Face feature matching based on semantic Information

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 feature matching problems have always been a very important research topic in the field of image processing. Face feature matching is always the most important part of common face recognition algorithms, but how to match features quickly and accurately is a quite essential and urgent problem to be solved. Therefore, we propose a face feature matching method which can effectively improve the matching efficiency and accuracy. For each feature point in the input image, face feature matching based on nearest neighbor ratio method requires global search for the best matching in the reference image. However, we find that such matching method can achieve better matching results, but it costs a lot of computing resources and time. Therefore, we use the feature vectors constructed by CNN convolution feature map to provide the local matching region based on semantic information for handcrafted descriptors after feature matching. Although two feature matching processes are used in this paper, the matching efficiency is improved several times compared with the traditional global search scheme. Experiments show that our proposed can be used as an important module of face recognition system to improve the accuracy of face recognition.

1. Introduction

Face feature matching is always a very important problem in the field of pattern recognition. Scholars and related enterprises in the field of computer recognition have invested a lot of energy and financial resources to improve the efficiency and accuracy of face feature matching. The nearest neighbour matching method [1] [2] is to find the one with the lowest matching score in all the feature points of the reference image for each feature point of the input image as the optimal matching. This method ensures that there is a corresponding matching feature point in the reference image for any feature point of the input image, but the matching quality is not guaranteed. Therefore, the improved nearest neighbour proportion method based on the nearest neighbour matching method [3] improves this defect. When the ratio of the optimal matching score and the suboptimal matching score meets a certain threshold, the optimal matching is determined. However, the above two schemes are based on global search, which consumes a lot of computing resources and time. Therefore, for the problems that need to match a large number of feature points, such as 3D face reconstruction, the computational loss of this scheme is catastrophic. In order to reduce the search range, Chen et al. [4] proposed a matching method based on K-nearest neighbour. The extracted feature points were clustered by K centers, and the clustering centers were determined. In feature matching, first of all, rough matching is carried out in the cluster center to determine the matching range, and fine matching is carried out in the cluster center with the highest matching score to find the most matching feature points in the class. However, this method lacks semantic information and cannot guarantee the correctness of matching. In addition to changing the search strategy, some scholars also proposed to change the matching score measurement method. Orb [5] proposed using Hamming distance to measure the similarity of features.
Facenet [6] proposed to project image features into Euclidean space, directly compare the scores, and determine the matching degree by threshold conditions. In order to reduce the search scope, this paper proposes a matching method based on semantic information guidance, in order to improve the matching efficiency and accuracy.

2. Feature matching based on vggface convolutional neural network

Traditional hand-designed features, such as sift [3], orb [5] or surf [7] [8], have some limitations in mining image information, such as not considering the relationship between feature points and feature points. The deep learning method based on CNN [9] [17] can learn many more specific features from a large number of face image data sets, such as more advanced features composed of edges and corners. The higher the characteristics, the stronger the specificity. Therefore, this paper uses the current mainstream CNN network vggface [10,18,19] to extract advanced features in the face image, provide semantic information for face feature matching, improve the number of matching and matching efficiency.

2.1. Feature Generator

The feature generator used in this system is based on the convolution feature map extracted by vggface [10]. In order to clarify the discussion, this paper first gives a brief introduction to vggface.

With its excellent recognition performance, alexnet [11] won the ilsvrc of Imagenet [12] project in 2012. It also attracts more and more researchers to study convolutional neural network. Vggnet [13] [14] achieves better recognition ability by increasing the number of convolution layers on the basis of alexnet. However, the experimental results of vggnet show that there are some trade-offs between recognition performance and computing resources. When the convolution layer of vggface reaches 19 layers, the performance improvement is very weak compared with 16 layers, but it consumes more computing resources and time. Therefore, after balancing computing resources and recognition performance, VggNet recommends using Vgg16, that is, 16 convolution layers. VggFace also uses Vgg16 to generate face image features. The specific network configuration of vggface is shown in Figure 1.

