Neurocomputing 378 (2020) 45–53

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Eye localization based on weight binarization cascade convolution neural network

Zhen-Tao Liu[a,b,*], Si-Han Li[a,b], Min Wu[a,b], Wei-Hua Cao[a,b], Man Hao[a,b], Lin-Bo Xian[c]

[a] School of Automation, China University of Geosciences, Wuhan 430074, China
[b] Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China
[c] Wuhan WXYZ Technologies Co. Ltd., Wuhan 430200, China

ARTICLE INFO

Article history:
Received 30 June 2019
Revised 22 September 2019
Accepted 5 October 2019
Available online 18 October 2019

Communicated by Bo Shen

Keywords:
Eye localization
Convolution neural network
Binary
Image processing
Coarse-to-fine

ABSTRACT

Eye localization is a key step in the field of face recognition and analysis, which is the premise and breakthrough of drowsiness estimation and auxiliary driving. An eye localization method based on Weight Binarization Cascade Convolution Neural Network (WBCCNN) is proposed, in which the WBCCNN is composed of four levels and the weight is constrained by binarization. It predicts eye positions from coarse-to-fine to improve the performance of eye localization, and binary network not only helps to reduce the storage size of the model, but also speeds up the operation. Experiments on eye localization are performed using Labeled Faces in the Wild (LFW), BioID, and Labeled Face Parts in the Wild (LFPW) Databases, from which the results show that the average detection errors of left eye and right eye by our method are 0.66% and 0.71% on LFW, respectively. The operation speed of binary network is approximately as twice as that of non-binary. In addition, our method requires less storage capacity, which maintains higher performance on BioID and LFPW, compared to some state-of-the-art works.

1. Introduction

Eye localization is to locate the center of human eyes for a given face image, which is also a challenging task due to the influence of posture and occlusion. It plays a significant role in many scientific research and applications, such as drowsiness estimation [1,2], face pose correction [3], and expression recognition [4–6]. Therefore, how to acquire high-precision eye localization has become a hot research topic in the fields of computer vision and pattern recognition.

Several approaches for eye localization have been put forward in last ten years, which can be generally divided into two categories, i.e., knowledge-based approach and learning-based approach [7]. The former adapts the priori knowledge of face pattern to establish the rules, while the latter regards the face image as a vector and transforms the problem of eye localization into the problem of signal detection in high-dimensional space [8,9]. Knowledge-based approach extracts geometric shape, gray scale, texture, and other features, subsequently verifies whether they conform to the prior knowledge of facial pattern. Kadour et al. [10] used the distinct difference between facial skin color and background to make gray projection in horizontal and vertical directions, so as to detect the position of human eyes. Although the computational complexity of this method is low, it is not suitable for complex situations such as rapid change. Li et al. [11] adopted difference of gaussians instead of gray image to improve the robustness to illumination, detecting position of human eyes via a novel integral projection function and radial symmetry transform. Nevertheless, it is greatly influenced by posture, shelter, and other factors. In addition, Active Shape Models (ASM) [12] and Active Appearance Models (AAM) [13] are used as classical models for eye localization, which are on the basis of point distribution model. Both of them detect human eyes by combining texture information with shape model, and achieve matching of global shape model through iterative optimization of local texture information.

The complexity of face patterns makes it difficult to describe facial features using knowledge-based approach. Therefore, learning-based approach is attracting much attention recently, in which deep learning has been paid more attention [14,15]. Cascaded Pose Regression (CPR) method treats eye localization as a regression process from the appearance of the face to the shape of the face, and it returns to the optimal eye localization through continuous iteration [16]. Sun et al. [17] applied deep convolutional neural network (DCNN) to facial landmark detection. Zhou et al. [18] improved the DCNN model and achieved high-accuracy localization of 68 facial landmarks. Zhang et al. [19] applied multi-task learning (MTL) to facial landmark localization and proposed

[*] Corresponding author.
E-mail address: liuzhentao@cug.edu.cn (Z.-T. Liu).

https://doi.org/10.1016/j.neucom.2019.10.048
0925-2312/© 2019 Elsevier B.V. All rights reserved.

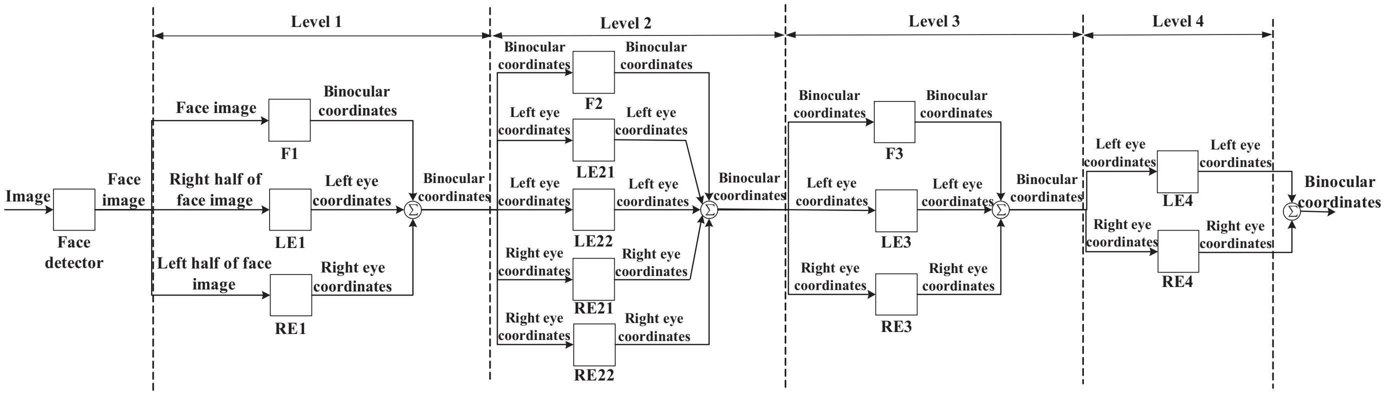


Fig. 1. Overall structure of WBCCNN for eye localization.

tasks-constrained deep convolutional network (TCDCN). Zhang et al. [20] proposed a multi-task cascaded convolutional networks (MTCNN) for simultaneously dealing with the problem of face detection and facial landmark localization. Wu et al. [21] proposed tweaked convolutional neural networks (TCNN) to study what features convolutional neural network learns from facial landmark localization tasks. Deep alignment network (DAN) was proposed in [22], which uses landmark heatmaps to increase the accuracy of landmark localization. Existing methods still can't achieve high accuracy and require high computational cost and memory capacity, since the general eye localization algorithms have limitations on eye localization with complex facial information and have complex network structure. In recent years, both the structure of the model and parameter setting become more esoteric to pursue higher accuracy. Following Occam's Razor principle, too complex model is inevitably difficult to be the optimal solution [23]. Therefore, we should simplify the model as much as possible while retaining comparative accuracy as latest researches.

