace to do feature matching between the input image and the candidate frames, and selects the candidate frame with the largest number of optimal feature matching as the matching image to achieve the recognition effect. We have done substantial experiments in a high-definition face image database and compared it with VggFace’s face recognition algorithm. Obviously, the method proposed in this paper can achieve higher accuracy and can be used for vggface and even other deep learning networks.

Acknowledgments
This work was supported in part by Guangdong Natural Science Youth Fund under Grant (No.2020A151510162), Key projects of social science and technology development in Dongguan under Grant (No. 2020507156156), in part by Special fund for science and technology innovation strategy of Guangdong Province under Grant (No. pdjh2020a1261).

References
[1] Turk M, Pentland A. Eigenfaces for recognition.[J]. J Cogn Neurosci, 1991, 3(1):71-86.
[2] Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 19(7):711-720.
[3] Ahonen T, Hadid A, Pietikäinen M. Face Recognition with Local Binary Patterns[C]// European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2004:469-481.
[4] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91-110.
[5] Li D, Lam K M. Design and learn distinctive features from pore-scaled facial keypoints[J]. Pattern Recognition, 2015, 48(3):732-745.
[6] Li D, Zhou H, Lam K M. High-Resolution Face Verification Using Pore-Scale Facial Features[J]. IEEE Transactions on Image Processing, 2015, 24(8):2317-2327.
[7] Li D, Xie X, Lam K M, et al. Elastic block set reconstruction for face recognition[C]// IEEE International Conference on Image Processing. IEEE Press, 2009:3293-3296.
[8] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786): 504-507.
[9] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[11] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition[C]//BMVC. 2015, 1(3): 6.
[12] Savran A, Alyüz N, Dibeklioğlu H, et al. Bosphorus database for 3D face analysis[C]//European Workshop on Biometrics and Identity Management. Springer, Berlin, Heidelberg, 2008: 47-56.
[13] Ding C, Choi J, Tao D, et al. Multi-directional multi-level dual-cross patterns for robust face recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(3): 518-531.
[14] Ding C, Tao D. Trunk-branch ensemble convolutional neural networks for video-based face recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 40(4): 1002-1014.
[15] Best-Rowden L, Jain A K. Longitudinal study of automatic face recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 40(1): 148-162.