Retraction Notice

The Editor-in-Chief and the publisher have retracted this article, which was submitted as part of a guest-edited special section. An investigation uncovered evidence of systematic manipulation of the publication process, including compromised peer review. The Editor and publisher no longer have confidence in the results and conclusions of the article.

TY either did not respond directly or could not be reached.

Estimation of human age by features of face and eyes based on multilevel feature convolutional neural network

Tangtang Yi*
Hunan Women's University, School of Information Science and Engineering, Changsha, Hunan, China

Abstract. Age estimation can be effectively used in security, human–computer interaction, entertainment, and audio-visual fields. However, current face-based age recognition algorithms are not highly accurate due to factors, such as face makeup and plastic surgery. Among the various feature points of a human face, the eyes are the most difficult part to be modified. And with the changes of age, the lens, cornea, and vitreous of eyes will also change accordingly. Therefore, we believe that paying attention to the local feature of the eye can improve the accuracy of age estimation to a certain extent. The multilevel feature convolutional neural network (MLFCNN) is proposed, which values eyes features and combines them with face features to jointly estimate human age. MLFCNN performs two rounds of estimation based on extracted features. First, the age range of the sample is estimated as the age group, and then on this basis, the fine age of the sample is further estimated. Field tests show that the mean absolute error of MLFCNN is 2.87, which is lower than other network models tested. When the MLFCNN estimated age did not differ more than 4 years from the actual age of the sample, the network predictions were considered correct (i.e., the tolerable age error was 4). The age estimation accuracy of MLFCNN under this condition was 91.14%. And 98.32% accuracy can be reached when the tolerable age error is 6. Ablation experiments verify that the fusion of ocular features and facial features can improve the performance of the age estimation network. In addition, it can still maintain a good performance under small training dataset. © *2022 SPIE and IS&T* [DOI: 10.1117/1.JEI.31.4.041208]

Keywords: age recognition; face; eye; convolutional neural network; age estimation.

Paper 210542SS received Aug. 17, 2021; accepted for publication Dec. 27, 2021; published online Jan. 11, 2022.

1 Introduction

As an important biological and social characteristic of human beings, age plays a vital role in human social interaction. Face age estimation refers to predicting the age corresponding to the face in the image. In recent years, with the rapid increase in the amount of monitoring equipment installed in public places, the problem of facial age estimation is closely related to actual demand. In addition, there are good application prospects in many intelligence fields, such as face age prediction, cross-age face recognition, intelligent security monitoring, harmonious human–computer interaction, and marketing analysis.[1,2]

Many scholars have done relevant research on age estimation. Face age estimation based on traditional machine learning is usually divided into two steps: face representation and age prediction. Face characterization usually uses shallow appearance, such as active appearance model, local binary pattern, and bionic features.[3–7] After obtaining the feature representation of each face image, age estimation can be regarded as a classification or regression problem to solve. Due to the limitations of machine learning, the performance of traditional age estimation algorithm is difficult to meet the requirements. First, features based on manual design usually require strong prior knowledge, and are therefore quite cumbersome. In addition, the traditional feature

*Address all correspondence to Tangtang Yi, tangyiyihn@126.com

1017-9909/2021/$28.00© 2021 SPIE and IS&T

extraction and fusion algorithm cannot make full use of the advantages of big data. With the expansion of training sample size, the improvement of network performance is not obvious.[8]

To break through the limitations of traditional algorithms, more and more age estimation models adopt the network architecture of in-depth learning. For example, Dong et al.[9] divided the age into seven groups using CNN + support vector machines (SVM) model. And DeepID network was used to extract features of face, while SVM was used as classifier. Niu et al.[10] proposed an end-to-end deep learning network to solve the problem of ordered regression in human age estimation. They used a multi-output CNN to deal with multiple binary classifications in ordered regression subquestions. Based on the Asian Face Age dataset, this method can achieve excellent human age estimation effect. Liu et al.[11] trained a classifier and a regressor on the basis of a pretrained model of a large-scale deep neural network (DNN), and then fused the classifier and regressor to achieve a superior performance in age estimation. This method is called AgeNet, which is one of the leading current fields of age estimation algorithm. The current age estimation algorithm generally uses the entire face image as the input of the network, but does not pay attention to some local features. As the eye changes with age, the lens, cornea, vitreous body, etc. will all change, and it is not easy to be modified. Therefore, this paper designs a neural network based on eye features and full-face features to estimate the age of humans, and verifies the performance of the network and the ability of eyes to represent human age through experiments.

The main contributions of this paper are as follows:

1. Development of a human age estimation system based on multilevel feature convolutional neural network (MLFCNN) has been proposed, which combines the global features and local features of the face for fine-grained age estimation.
2. Instead of only using facial features to estimate human age, importance is attached to the local features of the eyes, which are fused with facial features to comprehensively estimate the age. And the network is very extensible, other local features besides eye features can also be easily added to the network structure.
3. Two rounds of estimation were conducted based on extracted features. Firstl, the age range of the sample is estimated as the age group, and then on this basis, the actual age of the sample is further estimated.