To solve above problems, an eye localization method based on weight binarization cascade convolution neural network (WBCCNN) is proposed, which contains thirteen individual binary convolution neural networks with four levels. The weight of it is constrained to -1 or 1. This multiple-level structure is coarse-to-fine, which could increase the accuracy of eye localization effectively. Binary neural network not only helps to save the storage capacity of the model, but also reduces the computational cost. In addition, it can effectively localize human eyes even when low-level features from local regions are ambiguous or corrupted in challenging image examples. Experiments are performed on Labeled Faces in the Wild (LFW), the web data sets, BioID, and Labeled Face Parts in the Wild (LFPW) Databases. Our method achieves high performance both in accuracy and speed on LFW and the web test data sets. In addition, the average detection errors and the success rate of our method are superior performance on BioID and LFPW. The results demonstrate that our method can achieve higher accuracy compared with some state-of-the-art works.

The remainder of this paper is organized as follows. In Section 2, the framework of eye localization method based on WBCCNN is proposed. In Section 3, the analysis of WBCCNN is introduced. Experiments on human eye localization method and discussion are presented in Section 4.

2. Framework of WBCCNN for eye localization

An eye localization structure based on WBCCNN is constructed, as shown in Fig. 1. Firstly, a face detector such as libfacedetect is used to clip the images into face bounding boxes. Subsequently, they are converted to gray images which is the input of the whole binary cascaded convolution neural network. The binary cascaded convolutional neural network is divided into four levels, i.e., level 1, level 2, level 3, and level 4.

The squares present convolution neural network in Fig. 1. Level 1 contains three convolution neural networks F1, LE1, and RE1 (F1 means that the outputs of the convolution neural network are the binocular coordinates in level 1, LE1 means that the outputs of the convolution neural network are the left eye coordinates in level 1, RE1 means that the outputs of the convolution neural network are the right eye coordinates in level 1, and the following convolution neural networks are named in like manner). The inputs of F1 are the face images processed by the libfacedetect, and the outputs are the predictive binocular coordinates. The right half and the left half of the face images obtained by libfacedetect are clipped, and the images are inputs of LE1 and RE1, respectively. The outputs of LE1 and RE1 are predictive left eye and right eye coordinates, respectively. Then the outputs of LE1 and RE1 are superimposed, which plus the output of F1. And they are divided by 2 to get the predictive binocular coordinates of level 1. Finally, the smaller bounding boxes which are centered on the predictive coordinates of level 1 are clipped as the inputs of level 2.

Level 2 contains five convolution neural networks, i.e., F2, LE21, LE22, RE21, and RE22. The inputs of F2 are bounding boxes which are centered on the binocular coordinates predicted by the level 1, and it outputs more accurate binocular predictive coordinates. The inputs of LE21 and LE22 are bounding boxes which are centered on the left eye coordinates predicted by the level 1, and the outputs are more accurate left eye coordinates. The inputs of RE21 and RE22 are bounding boxes which are centered on the right eye coordinates predicted by the level 1, and the outputs are more accurate right eye coordinates. Similarly, the coordinates calculated are averaged and the predictive binocular coordinates of level 2 are obtained.

Level 3 contains three convolution neural networks, i.e., F3, LE3, and RE3. Similar to level 2, the inputs of level 3 are the smaller bounding boxes which are centered on the coordinates predicted by the level 2. The outputs of these convolutional neural networks are the predictive binocular coordinates, the left eye, and the right eye, respectively. The predictive binocular coordinates of third levels are obtained by superposing and averaging these coordinates again.

Level 4 contains two convolution neural networks, i.e., LE4 and RE4. Their inputs are bounding boxes centered on the predictive left eye coordinates of level 3 and right eye coordinates of level 3, and their outputs are the final predictive coordinates of left eye and right eye. Finally, human eye localization is calculated by superimposing these coordinates. By trial and error, it is shown that adding more level will not improve the accuracy observably, but increase the depth of the network and the cost of computation. Therefore, the levels of the network are set to 4.

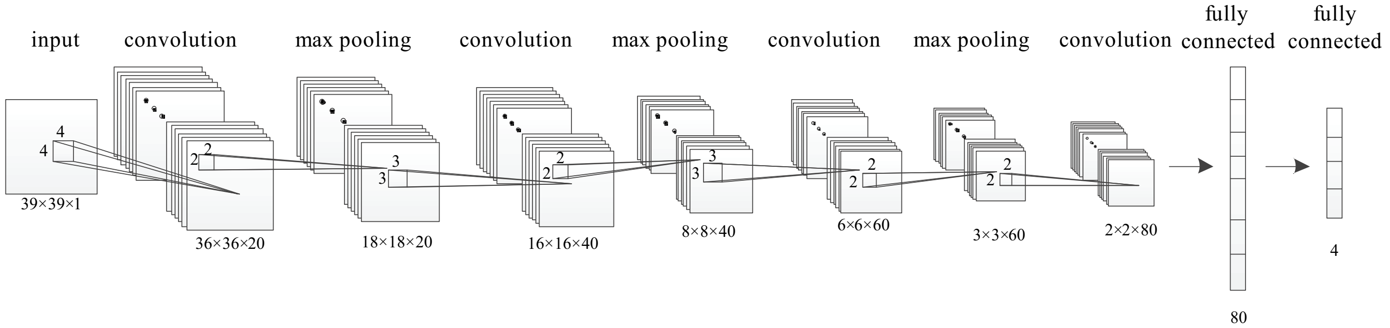


Fig. 2. Structure of F1 in WBCCNN.

This network structure adopts the localization idea from coarse-to-fine. A crude binocular coordinates are obtained by level 1. And then, search scope decreases, which makes it easier for the next level network to get more accurate coordinates. Similarly, level 2 and level 3 further reduce the search scope and the final predictive coordinates are calculated by level 4, which embodies the superiority of the cascade structure. Meanwhile, the weight of the whole network structure is constrained by binarization, i.e., the weight is either - 1 or 1. This substitution process runs through the whole forward and backward propagation. After network binarization, convolution can be expressed as simple addition and subtraction, and the computational cost can be greatly reduced.

3. Analysis of WBCCNN

Weight binarization convolutional neural network is a kind of multi-layer neural network, which is good at dealing with machine learning problems related to images, especially large images. And it can reduce dimensionality of image with large amounts of data. Convolutional neural network was proposed by LeCun and applied to handwriting recognition [24]. A typical convolution neural network includes convolution layer, pooling layer, and fully connected layer. The convolution layer cooperates with the pool layer to form several convolution groups, extracting features layer by layer. Subsequently, several fully connected layers follow on to do classification. Binary neural networks was proposed by Courbariaux et al. [25], and it was pointed out that the key of BinaryConnect algorithm is to binarize weights 1 or -1 only in forward and backward propagation. Courbariaux et al. [26] proposed a binary neural network model, which binarizes the weight and hidden layer activation simultaneously. The emergence of binary neural network accelerates the operation by reducing the size of the model and simplifying the computational difficulty.