![Figure 1. Vggface convolution neural network structure.](image)

The face image needs to be preprocessed before being input into VggFace, and the image is down sampled and cut to size $224 \times 224 \times 3$. Input the preprocessed image into VggFace, perform convolution operation and full connection layer calculation, and obtain $4096 \times 1$ -dimensional description vector. The specific process is shown in Fig. 2.
Figure 2. VggFace process for converting input image into description vector.

If the input image is set as $I_{in}$, the input image is convoluted, as shown in Formula (1):

$$I_{conv} = h_{16} \ast h_{15} \ast \cdots h_{2} \ast h_{1} \ast I_{in}$$  \hspace{1cm} (1)

Where $h_{i}, i \in [1,16]$ represents the $i$th convolution and $I_{conv}$ represents the final convolution output.

2.2. Convolution feature generator

The convolution feature map of convolution neural network contains rich high-level features and spatial distribution information. In this paper, the convolution feature map of VggFace is used as the feature generator to construct the deep learning feature vector.

Since VggFace [10] contains multiple convolution layers, based on Chen's method [15], this paper compares the performance of multiple convolution layers of VggFace and determines that the convolution layer $conv_5_3$ has the best recognition performance. Therefore, this paper uses the convolution feature graph of $conv_5_3$ as the feature generator.

The size of the convolution feature map output by the convolution layer $conv_5_3$ is $14 \times 14 \times 512$, the input image with size $224 \times 224$ is condensed into a convolution feature map of $14 \times 14$, and each pixel in the convolution feature map is represented by $n$-dimensional features. Therefore, we can construct $14 \times 14 \times 512$-dimensional depth learning feature vectors to describe the original input image.

2.3. Deep learning feature matching

The purpose of feature matching is to find an optimal matching feature in the reference image for each feature of the input image.

In this paper, the nearest neighbor matching method is used to measure the similarity by the cosine fraction between two feature vectors. Assuming that one of the deep learning features of the input image is $dF_{in}$ and the feature of the reference image is $dF_{ref}$, the matching score is calculated as shown in the formula.

$$ms = \frac{dF_{in} \bullet dF_{ref}}{||dF_{in}|| ||dF_{ref}||}$$  \hspace{1cm} (2)

For each feature $dF_{in}$ in the input image, there is a matching score:

$$matchScores = \{ms_1, ms_2, \cdots, ms_{196}\}$$  \hspace{1cm} (3)

The corresponding feature with the largest matching score is the optimal matching of the current feature.

3. Feature matching based on semantic information

In the field of feature matching, the most commonly used matching methods are nearest neighbor matching method [1] [2] and nearest neighbor proportion method [3], because these two methods are the simplest and most effective. However, no matter which of the above methods is used, it is inevitable to consume a lot of computing resources. When thousands of feature points are extracted from both images, each feature point of the input frame needs to be matched with all feature points in the reference frame to calculate the matching score. Although very good matching results can be obtained, the consumption of computing resources during the period is disastrous. Therefore, this
paper limits the matching area by adding semantic information, reduces the scope of matching, and then improves the matching efficiency and accuracy.

3.1. Build matching area mapping relationship

In order to determine the matching range, each depth learning feature vector is mapped back to the corresponding region of the original input image.

Any deep learning feature vector is \( dF_{ij} \in \mathbb{R}^{512 \times 1} \), where \( i \in [1, 14], j \in [1, 14] \). It is known that the original input image is \( I_{in} \in \mathbb{R}^{W \times H \times 3} \). The original input image is divided into \( 14 \times 14 \) regions, and the range of each region is: \( \text{region} = (w, h) = \left( \frac{w}{14}, \frac{h}{14} \right) \). Then the mapping relationship between each feature vector and the original input image is shown in Formula (4):

\[
dF_{ij} \Rightarrow I_{in}(w \cdot (i - 1):w \cdot i, h \cdot (j - 1):h \cdot j)
\]

(4)

Where, \( I(m, n) \) represents the pixel value with coordinate \( (m, n) \), \( I(m: m+k, n: n+l) \) represents the rectangular area surrounded by starting point coordinate \( (m, n) \) and ending point coordinate \( (m+k, n+l) \).