2 Related Work

The existing estimation methods of human age include three main categories: regression model, classification model, and rank model. The regression model regards age estimation as a regression task, and establishes a model representing facial age changes to estimate human age through regression analysis. According to whether the classification network is used in the regression task, the regression model can be divided into two categories: direct regression model and classification regression model. For the direct regression model, Yi et al.[12] proposed a multiscale deep convolutional neural network (CNN) to estimate human age. Multiple DNNs with different scales were used to extract facial features from different facial regions. According to results of feature extractions, the variance loss function is used to perform age regression. Ranjan et al.[13] proposed a 16-layer deep CNN[14] to extract features of face images, the apparent age was estimated by artificial neural network. And learning to aggregate and personalize was used to perform age regression. Although regression-based age estimation networks mentioned earlier have been greatly improved compared with the traditional model, the convergence performance and classification effect need to be improved. Researchers began to add a classification network to the regression model to improve the performance of age estimation networks. Rothe et al.[15] proposed a deep expectation (DEX) network to estimate apparent age. The DEX network changes the number of neurons in the last layer of the VGG-16 network to 101. The modified network is used to extract facial image features and classify facial ages, thereby obtaining 0 to 100 years old belonging to 101 categories. Finally, the class probability of the classifier is multiplied by the corresponding age to obtain the age predicted by the network. This method won the ChaLearn2015 facial appearance age estimation competition, but the network is too deep and the training is difficult. In addition, it is prone to overfitting.

The method based on the ranking model compares the estimated age with a series of age values to determine the position of the age target value in the age series. This method mainly addresses the problem of ignoring dynamic, fuzzy, and personalized features in the face aging process in traditional methods.[16–28] Chen et al.[19] proposed Rank-CNN, which converts the facial age estimation problem into a multiple sequential binary classification problem. As an extension and upgrade of the traditional algorithm, the ranking model can make up for the shortcomings of the traditional method, but the ranking model has a low fault tolerance rate and a high demand for network pruning, so the final effect is difficult to achieve the expected results.

As the accuracy of age estimation applications continues to rise, age group estimation under unrestricted conditions has become one of the focus of current research, and multiclass models are the main means to achieve detailed age estimation. Levi et al.[20] use diffusion-convolutional neural networks three convolutional layers and two fully connected layers to classify age and gender on the Adience dataset under unlimited conditions. To solve the problem that the Softmax cross entropy loss function cannot model the correlation between age groups, Hou et al.[21,22] proposed the EMD2 loss function instead of the traditional cross-entropy loss function, and performed age group classification in the VGG network.[23] Based on the age estimation network of the classification model, different degrees of age estimation effects can be achieved according to the set categories to meet different needs. Liu et al.[24] proposed a structure-based age estimation network, which first identifies gender, identifies age groups based on different genders, and finally performs fine classification within the group to obtain accurate age. Age estimation based on classification is a development trend, but factors, such as makeup and plastic surgery, largely limit the accuracy of classification tasks. Eyes are the most difficult part to be modified and can reflect the changes of age. In this paper, local features of age and global features of face are combined to carry out human age estimation, and the performance of network and the effect of eyes on age estimation are verified through experiments.

3 Architecture of Age Estimation System

The proposed age estimation system consists of four stages, and the architecture of the system proposed is shown in Fig. 1. The first stage is the face and eyes detection network. In practical applications, the background of the picture is generally complicated. To reduce environmental interference, this paper locates the face and eye area separately as the input of the subsequent network. In the second part, the features of the face and eyes are extracted. In the third part, the age group is estimated, and the facial features are further extracted. Finally, the facial features, ocular features, and age groups are integrated to comprehensively estimate the true age of the subject. This paper adopts a mechanism combining rough estimation and fine estimation. First, the age range of the detected person (the age group) is estimated based on the features of the face. Then, the rough estimation results are superimposed with the subsequent network to comprehensively estimate the age.

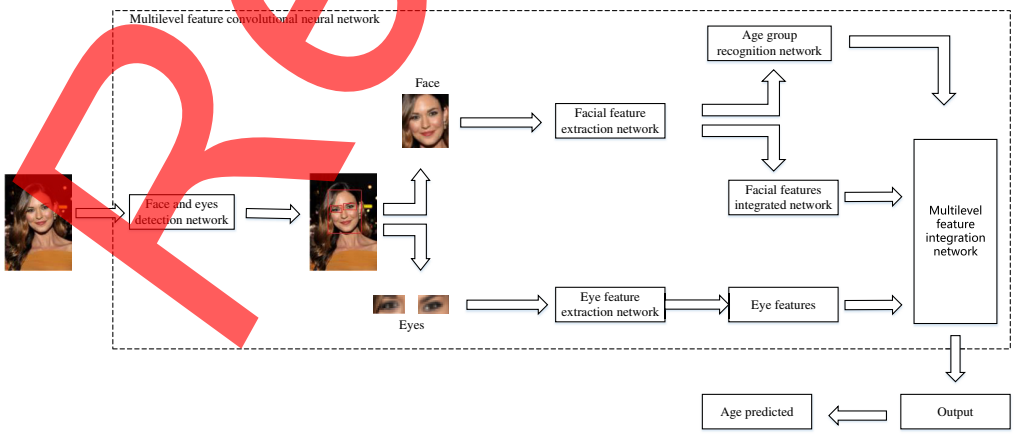

Fig. 1 The architecture of proposed age estimation system.

4 Multilevel Feature Convolutional Neural Network

The network structure diagram of the MLFCNN is shown in Fig. 8. And each part of the network will be explained separately.

4.1 *Face and Eyes Detection Network*

To avoid environmental interference on age estimation results, MLFCNN needs to locate and extract features from face and eye regions. In the area of target detection, excellent algorithms include Faster-region-convolutional neural networks (Faster-RCNN),[25] you only look once (YOLO),[26] single shot multibox detector (SSD),[27] etc. Because only eyes and faces need to be detected in this paper, direct use of these target detection algorithms will consume too much system resources and time. Based on the ideas of YOLO, SSD, and anchor box, a simplified eyes and face detection module is designed, which uses a lighter network to locate eyes and faces. In this paper, input image is divided into different blocks by grid refer to the idea of YOLO and Faster-RCNN. 10×10 grid lines are used to segment the input image, as shown in Fig. 2(b), and 11×11 grids and 10×10 intersection points can be obtained.