3.1. Brief review of cascade convolutional network

Overall network contains 13 individual convolution neural networks, of which F1, F2, and F3 have the same number of layers, including 4 convolution layers, 3 pooling layers, and 2 fully connected layers. But their specific parameters are slightly different. LE1 and RE1 have the same structures, i.e., 9 layers. Level 1, F2, and F3 have deeper structures because they have larger input regions. It's a high-level task that eye localization from lager input regions. And it's benefit for forming high-level features that deeper structures are designed. Other networks in level 2, level 3, and level 4 have 6 layers, which are designed to extract local features without deeper layers. The specific structures are shown in Table 1.

Table 1
Individual network structure.

	F1	F2, F3	LE1, RE1	Others
layer0	I(39,39)	I(21,36)	I(39,27)	I(15,15)
layer1	C(4,20)	C(3,20)	C(4,20)	C(4,20)
layer2	P(2,20)	P(2,20)	P(2,20)	P(2,20)
layer3	C(3,40)	C(2,40)	C(3,40)	C(3,40)
layer4	P(2,40)	P(2,40)	P(2,40)	P(2,40)
layer5	C(3,60)	C(2,60)	C(3,60)	F(60)
layer6	P(2,60)	P(2,60)	P(2,60)	F(2)
layer7	C(2,80)	C(1,80)	C(1,80)	
layer8	F(80)	F(80)	F(80)	
layer9	F(4)	F(4)	F(2)	

In Table 1, I(p,q) denotes the input of the convolution neural network with the image of the height p and width q, C(i,j) denotes that the side length of the square convolution kernels in the convolution layer is $i \times i$, and the number of maps in convolution layers is j. Pooling layer is denoted by P(m,l), where m is the side length of square pooling regions and l is the number of maps in pooling layers. Fully connected layer is denoted by F(k), and the number of neurons is k.

The structure of F1 is given as an example in Fig. 2, which illustrates size of the input, the convolutions, the kernels, and the max poolings. The number of maps and neurons are illustrated as well.

In the convolution layers, outputs are calculated by convolution with the weight and shifted by the bias, followed by an activation function Rectified Linear Unit (ReLU). For different output maps and different regions in the maps, the set of kernels and the bias are different. Selection of initial weight and bias has influence on the localization effect and the initial weight is the random numbers by truncated normal distribution function. For neurons in the convolution layers, absolute value is corrected after activation function ReLU. Activation function ReLU can make calculation very fast, because neither function nor its derivatives contain complex mathematical operations [27].

Pooling layers use max pooling and pooling regions aren't overlapped. In addition, max pooling results are followed by an activation function ReLU. Fully connected layers contain the weight and the bias whose initial value is set in the same way as the convolution layers. And then, output value is obtained by an activation function ReLU which makes gradient descent more efficient and avoids the problem of vanishing gradient and exploding gradient.

3.2. Weight binarization

All weights in cascade convolutional neural network are binarized, including the weights of convolution layer and fully connected layer. In forward and backward propagation, the weights are binarized 1 or - 1, but the original weights are used when updating parameters. The main reason is that the gradient magnitude is usually very small, and when binary convolution kernel is updated directly, the derivative value is 0, which causes gradient disappearance. And the requirement of the accuracy of updating parameters is relatively high.

The operation of a convolution layer can be represented by $I{*}W$, where I represents input and W represents convolution kernel, and their dimensions are $c \times w \times h$, which are the number of channels, width and height of convolution kernel, respectively. Original convolution kernel W is replaced by the binary convolution kernel B and the scale parameter α, i.e.,

$$I * W \approx (I \oplus B)\alpha, \tag{1}$$

where $\oplus$ denotes convolution without multiplication and α defaults to a positive number.

To replace the original convolution kernel W with the scale parameter α and the binary convolution kernel B, the former should be equal to the latter as much as possible. Therefore, the following optimization is done.

$$J(B, \alpha) = ||W - \alpha B||^2, \tag{2}$$

$$\alpha^*, B^* = \text{argmin}_{\alpha, B} J(B, \alpha), \tag{3}$$

where $J(B, \alpha)$ is optimization objective function, and α^*, B^* are the values of the parameters when the function takes the minimum value. Eq. (2) can be expanded as

$$J(B, \alpha) = \alpha^2 B^T B - 2\alpha W^T B + W^T W, \tag{4}$$

where the values in B are either 1 or -1, therefore in a definite convolution layer, $B^T B = n = c \times w \times h$ is a constant. Because W is known, $W^T W$ is also a constant. In addition, α is a positive number. These constants can be removed from the optimization function without affecting the optimization results. The optimal solution for B can be calculated as

$$B^* = \text{argmax}_B \{W^T B\} \quad \text{s.t.}\, B \in \{+1, -1\}^n. \tag{5}$$

According to Eq. (5), the optimal value of B is the symbol of the value of W, i.e., $B^* = \text{sign}(W)$. Subsequently, finding the optimal value of α. The derivative of $J(B, \alpha)$ with respect to α is taken, which is set to be zero.

$$\alpha^* = \frac{W^T B^*}{n}, \tag{6}$$

the following for calculating α^* can be obtained by simple conversion and using the B^* calculated

$$\alpha^* = \frac{W^T \text{sign}(W)}{n} = \frac{\sum |W_i|}{n} = \frac{1}{n}||W||_{l1}, \tag{7}$$

where $||W||_{l1}$ is L1 norm, i.e., the optimal value of α is the average of absolute weight values.

Because all weights in cascade convolutional neural network are binarized, the storage size of the network can be saved and the operation speed of the network can be increased.

3.3. Coordinate transformation

In the WBCCNN, the size of the input images of different networks is different. Therefore, in the training phase, normalized coordinates must be obtained according to the range of the clipped images. The specific ranges of the clipped images (inputs of the networks) are illustrated in Table 2, in which left, right, top, and bottom are used to describe the ranges of the input of the networks. In the data set of training phase, the coordinates of the large bounding boxes containing the face are given. For networks F1, LE1, and RE1, the four boundary coordinates are relative to the normalized face bounding box with boundary coordinates (0, 1, 0, 1). But for F2 and F3, the four boundary coordinates are relative to the predicted coordinates of binocular geometric center. For the other networks, the four boundary coordinates are relative to the predicted coordinates of left eye or right eye. The specific boundary coordinates of the input can be formally expressed as

Table 2
Ranges of the input of the networks.