3.2. Feature matching based on semantic information

Through the matching filtering of deep learning feature vectors and the mapping relationship constructed above, we can narrow the matching range of feature points, as shown in Figure 3. Based on the guidance of semantic information, the number and speed of feature point matching are improved.

The manually designed SIFT feature vectors [3] are extracted from all the mapped regions. Due to the different textures of each region of the input image, the number of SIFT features extracted from each region is also different. For any SIFT feature vector \( dS^R_{i} \in \mathbb{R}^{128 \times 1} \) in the \( R \) region of the input image, the optimal matching is only found in the \( K \) region of the reference image. After the first round of semantic information screening, we found a better matching region. Therefore, feature matching in the specified matching area can greatly improve the matching speed.

The method used in this paper to measure the similarity of feature points is also the nearest neighbor proportion method, but compared with the traditional matching method based on global search, this paper reduces the matching area and improves the matching efficiency. The more specific screening method is shown in formula (5)-(7):

\[
ms = \frac{dF_{in} \cdot dF_{re}^{max}}{|dF_{in}||dF_{re}^{max}|}_{max}
\]

(5)
Where \( dF_{in} \) represents a SIFT feature vector of the input image \( I_{in} \), and \( dF_{re}^{max} \) represents the feature vector with the highest matching score between the corresponding matching region and the feature vector in the reference frame. \( ms_{max} \) represents the highest matching score.

\[
ms_{sec} = \frac{dF_{in} \cdot dF_{re}^{sec}}{|dF_{in}| |dF_{re}^{sec}|}
\]  

(6)

Similarly, \( dF_{re}^{sec} \) represents the feature vector with the second highest matching score between the corresponding matching region in the reference frame and the feature vector, and \( ms_{sec} \) represents the second highest matching score. Assuming that the preset matching score threshold is \( S_{th} \), the judgment method of whether the eigenvector matching is the best is shown in formula (7):

\[
bestMatch = \begin{cases} 
0, & S_{th} \geq \frac{ms_{max}}{ms_{sec}} \\
1, & S_{th} < \frac{ms_{max}}{ms_{sec}}
\end{cases}
\]  

(7)

Matching pairs larger than the threshold condition are regarded as the optimal matching. Otherwise, they do not meet the conditions and are excluded.

4. Experimental comparison and analysis

In order to verify the correctness of this paper, we have done a lot of experiments on facial expression changes on a public high-definition face data set Bosphorus [16]. The general size of the database is 1000 x 1000 high-definition face images. We have carried out the relevant experiments in this database and achieved excellent matching results. The experiment takes 105 people’s face expressionless as the standard database and each person’s expression as the test set. By comparing the SIFT feature vector matching effect of the nearest neighbour proportion method with semantic information and under normal conditions, the number of feature matching pairs and matching efficiency are taken as the measurement of the algorithm.

In order to better illustrate the practicability of this paper, this paper uses the same response threshold and the same matching threshold to extract SIFT features and feature matching from the input image. The only difference in the comparative experiment is the matching method. The method proposed in this paper is that the feature matching only searches for the optimal matching in the specified area under the guidance of semantic information, while the traditional nearest neighbour proportion method searches for the optimal matching in the whole graph.

The feature matching effect of the nearest neighbour proportion method based on semantic information to shorten the matching range and the nearest neighbour proportion method of traditional global search on different expressions of the same person is shown in Figure 4. Obviously, the matching quantity and quality of the method proposed in this paper are better than the traditional global search method.
Figure 4. The left two images are the matching results of the same person's different expressions based on semantic information matching method, and the right two images are the matching results corresponding to the traditional global search.

The feature matching effect of the nearest neighbour proportion method based on semantic information to shorten the matching range and the nearest neighbour proportion method of traditional global search on the same expression of different people is shown in Figure 5. Obviously, the matching effect of the scheme in this paper is better than that of the traditional scheme based on global search.