As shown in Fig. 2(c), there are a total of $10 \times 10 = 100$ intersecting points of grid lines, called anchor points, around which a set of candidate boxes are generated. Since the shapes of face and eye are basically fixed, six anchor boxes are generated for each anchor point, including two shapes (1:1, 2:1) and three size. The schematic diagram of the anchor box is shown in Fig. 3.

As shown in Fig. 3, the central point of each anchor box is distributed on or near its corresponding anchor point. The parameters contained in an anchor box are shown in Eq. (1)

$$\text{anch} = (\ \text{CF} \quad w_i \quad h_i \quad \sigma_x \quad \sigma_y \quad x_h \quad y_h \quad C_1 \quad C_2\ ), \tag{1}$$

where CF is the confidence of the anchor box, indicating the probability of the existence of objects in the current anchor box and the coincidence rate between the ground truth and anchor box. The length and width of current anchor box are represented by w_i and h_i. As shown in

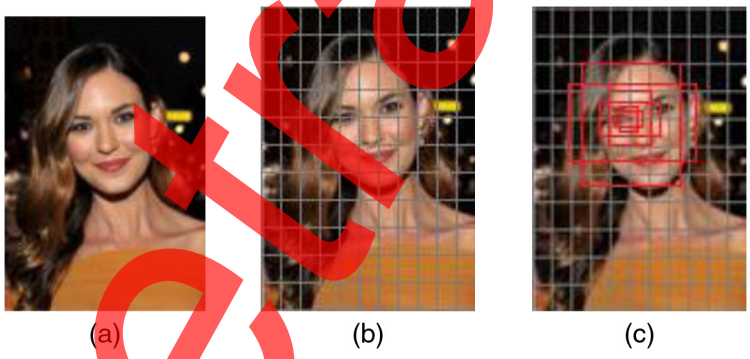

Fig. 2 Schematic diagram of input image with grid.

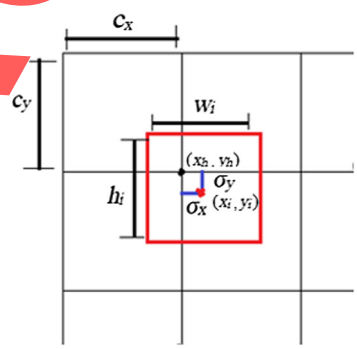

Fig. 3 Schematic diagram of the anchor box.

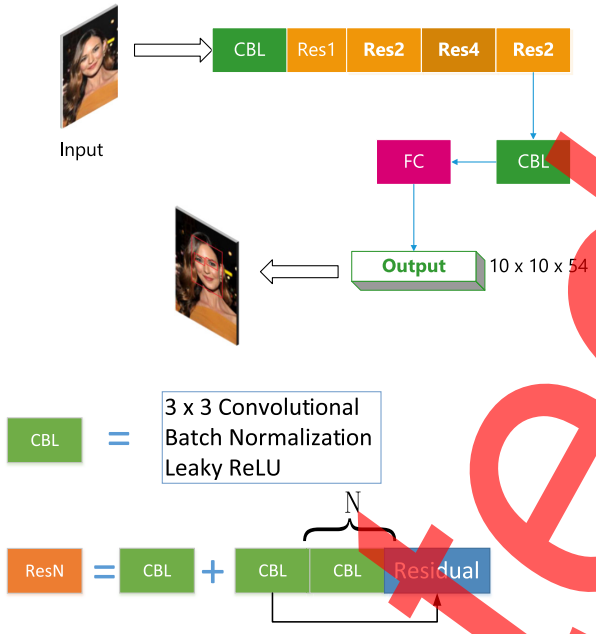

Fig. 4 The structure of the face and eyes detection network.

Eq. (2), σ_x and σ_y represent the offset degree of the center of the anchor box relative to the anchor point. x_h and y_h are the coordinates of the anchor point corresponding to current anchor box. C_1 and C_2 are the probability values of eyes and faces if there is a target in the anchor box

$$\begin{cases} \sigma_x = \frac{x_i - x_h}{c_x} \\ \sigma_y = \frac{y_i - y_h}{c_y} \end{cases}. \tag{2}$$

As shown in Eq. (1), an anchor box contains nine parameters, while an anchor corresponds to six anchor boxes, so each anchor has a total of $6 \times 9 = 54$ parameters. An input image contains 10×10 anchors, so the output of the face and eyes detection network is a vector with size of $10 \times 10 \times 54$. According to the confidence, anchor boxes representing the face and eyes are selected as the input to subsequent networks. The structure of the face and eyes detection network is shown in Fig. 4.

To prevent network from gradient disappearance and accelerate the convergence speed of the network, the Batch Normalization layer is widely used in this paper to normalize the network and ensure the nonlinear expression ability of the network model. For deep convolutional networks, with the deepening of the network, the training effect becomes worse and worse, and the network is difficult to converge. Referring to He et al.'s deep residual network model,[28] this paper uses the residual structure as the main structure of the network.

Figure 5 is a simple residual unit structure model. The residual network draws on the idea of cross-layer linking of high-speed networks, and directly passes the input x to the output as the

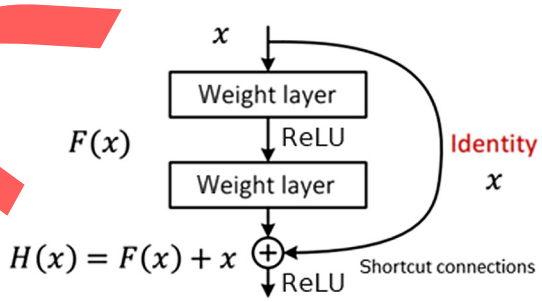

Fig. 5 A simple residual unit structure model.