Network	left	right	top	bottom
F1	−0.05	+1.05	−0.05	+1.05
LE1	−0.03	+0.73	−0.05	+1.05
RE1	+0.27	+1.03	−0.05	+1.05
F2	−0.50	+0.50	−0.29	+0.29
LE21	−0.17	+0.17	−0.17	+0.17
LE22	−0.19	+0.19	−0.19	+0.19
RE21	−0.17	+0.17	−0.17	+0.17
RE22	−0.19	+0.19	−0.19	+0.19
F3	−0.46	+0.46	−0.27	+0.27
LE3	−0.13	+0.13	−0.13	+0.13
RE3	−0.13	+0.13	−0.13	+0.13
LE4	−0.10	+0.10	−0.10	+0.10
RE4	−0.10	+0.10	−0.10	+0.10

$$x_l = x + left \cdot width, \tag{8}$$

$$x_r = x + right \cdot width, \tag{9}$$

$$y_t = y + top \cdot height, \tag{10}$$

$$y_b = y + bottom \cdot height, \tag{11}$$

where x_l, x_r, y_t, y_b are the coordinates of the left boundary, the right boundary, the top boundary, and the bottom boundary, respectively. (x, y) are the coordinates of the left boundary and the top boundary of the large bounding boxes in F1, LE1 and RE1, the truth coordinates of binocular geometric center in F2 and F3, and the truth coordinates of left eye or right eye in the other networks. *left, right, top* and *bottom* are the values of left, right, top, and bottom in Table 2. In addition, *width* and *height* denote the width and the height of the large bounding boxes which are given in the data set.

To avoid the coordinates of the human eye being always in the center of the clipped picture in level 2, level 3, and level 4, training patches centered at coordinates randomly shifted from the truth coordinates which are labeled in the training set are taken. The shift in vertical and horizontal directions doesn't exceed 0.5 in level 2, and 0.2 in level 3 and level 4. The new boundary coordinates are expressed as

$$x_{nl} = x_l + random \cdot left \cdot width, \tag{12}$$

$$x_{nr} = x_r + random \cdot right \cdot width, \tag{13}$$

$$y_{nt} = y_t + random \cdot top \cdot height, \tag{14}$$

$$y_{nb} = y_b + random \cdot bottom \cdot height, \tag{15}$$

where x_{nl}, x_{nr}, y_{nt}, y_{nb} are the new coordinates of the left boundary, the right boundary, the top boundary, and the bottom boundary, which are randomly shifted. *random* is an random number from *shift*|*left* to *shift*|*right*. *shift* is the maximum shift.

After clipping the images, normalized coordinates are calculated as

$$x_n = (x_{raw} - x_l)/[(right - left) \cdot width] \cdot W, \tag{16}$$

$$y_n = (y_{raw} - y_t)/[(bottom - top) \cdot height] \cdot H, \tag{17}$$

where (x_n, y_n) and (x_{raw}, y_{raw}) are the normalized coordinates and the raw coordinates, i.e., ground-truth coordinates. x_l and y_t are the coordinates of the left and the top boundary or the new coordinates of the left and the top boundary. W and H denote the width and the height of the input images which are illustrated in Table 1. In the training phase, network parameters are learnt by stochastic gradient descent.

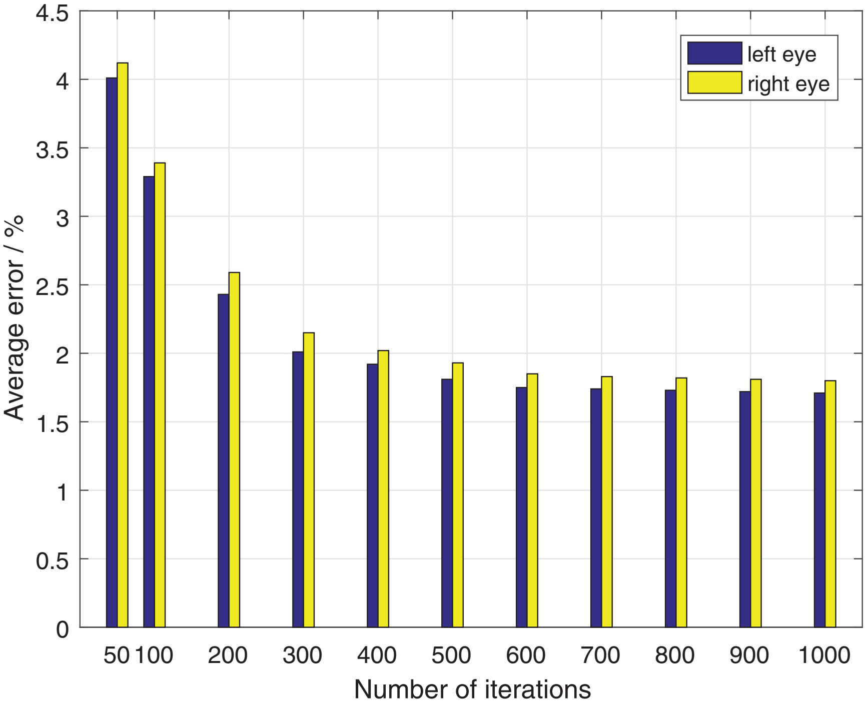


Fig. 3. Average detection errors under the different number of iterations in F1.

4. Experiments

Several experiments are performed in this section. In 4.1, the experimental settings are introduced. Subsequently, we compare the performance of using the different number of iterations, the performance of each level, and the performance of binary and non-binary networks in 4.2. The performance using some state-of-the-art methods and commercial software are presented in 4.3.

4.1. Experimental setting

Experiments on eye localization are performed on a computer with 64 bits system, Intel Core i5 CPU processor, and NVIDIA 1050TI GPU. All the experiments have two phases, i.e., learning phase and testing phase. The data set is divided into two parts, i.e., training data set and testing data set, which are provided by Labeled Faces in the Wild (LFW) [28], and the web [17].

LFW contains more than 13,000 images of human face, which is a international authoritative face recognition database. 1680 of the people pictured have two or more distinct photos in the data set. Each image has been labeled with the name of the person and the coordinates of the human eyes and the large bounding boxes containing the face. 5590 images in the data set are selected as part of the training set.

Together with LFW, 7876 images which are downloaded from the web constitute the data set used in the experiment, 13,466 images in total. The images are also labeled with the coordinates of the human eyes and the large bounding boxes containing the face by Sun et al. [17]. Subsequently, 10,000 images are selected for training, and the remaining 3466 images for testing.

We measure the performance by the average detection error of each binocular coordinates, which can indicate the accuracy of our method. The detection error is measured as

$$err=\sqrt{(x-x')^2+(y-y')^2}/l, \tag{18}$$

where (x, y) and (x', y') are the ground-truth coordinates and the predicted coordinates, and l is the width of the large bounding box given in the data set. In the training phase, a loss function is defined as

$$loss=[(x-x')^2+(y-y')^2]/2, \tag{19}$$

which measures the difference between the predicted coordinates and the ground-truth coordinates.

4.2. Comparison of different number of iterations, levels, and binarization

Before the training, all weights are initialized with different random numbers. Truncated normal distribution function sets the initial value of the weight. In general, the greater the number of iterations is, the higher the accuracy will be. To explore the relationship between the number of iterations and the accuracy, we compare the average detection error using different number of iterations in F1, as shown in Fig. 3, in which the average detection error of the binocular coordinates is described. It shows that when the number of iterations is less than 500, the accuracy increases as the number of iterations increases. When the number of iterations is larger than 500, however, the improvement of accuracy is not obvious as the number of iterations increases. Therefore, to balance the performance and the efficiency, the number of iterations is set to be 1000, which means that the each image of the 10,000 input images is used 1000 times. And the batch size is 16, so that a total of 625,000 batches are run.