This paper analyzes the comparison of the number of optimal matching pairs between the traditional global search and the proposed method under the same matching threshold under different expressions. The experimental results are shown in Table 1. Obviously, the number of feature
matching in the proposed method is nearly 30% to 50% higher than the traditional global search matching method.

Table 1. Comparison of average optimal matching pairs of different expressions between traditional global matching and our method.

| Method / expression | Angry  | Anxious | Happy  | Sad   |
|---------------------|--------|---------|--------|-------|
| Traditional global method | 200.34 | 232.24  | 148.51 | 244.65 |
| Our approach         | 287.69 | 310.24  | 226.13 | 343.83 |

In order to prove that the method in this paper improves the matching efficiency compared with the traditional method, this paper also compares the difference of the two methods in average matching time. The experimental results are shown in Table 2. Obviously, compared with the traditional matching scheme based on global search, the matching time of the proposed matching scheme is nearly 4 times higher and nearly 1 second is saved. Although we use two matching processes, the overall matching time is still much less than the matching scheme based on global search.

Table 2. Comparison of average matching time of different expressions between traditional global matching and our method.

| Method / expression | Angry (s) | Anxious (s) | Happy (s) | Sad (s) |
|---------------------|-----------|-------------|-----------|--------|
| Traditional global method | 1.2741    | 1.3521      | 1.3779    | 1.3314 |
| Our approach         | 0.3586    | 0.3583      | 0.3758    | 0.3576 |

From the experiment, we can see that although the same threshold conditions are adopted, the matching method based on semantic information proposed in this paper has significantly improved the matching number and matching efficiency compared with the traditional matching method based on global search. The matching method is limited by semantic information. In fact, a matching has been completed, and this matching is based on higher-level semantic features. Obviously, the quality of matching is higher. Then, through the mapping relationship, the specific areas with semantic guidance are matched. Due to the reduction of the search scope, the matching efficiency is naturally improved.

5. Conclusion
We propose a face feature matching method based on semantic information, which is a very reliable face feature matching method. In this method, the convolution feature map extracted by vggface is used as a feature generator to construct deep learning feature vector, and the advanced features are matched. Then, the deep learning feature vector is associated with the specific region of the original input image through the mapping relationship. Hand designed SIFT features are extracted from all the specified matching regions, and the matching time is greatly shortened by matching the corresponding regions based on semantic information. And because the first matching is based on high-quality advanced features, the quality of matching pairs is also improved when re matching. We have done enough experiments on a high-definition face image database Bosphorus, and compared with the traditional global search matching methods, it is obvious that the proposed method can produce more correct matching pairs, and the matching efficiency is also greatly improved.

Acknowledgments
This work was supported in part by Guangdong Natural Science Youth Fund under Grant (No.2020A1515110162), 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] Tomasi C, Manduchi R. Stereo matching as a nearest-neighbor problem[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(3): 333-340.
[2] Muja M, Lowe D G. Fast approximate nearest neighbors with automatic algorithm configuration[J]. VISAPP (1), 2009, 2(331-340): 2.
[3] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110
[4] Chen Q, Li D, Tang C K. KNN matting[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(9): 2175-2188.
[5] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF[C]//ICCV. 2011, 11(1): 2.
[6] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 815-823.
[7] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2006: 404-417.
[8] Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (SURF)[J]. Computer vision and image understanding, 2008, 110(3): 346-359.
[9] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436.
[10] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition[C]//bmvc. 2015, 1(3): 6.
[11] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
[12] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
[13] Ke H, Chen D, Li X, et al. Towards brain big data classification: epileptic EEG identification with a lightweight VGGNet on global MIC[J]. IEEE Access, 2018, 6: 14722-14733.
[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[15] Chen Z, Lam O, Jacobson A, et al. Convolutional neural network-based place recognition[J]. arXiv preprint arXiv:1411.1509, 2014.
[16] Savran A, Aly ü z 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.
[17] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
[18] 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.
[19] 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.