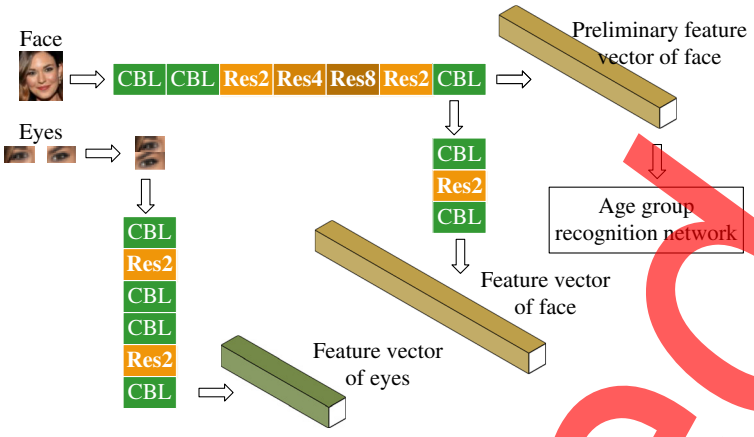

Fig. 6 The structure of the feature exaction network.

initial result through shortcut connection. The network output becomes $H(x) = F(x) + x$, and $F(x) = H(x) - x$ is called the residual. When $F(x) = 0$, $H(x) = x$ can be obtained. Therefore, the goal of network training is to approximate the residual result to zero.

4.2 *Features Extraction Network*

When the face and eyes are located, feature extraction is performed on the face area and the eye area, respectively. Combined with the features of the whole face and the local features of the eyes, a comprehensive assessment of age can be made. This paper uses the network model shown in Fig. 6 to extract features. Cut the detected human face and eyes into separate pictures, resize the picture containing the face to a standard size of 512×512, and resize each eye to a standard size of 128×64. Since there are left and right eyes in the eye area, to facilitate network processing, the left and right eyes are superimposed up and down to form a 128×128 complete image. For some images with special angles, if the picture contains only one eye, this eye is used to superimpose twice. After obtaining facial feature data, it is necessary to perform age group recognition and further feature extraction.

4.3 *Age Group Recognition Network*

This paper designed an age group recognition network to improve the recognition accuracy and enhance the stability of the network. Estimate age range before making a fine age estimate. This paper divides age into 16 groups as shown in Table 1. Different groups cross each other to improve the robustness of the estimate.

Table 1 Age groups.

Group number	Age range (years old)	Group number	Age range (years old)
1	0 to 2	9	24 to 32
2	1 to 6	10	28 to 36
3	4 to 10	11	32 to 42
4	6 to 14	12	36 to 48
5	10 to 18	13	40 to 52
6	14 to 20	14	44 to 58
7	16 to 24	15	50 to 59
8	20 to 28	16	60+

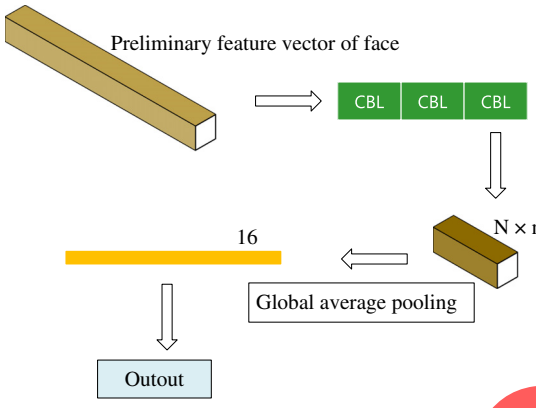

Fig. 7 The structure of the age group recognition network.

According to the age group recognition network shown in Fig. 7, the preliminary feature vector of face is analyzed, and the network output is a vector with a length of 16. And the vector is composed of the probability values of the current target belonging to 16 age groups, from which the group with the highest probability value is selected. After the age group of current target is obtained, when the age is estimated, the weight of the age in the age group can be increased.

There are too many parameters contained by the full connection layer (FCN), so that the global average pooling (GAP) layer is used instead of FCN. GAP can significantly reduce the amount of network calculations and speed up training, and can also prevent overfitting.

4.4 *Multilevel Feature Integration Network*

This paper can estimate a total of 60 categories from 1 to 59 years old and over 60 years old (≥ 60). The overall structure of MLFCNN is shown in Fig. 8. The multilevel feature integration network can comprehensively estimate the age based on the face feature vector, eye feature vector, and age group. First, the face and eye features are integrated to obtain a comprehensive feature vector. Then, according to age groups, a set of weight parameters is formed and superimposed with comprehensive feature vectors to obtain the comprehensive result of age estimation.