To demonstrate the superiority of WBCCNN in accuracy, the performance of each convolution neural network is compared in Fig. 4 and the performance of each level is compared in Fig. 5, from which it can be seen that the average errors of those networks in high level are generally low. The average errors of the binocular coordinates in level 1 are 1.45% and 1.49%, respectively. The errors are greatly reduced in the level 2, which are 0.96% and 1.01%, respectively. In level 3 and level 4, however, the reduction of the errors is not obvious. The final detection errors are 0.66% and 0.71%, respectively. In view of the results mentioned above, we can draw that the cascade method can effectively reduce the

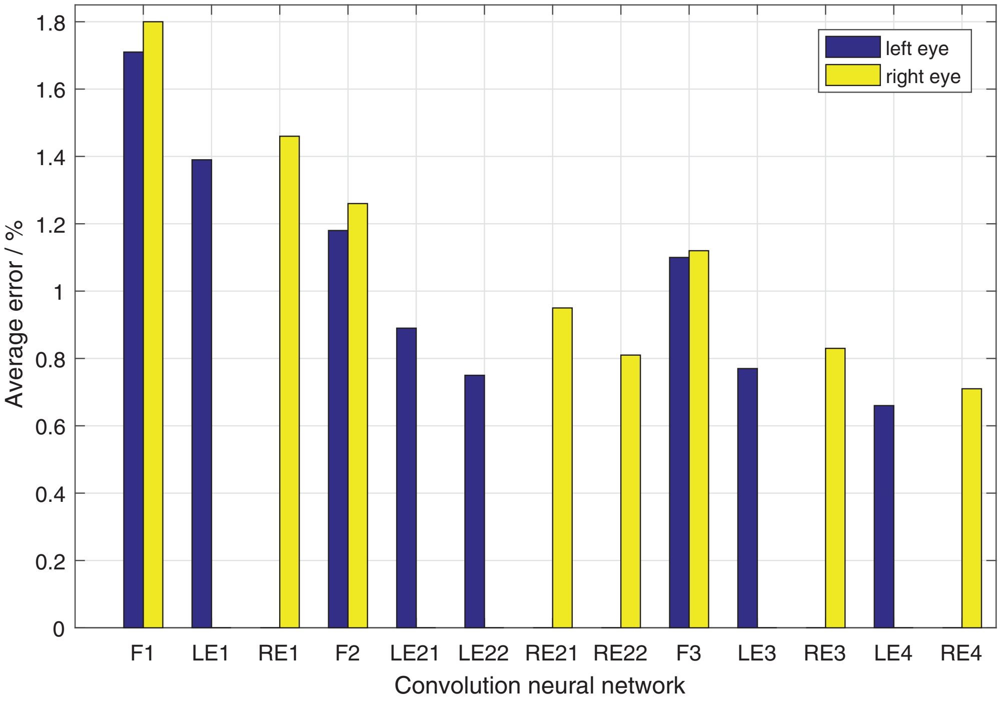


Fig. 4. Average detection errors of the each network.

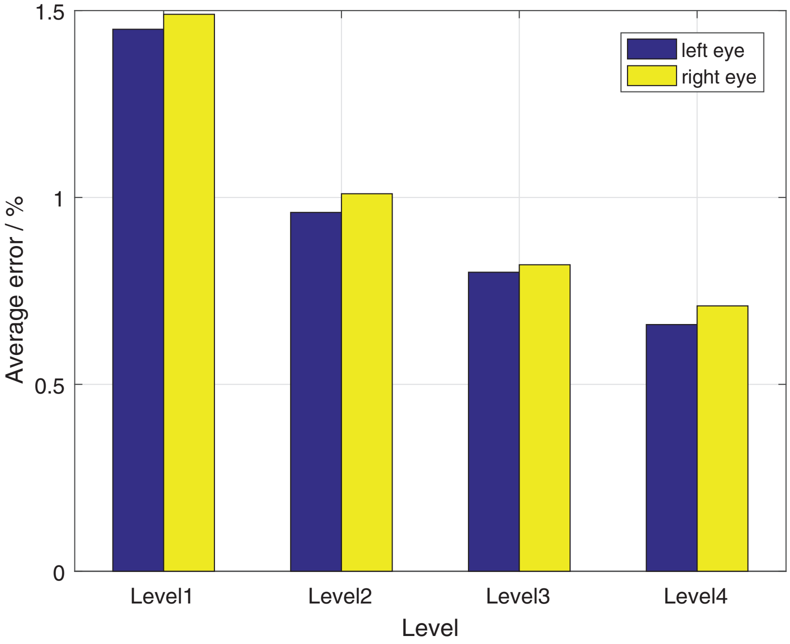


Fig. 5. Average detection errors of the each level.

errors because the range of the images in high level is more accurate.

To demonstrate the superiority of WBCCNN in operation speed and storage capacity, we compare the performance of binary and non-binary networks. The operation speed and storage capacity of F1, and the final detection errors are shown in Table 3. The operation speed of binary F1 is approximately as twice as that of non-binary F1. At the same time, binary F1 requires less storage capacity. In addition, the detection errors of binary and non-binary networks are not much difference. Therefore, WBCCNN whose weights are binarized, can be excellent in terms of operation speed and storage capacity.

4.3. Comparison with some state-of-the-Art works and discussion

We compare with a deep convolutional network cascade for facial point detection method which was proposed by Sun et al. [17] on LFW and the web data set. It used deep convolutional network cascade to detect facial point, which was also tested on LFW and the web data set. The average errors of the binocular coordinates of the two methods are shown in Table 4.

Table 4 shows that the average detection errors of our method is definitely less than those of [17]. Furthermore, the average errors of our method is 0.66% and 0.71%, respectively, and it decreases approximately 27% and 18% compared with those of [17].

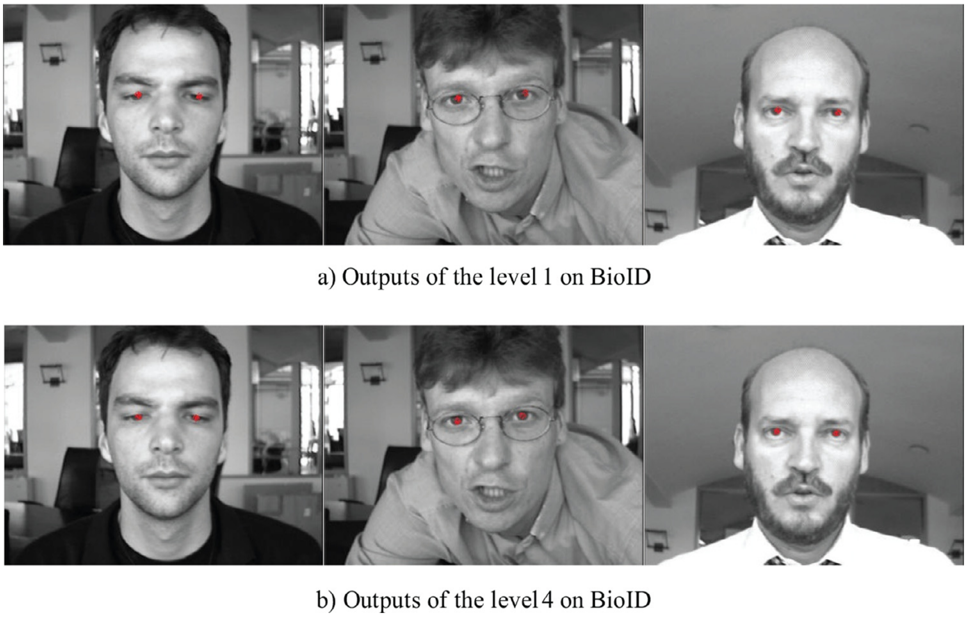

a) Outputs of the level 1 on BioID
b) Outputs of the level4 on BioID

Fig. 6. Examples of eye localization in level 1 and level 4 on BioID.

Table 3
Comparison between binary and non-binary networks on LFW and the web data set.

	Binary	Non-binary
Running time(F1)	10.98s	20.79s
Storage capacity(F1)	2.86M	5.95M
Average error of left eye	0.66%	0.49%
Average error of right eye	0.71%	0.46%

Table 4
Comparison of average error on LFW and the web data set.

Method	Average error of left eye	Average error of right eye
Our method	0.66%	0.71%
[17]	0.90%	0.87%

Table 5
Comparison of average error on BioID.

Method	Average error of left eye (%)	Average error of right eye (%)
Our method	1.88	2.06
[30]	7.10	7.80
[31]	3.90	3.70
Luxand	3.80	3.40

To demonstrate the competitiveness of our method, more related works and commercial softwares are compared with our method on BioID Face Database [29]. Because most of previous works used the binocular distance to normalize detection errors, we normalize the errors in the same way. BioID Face Database consists of 1521 gray level images with a resolution of 384 × 286 pixel, in which each image presents the frontal view of a face of one out of 23 different test persons. The coordinates of the human eyes and the large bounding boxes containing the face are labeled. Some competitive and commercial software containing Boosted Regression with Markov Networks, Component based Discriminative Search and Luxand Face SDK [30,31]. Comparison of average error on BioID are shown in Table 5, which shows that the average errors of our method is obviously lower than those of other methods, i.e., 1.88% and 2.06%.

Furthermore, we define the success rate as the proportion of cases whose normalized errors are less than 10%. To be consistent with most previous works, we calculate the maximum in the errors of the left eye and the right eye only. Comparison with ten state-of-the-art methods on BioID Face Database is given in Table 6, which shows that our method gains the best results of success rate, i.e., 96.58%. In addition, some examples of eye localization are presented in Fig. 6, in which the red dots in the images in the first row are the outputs of the level 1 and those in the second row are the outputs of the level 4.

Table 6
Comparison of success rate on BioID.

Literature	Method	Success rate (%)
[32]	Differential Geometry and Local Self-Similarity Matching	87.31
[33]	Invariant Isocentric Patterns	91.67
[34]	Means of Gradients	93.40
[35]	Face Detection and CDF Analysis	86.00
[36]	Several Base-Line Eye Localizers and Face Matching Algorithms	87.00
[37]	General-to-Specific Model Definition	85.20
[38]	2D Cascaded Adaboost	93.00
[39]	Feature-Based Affine-Invariant	76.00
[40]	Multi-Stage Approach	96.00
[41]	Isophote and Gradient Features	93.68
–	Our method	99.47

To make our method more challenging, we compare with more state-of-the-art methods on Labeled Face Parts in the Wild (LFPW) Dataset [42]. LFPW contains 1432 face images downloaded from the web and it is divided into two parts, i.e., 1132 training images and 300 testing images. LFPW is a multi-pose and multi-view facial fiducial point dataset, which includes images affected by various pose, expressions, illumination, and other factors. Because it just offers image URLs and some links have been invalid, we only download 781 training images and 249 testing images, which are all our test images. In Fig. 7, we present a comparison with the methods introduced in [43] and [44] by average error. Furthermore, we compare with the methods introduced in [45] and [46] by success rate in Table 7. From these results, we can see that WBCCNN maintains a high competitiveness in accuracy.

Both the average error and success rate of eye localization method based on WBCCNN are superior to above state-of-the-art works, since it has multiple-level structure of coarse-to-fine, which can gradually reduce the interference of external background factors. In addition, it can effectively localize human eyes even

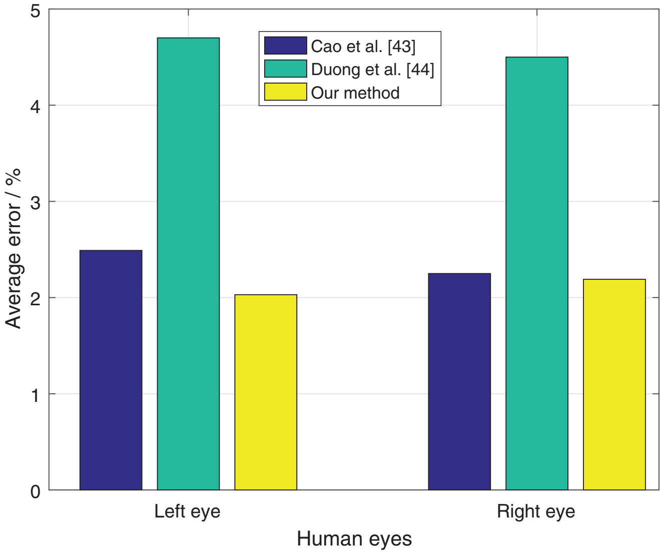


Fig. 7. Comparison of average error on LFPW.

Table 7
Comparison of success rate on LFPW.

Literature	Method	Success rate (%)
[45]	Deep feature learning	96.85
[46]	Inner product detector	84.80
–	Our method	99.03

when low-level features from local regions are ambiguous or corrupted. Besides, WBCCNN has superiority in operation speed and storage capacity due to the binary network.

5. Conclusion

An eye localization method based on WBCCNN was proposed, which was comprised of four levels containing thirteen individual binary convolution networks. Experiments using LFW, the web, BioID, LFPW data sets were performed. Average detection error of each binocular coordinates was the criteria for testing accuracy. Level 1 of cascade convolution neural network provided highly robust initial estimations. Level 2 improved the accuracy significantly, which played a role of precision adjustment. Final results were obtained by fine-tuning at level 3 and level 4, which were 0.66% and 0.71% on LFW and the web data set, respectively. In addition, the weights in network were binarized which simplified the computing process, optimized the operation speed and the storage size, e.g., operation speed of binary F1 was approximately as twice as that of non-binary F1. Furthermore, WBCCNN achieved higher accuracy compared with some state-of-the-art methods on BioID and LFPW Database.