4.5 *Multilevel Feature Integration Network*

According to the structure diagram of MLFCNN, the loss function of the network consists of three parts. The first part is the loss of the face and eyes detection network, the second part is the loss of the age group recognition network, and the final part is the loss of age estimation. The loss of the face and eyes detection system includes regression loss and classification loss, as shown in Eq. (3)

$$\begin{aligned}\text{Loss}_1 = {} & \delta_{\text{obj}} \sum_{i=0}^{10\times10} \sum_{j=0}^{6} E_i^j(\sigma_x^2 + \sigma_y^2) + \delta_{\text{obj}} \sum_{i=0}^{10\times10} \sum_{j=0}^{6} E_i^j[(\sqrt{w_i} - \sqrt{\widehat{w_i}})^2 + (\sqrt{h_i} - \sqrt{\widehat{h_i}})^2] \\ & + \sum_{i=0}^{10\times10} \sum_{j=0}^{6} E_i^j(CF - \hat{CF})^2 + \delta_{\text{noobj}} \sum_{i=0}^{10\times10} \sum_{j=0}^{6} \bar{E}_i^j(CF - \hat{CF})^2 \\ & + \sum_{i=0}^{10\times10} E_i \sum_{c\in\text{classes}} (p_i(c) - \widehat{p}_i(c))^2 \end{aligned} \tag{3}$$

where E_i^j indicates whether there is a target object in the j'th anchor box of the i'th anchor, if it exists, it is 1, and if it does not exist, it is 0. $\overline{E_i^j}$ is opposite to E_i^j, it is 1 when there is no object in

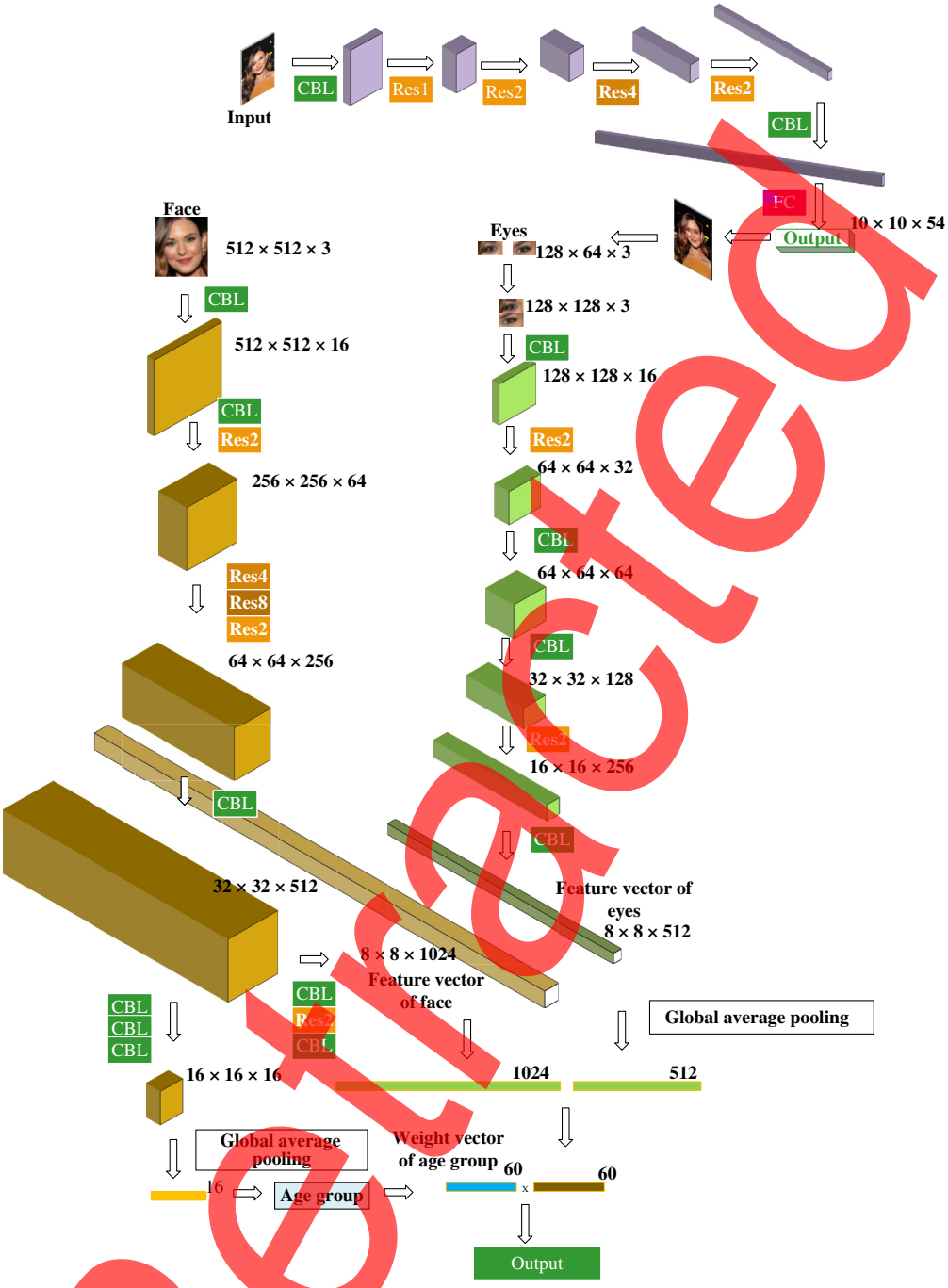

Fig. 8 Network model structure of MLFCNN.

the anchor box. E_i refers to whether the target center falls near the i'th anchor. The loss of the age group recognition and age estimation only includes classification loss, as shown in Eq. (4)

$$\text{Loss}_2 = -\sum_{i}^{n}\sum_{j}^{16} p_{ij}\,\log(q_{ij}) - \sum_{i}^{n}\sum_{j}^{60} p_{ij}\log(q_{ij}), \tag{4}$$

where n represents the number of faces contained in current image, p_{ij} represents the i'th sample belongs to the label of category j, it is 0 or 1. q_{ij} represents the probability that the network predicts that the i'th sample is classified as j. The overall loss function of the network is $\text{Loss}_1 + \text{Loss}_2$.

Fig. 9 Samples from FGNet dataset.