In the future, to further improve the efficiency of the algorithm, binarizing both of inputs and weights will be considered. Reliable eye localization can greatly improve the accuracy of drowsiness estimation, therefore, WBCCNN is a promising choice for drowsiness estimation. Recently, most of traditional type of service robots are implemented by users instruction. People hope that robots have the ability to perceive human emotions [47,48], intentions, and serve people initiatively. The proposed eye localization method will be being applied to initiative service of robots for drowsiness estimation to help people to increase the work efficiency [49].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 61976197, 61403422, and 61273102, the Hubei Provincial Natural Science Foundation of China under Grants 2018CFB447 and 2015CFA010, the Wuhan Science and Technology Project under Grant 2017010201010133, the 111 project under Grant B17040, and the Fundamental Research Funds for National University, China University of Geosciences (Wuhan) under Grant 1810491T07.

References

[1] B. Cyganek, S. Gruszczyski, Hybrid computer vision system for drivers' eye recognition and fatigue monitoring, Neurocomputing 126 (2014) 78–94.
[2] Z.T. Liu, S.H. Li, W.H.C.e. al., An initiative service method with regard to degree of sleepiness for drinking service robot, in: Proceedings of the 37th Chinese Control Conference, 2018, pp. 5487–5491.
[3] Y. Tan, J. Yang, Y.G.Z.e. al., Face recognition with pose variations and misalignment via orthogonal procrustes regression, IEEE Trans. Image Process. 25 (6) (2016) 2673–2683.
[4] Y. Liang, B. Liu, J.X.e. al., Decoding facial expressions based on face-selective and motion-sensitive areas, Hum. Brain Mapp. 38 (6) (2016) 3113–3125.
[5] K. Lee, E.C. Lee, Comparison of facial expression recognition performance according to the use of depth information of structured-light type RGB-d camera, J. Ambient Intell. Hum. Comput. (2019), doi:10.1007/s12652-019-01278-2.
[6] N.Y. Zeng, H. Zhang, B.Y.S.e. al., Facial expression recognition via learning deep sparse autoencoders, Neurocomputing 273 (2018) 643–649.
[7] M. Lopar, S. Ribaric, An overview and evaluation of various face and eyes detection algorithms for driver fatigue monitoring systems, in: Proceedings of the Croatian Computer Vision Workshop, 2013, pp. 15–18.
[8] Z.T. Liu, S.H. Li, W.H.C.e. al., Combining 2D Gabor and local binary pattern for facial expression recognition using extreme learning machine, J. Adv. Comput. Intell. Intell. Inform. 23 (3) (2019) 444–455.
[9] Z.H. Feng, J. Kittler, X.J. Wu, Mining hard augmented samples for robust facial landmark localization with CNNs, IEEE Signal Process. Lett. 26 (3) (2019) 450–454.
[10] S. Kadour, M.D. Levine, Face detection in gray scale images using locally linear embeddings, Comput. Vis. Image Underst. 105 (1) (2007) 1–20.
[11] W.H. Li, Y. Wang, Y. Wang, Eye location via a novel integral projection function and radial symmetry transform, Int. J. Digit. Content Technol. Appl. 5 (8) (2011) 70–80.
[12] T.F. Cootes, C.J. Twining, K.O.B.e. al., Diffeomorphic statistical shape models, Image Vis. Comput. 26 (3) (2008) 326–332.
[13] D. Cristinacce, T.F. Cootes, Feature detection and tracking with constrained local models, Pattern Recognit. 41 (10) (2008) 3054–3067.
[14] N.Y. Zeng, Z.D. Wang, H.Z.e. al., Deep belief networks for quantitative analysis of gold immunochromatographic strip, Cognit. Comput. 8 (4) (2016) 684–692.
[15] N.Y. Zeng, H. Qiu, Z.D.W.e. al., A new switching-delayed-PSO-based optimized SVM algorithm for diagnosis of Alzheimer's disease, Neurocomputing 320 (2018) 195–202.
[16] P. Dollar, P. Welinder, P. Perona, Cascaded pose regression, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 1078–1085.
[17] Y. Sun, X.G. Wang, X.O. Tang, Deep convolutional network cascade for facial point detection, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2013), 2013, pp. 3476–3483.
[18] E. Zhou, H. Fan, Z.C.e. al., Extensive facial landmark localization with coarse–to-fine convolutional network cascade, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2014, pp. 386–391.
[19] Z. Zhang, P. Luo, C.L.C.e. al., Facial landmark detection by deep multi-task learning, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 94–108.
[20] K. Zhang, Z. Zhang, Z.L.e. al., Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process Lett. 23 (10) (2016) 1499–1503.
[21] Y. Wu, T. Hassner, K.K.e. al., Facial landmark detection with tweaked convolutional neural networks, IEEE Trans. Pattern Anal. Mach. Intell. 40 (12) (2018) 3067–3074.
[22] M. Kowalski, J. Naruniec, T. Trzcinski, Deep alignment network: a convolutional neural network for robust face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 2034–2043.
[23] A. Blumer, Occam'S razor, Inf. Process Lett. 24 (6) (1987) 377–380.
[24] Y. Bengio, Y. Lecun, C.N.e. al., Lerec: a NN/HMM hybrid for on-line handwriting recognition, Neural Comput. 7 (6) (1995) 1289–1303.
[25] M. Courbariaux, Y. Bengio, J.P. David, Binaryconnect: training deep neural networks with binary weights during propagations, 2015, ArXiv:1511.00363.
[26] M. Courbariaux, I. Hubara, D.S.e. al., Binarized neural networks: training deep neural networks with weights and activations constrained to +1 or −1, 2016, arxiv:1602.02830.

[27] R.H. Hahnloser, R. Sarpeshkar, M.A.M.e. al., Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit, Nature 405 (6789) (2000) 947–951.

[28] G.B. Huang, M. Mattar, T.B.e. al., Labeled faces in the wild: a database for studying face recognition in unconstrained enviroments, in: Proceedings of the Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, 2008.

[29] B.F. Database, Dataset for face detection, Facedb - BioID, https://www.bioid.com/facedb/.

[30] M. Valstar, B. Martinez, X. Binefa, et al., Facial point detection using boosted regression and graph models, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2010), 2010, pp. 2729–2736.

[31] L. Liang, R. Xiao, F. Wen, et al., Face alignment via component-based discriminative search, in: Proceedings of the European Conference on Computer Vision(2008), 2008, pp. 72–85.

[32] M. Leo, D. Cazzato, T. DeMarco, C.D.e. al., Unsupervised eye pupil localization through differential geometry and local self-similarity matching, PLoS ONE 9 (2014).

[33] R. Valenti, T. Gevers, Accurate eye center location through invariant isocentric patterns, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1785–1798.

[34] F. Timm, E. Barth, Accurate eye centre localisation by means of gradients, in: Proceedings of the International Conference on Computer Vision Theory and Applications, 2011.