5 Tests and Results Analysis

5.1 *Datasets*

To enhance the diversity of dataset, the FGNet dataset[29] and the IMDB-WIKI dataset[15] are mainly used in this paper. The FGNet dataset was released in 2000 and was the first significant age dataset. It contains 1002 pictures of 82 people with an age from 0 to 69, which is currently one of the real age datasets that contain the youngest population. Figure 9 is some sample pictures of the FGNet dataset.

Published in 2015, the IMDB-WIKI dataset is composed of the IMDB database and the Wikipedia database. The IMDB face database contains 460,723 face images, whereas the Wikipedia face database contains 62,328 face images, a total of 523,051 face pictures. The dataset comes from pictures of celebrities crawled from IMDB and Wikipedia, which is currently the largest dataset for age and gender identification. And the age information is calculated based on the time stamp and date of birth of the photo. Figure 10 shows some sample images of the IMDB-WIKI dataset.

Annotation information of FGNet does not include the location information of eyes and faces, the annotation of IMDB-WIKI dataset contains the location information of faces, but does not contain the location information of eyes. This paper takes 2 months to annotate the dataset, and a total of 21,084 images of different age groups are labeled. And the training set and the test set are divided in ratio of 8:2. According to the network structure of MLFCNN, the network output contains only 60 categories. For samples in the dataset that are 60 years old or older, this paper regards them as the same category and uses a unified label for labeling. The hardware environment of the experiment is Intel Core i7-9900K CPU, NVIDIA Titan X GPU.

5.2 *Experiments and Analysis*

5.2.1 *Evaluation index of experiments*

This paper uses mean absolute error (MAE) and cumulative scores (CS) to measure the performance of the age estimation network. The calculation method of MAE is shown in Eq. (5)

Fig. 10 Samples from IMDB-WIKI dataset.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|P_i - L_i|, \tag{5}$$

where N is the total number of samples participating in the test, P_i is the result of the age estimation of the i'th sample by the network, and L_i represents the real age of the i'th sample picture. MAE can effectively reflect the average level of network age estimation. The calculation method of CS is shown in Eq. (6)

$$\mathrm{CS} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{bool}(|P_i - L_i| \le m), \tag{6}$$

where m means tolerable age error. CS represents the accuracy rate of network age estimation under the condition that the tolerable age error is m.

5.2.2 *Comparative experiments with advanced models*

To verify the performance of MLFCNN in age estimation, a set of comparative experiments was designed to compare MLFCNN with some advanced algorithms in the field of age estimation. Deeply learned feature[11] divided the age estimation network into a classifier and a repressor, and then fused the two models complementarily to obtain better performance. Structure-Based[24] proposed a structure-based age estimation network, which identified the gender of the samples initially, and used different age estimation methods for different genders to improve the accuracy. Multiscale[12] first divides the face image into different sizes, then divides multiple local aligned patches in it, and finally sends the resulting blocks to a four-layers CNN. Multi-region convolutional neural networks (MR-CNN)[10] believes that age is continuous, so that the number of neurons in the output layer is 1, thus obtaining the specific age value. The results of comparative experiments are shown in Fig. 11 and Table 2.

It can be seen from Table 2 that MLFCNN has an MAE of 2.87 in the mixed dataset test, which is the lowest among all tested networks. In addition, according to Fig. 11, MLFCNN can achieve 91.14% accuracy when tolerable age error is 4, and 98.32% accuracy when tolerable age error is 6, which is better than other networks.

5.2.3 *Ablation experiments*

The previous experiment shows that MLFCNN has some advantages in age estimation. MLFCNN incorporates the features of the eyes on the basis of the full-face features. To verify

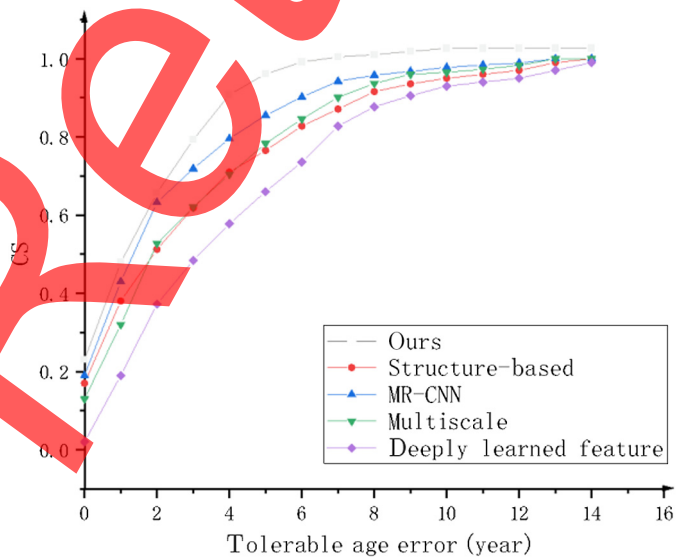

Fig. 11 A comparison with the current advanced age estimation methods for CS.

Table 2 A comparison with the current advanced age estimation methods for MAE.

No.	Age estimation methods	MAE
1	Deeply learned feature[11]	4.53
2	Structure-based[24]	3.86
3	Multiscale[12]	3.67
4	MR-CNN[10]	3.25
5	Ours	2.87

Table 3 Validation experiment on the importance of eye features.

No.	Age estimation methods	MAE
1	Face + Eyes MLFCNN	2.89
2	Face Only MLFCNN	3.18
3	No Eyes MLFCNN	3.72

whether the features of the eyes play an important role in the age estimation task, a set of comparison tests are designed. There are three network models involved in the comparison. The first model is a normal MLFCNN model, which uses full-face features and eye features, called "Face + Eyes MLFCNN." The second model uses only full-face features without eye features. The MLFCNN is called "Face Only MLFCNN." The third model uses a 0 vector to replace the eyes of the person in the input image, so that the model does not use any eye features when estimating the age, and is called "No Eyes MLFCNN." The experimental results are shown in Table 3.

MLFCNN will perform an age group estimation before accurate age estimation. To verify the influence of this mechanism on the effect of age estimation, we further designed a comparative experiment on the basis of the previous experiment. We removed the age group estimation network based on the normal MLFCNN, which is called MLFCNN-R. Combining different degrees of eye characteristics, six groups of experimental models are formed, namely: "Face + Eyes MLFCNN," "Face Only MLFCNN," "No Eyes MLFCNN," "Face + Eyes MLFCNN-R," "Face Only MLFCNN-R," and "No Eyes MLFCNN-R." The convergence of different networks in the training process and the accuracy of the network, when the allowable error is 5, are shown in Figs. 12 and 13, and the MAE of different networks is shown in Table 4.

Figures 12 and 13 show that eye features have a greater impact on the overall accuracy of the network, and full use of eye features can improve the performance of the network. The age group recognition network has a great influence on the convergence of the network and the final recognition rate. When the age group recognition is not performed, the network convergence is more difficult and the overall accuracy rate will decrease. Therefore, age characteristics and age group recognition networks are of great significance to MLFCNN and can improve the performance of the network.

5.2.4 *Small sample comparison experiment*

To verify the stability of MLFCNN, a set of small sample comparison experiments were designed, using 100%, 60%, 40%, 20%, and 10% of the existing training set to train MLFCNN. Compare the performance of the models obtained with different training datasets. The experimental results are shown in Table 5. Experimental results show that MLFCNN can still maintain a relatively stable performance when training samples are reduced. The stability and anti-interference of the network model are better than other network models participating in the experiment.

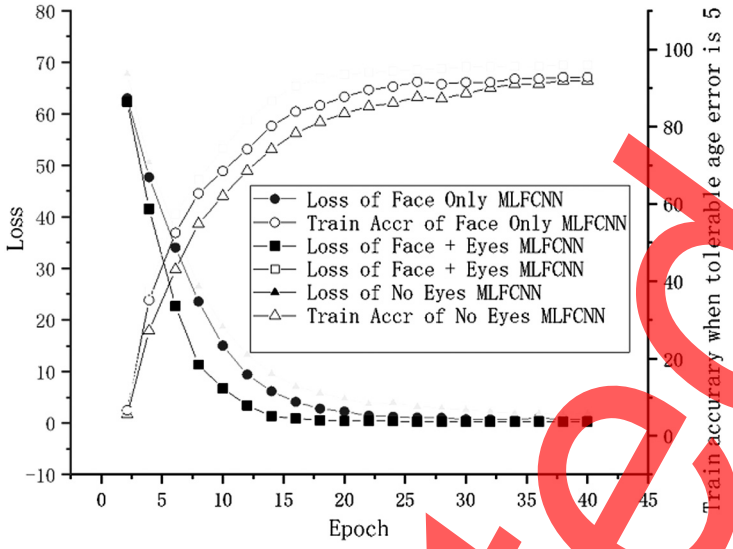

Fig. 12 Convergence and training accuracy of "Face + Eyes MLFCNN," "Face Only MLFCNN," and "No Eyes MLFCNN."

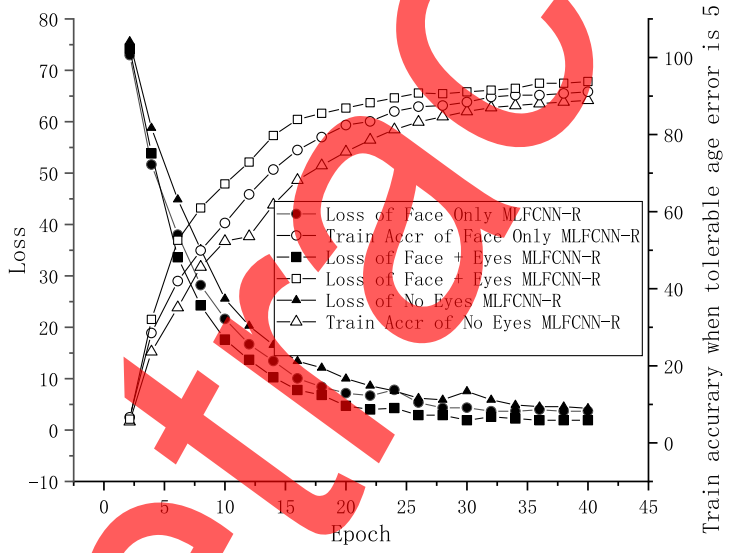

Fig. 13 Convergence and training accuracy of "Face + Eyes MLFCNN-R," "Face Only MLFCNN-R," and "No Eyes MLFCNN-R."

Table 4 MAE of different network combinations.

No.	Age estimation methods	MAE
1	Face + Eyes MLFCNN	2.84
2	Face Only MLFCNN	3.13
3	No Eyes MLFCNN	3.64
4	Face + Eyes MLFCNN-R	3.15
5	Face Only MLFCNN-R	3.83
6	No Eyes MLFCNN-R	5.98

Table 5 Validation experiment on the importance of eye features.