[35] M. Asadifard, J. Shanbezadeh, Automatic adaptive center of pupil detection using face detection and CDF analysis, in: Proceedings of the International MultiConference of Engineers and Computer Scientists, 2010.

[36] B. Kroon, A. Hanjalic, S.M. Maas, Eye localization for face matching: is it always useful and under what conditions? in: Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval, 2008, pp. 379–388.

[37] P. Campadelli, R. Lanzarotti, G. Lipori, Precise eye localization through a general-to-specific model definition, in: Proceedings of the British Machine Vision Conference, 2006, pp. 187–196.

[38] Z. Niu, S. Shan, S.Y.e. al., 2D cascaded adaboost for eye localization, in: Proceedings of the 18th International Conference on Pattern Recognition, 2006, pp. 1216–1219.

[39] M. Hamouz, J. Kittler, J.K.K.e. al., Feature-based affine-invariant localization of faces, Proceedings of the IEEE Transactions on Pattern Analysis & Machine Intelligence 27 (2005) 1490–1495.

[40] D. Cristinacce, T.F. Cootes, I.M. Scott, A multi-stage approach to facial feature detection, in: Proceedings of the British Machine Vision Conference, 2004, pp. 1–10.

[41] W. Zhang, M.L. Smith, L.N.S.e. al., Eye center localization and gaze gesture recognition for human-computer interaction, J. Opt. Soc. Am. A 33 (3) (2016) 314–325.

[42] P.N. Belhumeur, D.W. Jacobs, D.J.K.e. al., Localizing parts of faces using a consensus of exemplars, IEEE Trans. Pattern Anal Mach. Intell. 35 (12) (2013) 2930–2940.

[43] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, Int. J. Comput. Vis. 107 (2) (2014) 177–190.

[44] C.N. Duong, K. Luu, K.G.Q.e. al., Deep appearance models: a deep Boltzmann machine approach for face modeling, Int. J. Comput. Vis. 127 (5) (2019) 437–455.

[45] Y. Wu, Q. Ji, Learning the deep features for eye detection in uncontrolled conditions, in: Proceedings of the IEEE International Conference on Pattern Recognition, 2014, pp. 455–459.

[46] G.M. Araujo, F.M.L. Ribeiro, W.S.J.e. al., Weak classifier for density estimation in eye localization and tracking, IEEE Trans. Image Process. 26 (7) (2017) 3410–3424.

[47] Z.T. Liu, Q. Xie, M.W.e. al., Speech emotion recognition based on an improved brain emotion learning model, Neurocomputing 309 (2018) 145–156.

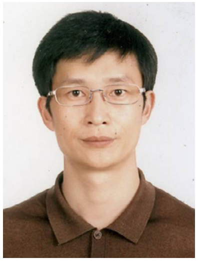
[48] Z.T. Liu, M. Wu, W.H.C.e. al., Speech emotion recognition based on feature selection and extreme learning machine decision tree, Neurocomputing 273 (2018) 271–280.

[49] M. Hao, W.H. Cao, M.W.e. al., Proposal of initiative service model for service robot, CAAI Trans. Intell. Technol. 2 (4) (2017) 253–261.

Zhen-Tao Liu received the B.E. and M.E. degrees from Central South University, Changsha, China, in 2004 and 2008, respectively, and Dr. E. degree from Tokyo Institute of Technology, Tokyo, Japan, in 2013. From 2013 to 2014, he was with Central South University, Changsha, China. Since Sept. 2014, he has been with School of Automation, China University of Geosciences, Wuhan, China. His research interests include affective computing, fuzzy systems, and intelligent robot. He is a member of IEEE-IES (Industrial Electronics Society, Institute of Electrical and Electronics Engineers) Technical Committee on Human Factors, CAAI (Chinese Association for Artificial Intelligence) Technical Committee on Intelligent Service, and SOFT (Japan Society for Fuzzy Theory and Systems). He is an Associate Editor of Int. J. of Advanced Computational Intelligence and Intelligent Informatics. He received Best Paper Awards of Int. J. of Advanced Computational Intelligence and Intelligent Informatics in 2018 and 2017, respectively, Best Presentation Award in IWACIII2017, Young Researcher Award of Int. J. of Advanced Computational Intelligence and Intelligent Informatics in 2014, Best Paper Award in ASPIRE League Symposium 2012, Excellent Presentation Award in IWACIII2009, and Zhang Zhongjun Best Paper Nomination Award in CPCC 2009.

Si-Han Li received the B.E. degree from Wuhan Polytechnic University, Wuhan, China in 2017. He is currently pursuing the master's degree with the School of Automation, China University of Geosciences. His current research interests include eye localization, initiative service robot, and human-robot interaction system. He is a student member of Chinese Association for Artificial Intelligence.

Min Wu received his B.S. and M.S. degrees in engineering from Central South University, Changsha, China, in 1983 and 1986, respectively, and his Ph.D. degree in engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1999. He was a faculty member of the School of Information Science and Engineering at Central South University from 1986 to 2014, and was promoted to Professor in 1994. In 2014, he moved to China University of Geosciences, Wuhan, China, where he is a professor in the School of Automation. He was a visiting scholar with the Department of Electrical Engineering, Tohoku University, Sendai, Japan, from 1989 to 1990, and a visiting research scholar with the Department of Control and Systems Engineering, Tokyo Institute of Technology, from 1996 to 1999. He was a visiting professor at the School of Mechanical, Materials, Manufacturing Engineering and Management, University of Nottingham, Nottingham, UK, from 2001 to 2002. His current research interests include process control, robust control, and intelligent systems. Dr. Wu is a member of the Chinese Association of Automation, and a fellow of IEEE. He received the IFAC Control Engineering Practice Prize Paper Award in 1999 (together with M. Nakano and J. She).



Wei-Hua Cao received his B.S., M.S., and Ph.D. degrees in Engineering from Central South University, Changsha, China, in 1994, 1997, and 2007, respectively. He is a Professor in the School of Automation, China University of Geosciences. He was a visiting scholar with the Department of Electrical Engineering, Alberta University, Edmonton, Canada, from 2007 to 2008. His research interest covers intelligent control and process control.

Man Hao received the B.E. degree from China University of Geosciences, Wuhan, China in 2016. She is currently a Ph.D. student in School of Automation, China University of Geosciences. Her current research interests include initiative service robot, emotion recognition, and human-robot interaction system. She is a student member of Chinese Association for Artificial Intelligence.

Lin-Bo Xian received the B.E. and M.E. degrees from Huazhong University of Science and Technology, Wuhan, China, in 2004 and 2007, respectively. From 2007 to 2009, he was at the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore, and worked as a researcher at the Robot Technology Research Center. In 2009, he founded Wuhan WXYZ Technologies Co. Ltd., Wuhan, China. His research interests include intelligent robot, commercial service robot, and education robot. Mr. Xian is the secretary-general of the Strategic Alliance for Robot Innovation in Hubei Province.