	The proportion of training samples in the training set				
	100%	60%	40%	20%	10%
Networks			MAE		
Multiscale[12]	3.66	Multiscale[12]	3.66	Multiscale[12]	3.66
MR-CNN[10]	3.24	MR-CNN[10]	3.24	MR-CNN[10]	3.24
Ours	2.83	Ours	2.83	Ours	2.83

6 Conclusion

In this paper, a human age estimation algorithm based on MLFCNN and face images is proposed, which values eye features and combines them with face features to jointly estimate human age. To reduce the error, MLFCNN adopts the combination of rough estimation and fine estimation, first estimating the age group of the sample, and then estimating the specific age. The experimental results show that the MAE of MLFCNN is 2.87, which is lower than other network models tested. When the allowable age error is 4 years, the estimated accuracy rate is as high as 91.14%, and 98.32% accuracy when tolerable age error is 6. In addition, it can still maintain a good performance under the small sample test. The robustness and anti-interference of the network are strong. And ablation experiments have verified the importance of ocular features and age recognition network for the age estimation task in this paper.

Acknowledgments

This work is supported by the Scientific Research Project of Hunan Provincial Department of Education (No. 18c1062). The authors declare that there is no conflict of interest regarding the publication of this paper.

Code, Data, and Materials Availability

The data included in this paper are available without any restriction.

References

1. Z. Hu et al., "Facial age estimation with age difference," *IEEE Trans. Image Process.* **26**(7), 3087–3097 (2017).
2. H. Liu et al., "Group-aware deep feature learning for facial age estimation," *Pattern Recognit.* **66**, 82–94 (2017).
3. A. Günay et al., "Age estimation based on AAM and 2D-DCT features of facial images," *Int. J. Comput. Sci. Appl.* **6**(2), 559–571 (2015).
4. J. Lu et al., "Cost-sensitive local binary feature learning for facial age estimation," *IEEE Trans. Image Process.* **24**(12), 5356–5368 (2015).
5. I. Huerta et al., "A deep analysis on age estimation," *Pattern Recognit. Lett.* **68**, 239–249 (2015).
6. A. Spizhevoi et al., "Estimating human age using bio-inspired features and the ranking method," *Pattern Recognit Image Anal.* **25**(3), 547–552 (2015).
7. Y. Zhu et al., "A study on apparent age estimation," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, pp. 25–31 (2015).
8. R. Meng et al., "A survey of image information hiding algorithms based on deep learning," *Comput. Model. Eng. Sci.* **117**(3), 425–454 (2018).

9. Y. Dong et al., "Automatic age estimation based on deep learning algorithm," *Neurocomputing* **187**, 4–10 (2016).
10. Z. Niu et al., "Ordinal regression with multiple output CNN forage estimation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 4920–4928 (2016).
11. X. Liu et al., "Agenet: deeply learned regressor and classifier for robust apparent age estimation," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, pp. 16–24 (2015).
12. D. Yi et al., "Age estimation by multi-scale convolutional network," *Lect. Notes Comput. Sci.* **9005**, 144–158 (2015).
13. R. Ranjan et al., "Unconstrained age estimation with deep convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, pp. 109–117 (2015).
14. J. C. Chen et al., "Unconstrained face verification using deep CNN features," in *Proc. IEEE Winter Conf. Appl. Comput. Vision*, IEEE, pp. 1–9 (2016).
15. R. Rothe et al., "DEX: deep expectation of apparent age from a single image," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, pp. 10–15 (2015).
16. E. Agustsson et al., "Apparent and real age estimation in still images with deep residual regressors on appa-real database," in *Proc. 12th IEEE Int. Conf. Autom. Face and Gesture Recognit.*, IEEE, pp. 87–94 (2017).
17. S. Feng et al., "Human facial age estimation by cost-sensitive label ranking and trace norm regularization," *IEEE Trans. Multimedia* **19**(1), 136–148 (2017).
18. K. Y. Chang et al., "A learning framework for age rank estimation based on face images with scattering transform," *IEEE Trans. Image Process.* **24**(3), 785–798 (2015).
19. S. Chen et al., "Using ranking-CNN for age estimation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 5183–5192 (2017).
20. G. Levi et al., "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. Workshops*, pp. 34–42 (2017).
21. L. Hou et al., "Squared earth mover's distance-based loss for training deep neural networks," arXiv preprint: 1611.05916 (2016).
22. Z. P. Ge et al., "Age estimation based on pulp cavity/chamber volume of 13 types of tooth from cone beam computed tomography images," *Int. J. Legal Med.* **130**(4), 1159–1167 (2016).
23. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2018).
24. K. H. Liu et al., "A structure-based human facial age estimation framework under a constrained condition," *IEEE Trans. Image Process.* **28**(10), 5187–5200 (2019).
25. S. Ren et al., "Faster R-CNN: towards real-time object detection with region proposal networks," in *Adv. Neural Inf. Process. Syst.*, pp. 91–99 (2015).
26. J. Redmon et al., "You only look once: unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 779–788 (2016).
27. W. Liu et al., "SSD: single shot multibox detector," *Lect. Notes Comput. Sci.* **9905**, 21–37 (2016).
28. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
29. A. Lanitis et al., "Toward automatic simulation of aging effects on face images," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(4), 442–455 (2002).

Tangtang Yi received her master's degree in computer application technology. She graduated from the School of Information Engineering, Xiangtan University in 2008. She is currently working at the School of Information Science and Engineering, Hunan Women's College. She has more than 10 years of research experience in the field of artificial intelligence and image processing. And she has published more than 10 academic papers in this field in peer-reviewed journals at home and abroad. She was a recipient of the International Association of Geomagnetism and Aeronomy Young Scientist Award for Excellence in 2008, and the IEEE Electromagnetic Compatibility Society Best Symposium Paper Award in 2011.