This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2021.3064280, IEEE Transactions on Cognitive and Developmental Systems

Attention Mechanism based Real-Time Gaze Tracking in Natural Scenes with Residual Blocks

Lihong Dai, Jinguo Liu, *Senior Member, IEEE*, Zhaojie Ju, *Senior Member, IEEE*, and Yang Gao, *Senior Member, IEEE*

***Abstract*—Gaze tracking is widely used in fatigue driving detection, eye disease diagnosis, mental illness diagnosis, website or advertising design, virtual reality, gaze-control devices and human-computer interaction. However, the influence of light, specular reflection and occlusion, the change of head pose, especially the ever-changing human pose in natural scenes, have brought great challenges to the accurate gaze tracking. In this paper, gaze tracking in natural scenes is studied, and a method based on Convolutional Neural Network (CNN) with residual blocks is proposed, in which attention mechanism is integrated into the network to improve the accuracy of gaze tracking. Furthermore, it is tested on the GazeFollow database which contains six kinds of databases. The results show that the performance of proposed method outperforms that of other state-of-the-art methods in natural scenes. Moreover, the proposed method has better real-time performance and is more suitable for practical applications.**

***Index Terms*—Gaze tracking, attention mechanism, residual blocks, CNN**

Manuscript received; revised; accepted.

Date of publication; date of current version.

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1304600, in part by the Natural Science Foundation of China (Grant 51775541,51575412,52075530), in part by the CAS Interdisciplinary Innovation Team under Grant JCTD-2018-11, and in part by the AiBle project co-financed by the European Regional Development Fund. *(Corresponding author: Jinguo Liu, Zhaojie Ju.)*

L. Dai is with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, and also with institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China. Moreover, L. Dai is with University of the Chinese Academy of Sciences, Beijing 100049, and also with School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: dailihong2004@163.com).

J. Liu is with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, and also with institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China (e-mail: liujinguo@sia.cn).

Z. Ju is with School of Computing, University of Portsmouth, Portsmouth P01 3HE, U.K. (e-mail: zhaojie.ju@port.ac.uk).

Y. Gao is with Space Technology for Autonomous and Robotic Systems Laboratory (STAR LAB), Surrey Space Centre, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: yang.gao@surrey.ac.uk).

I. Introduction

Gaze tracking is widely used and faces great challenges at the same time, so it is necessary to study gaze tracking deeply.

A. Significance of Gaze Tracking

1) Application of Gaze Tracking: Gaze tracking is broadly used in fatigue driving detection, eye disease diagnosis, mental illness diagnosis, website or advertising design, virtual reality, gaze-control devices and human-computer interaction [1]. In fatigue driving detection, the gaze of drivers is tracked. Once the driver is distracted, drunk or tired, warning signals will be sent out or the car will be stopped, which can improve traffic safety [2, 3]. In medicine, gaze tracking can be used for the diagnosis of various eye diseases, such as strabismus [4]. In psychology, gaze tracking can be used to diagnose neurological disorders [5], Alzheimer, and autism [6], etc. Psychologists often use psychological tests, MRI and CT scans to make the diagnosis. However, these tests are time-consuming and expensive. By contrast, if automatic gaze tracking is adopted, it will take psychologists less time and effort to evaluate the mental state of patients. In the aspect of website and advertisement design, by tracking the customers' gaze, the product information they are interested in and concerned about can be obtained, which will facilitate the design. In virtual reality, gaze is used to control the position of the cursor on the computer screen to aid to edit email, browse photos, play music and play games. Furthermore, robots or other devices can be controlled by gaze to achieve human-computer interaction. For example, the robot with hand-held endoscope is controlled by gaze to assist the surgery [7], and another robot is controlled by gaze to help the disabled with daily service [8].

2) Challenge of Gaze Tracking: There are many problems and challenges for remote gaze tracking in natural light. For example, in the outdoor environment, the image quality is often poor due to the influence of illumination, occlusion and distance from the camera. The changes of facial expression and head pose, especially the ever-changing human pose in natural scenes, bring great challenges to gaze tracking.

(1) Low Quality Human Eye Image: With the development of camera hardware, the resolution of the camera is gradually improved, and the images taken are more and more clear. However, with the increase of the distance between people and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2021.3064280, IEEE Transactions on Cognitive and Developmental Systems

cameras, the area of face and eye will decrease, and the quality of image in eye area will degrade. In addition, due to the influence of illumination, specular reflection and occlusion, it is difficult to capture a clear eye image. In such a low-quality image, it is difficult to properly locate the centers of pupil and iris, which increases the difficulty for accurate gaze tracking.

(2) Head Pose Change: For gaze tracking with free head movement, it is often necessary to estimate the head pose. While the head pose is usually estimated by locating the facial landmarks. However, the positions of the facial landmarks are affected by many factors. For example, the changes of human expression and the head pose, or occlusion, will cause the facial landmarks localization inaccurate, which will affect the accuracy of head pose estimation, then bring large deviations to the gaze tracking.

(3) Challenges in Natural Scenes: Gaze tracking in natural scenes is more challenging than that on the screen of a computer or mobile phone. The image taken may be a person's side face or back, in which the eyes may be closed or blocked, or only one eye is photographed, or even no eye is photographed. Therefore, the methods that rely on human eyes for gaze estimation are no longer applicable, and it also brings great challenges to accurate gaze tracking.

B. *Related Work*

Gaze tracking can usually be classified into feature-based methods and appearance-based methods.

In feature-based gaze tracking methods, the specific local features of the eye region of the image are identified, and then the relationship between these features and 2D gaze position is established, so as to carry out gaze tracking. The feature-based methods can be further divided into regression-based ones and model-based ones. A regression-based method is adopted in [9]. Firstly, the vector between the iris center and the reference point (inner canthus point) is taken as the input and the gaze point as the output, and the relationship between them is fitted by a polynomial. Then the pitching angle and yaw angle of the head pose are used to compensate the error so as to realize the gaze tracking. A model-based approach is adopted in [10]. The visual axis is approximately as the line from the iris center to the gaze point. The eye-related parameters are obtained by calibration, and the gaze point is calculated by the 3D geometric model of the human eye. Furthermore, feature-based gaze tracking methods can also be divided into head-mounted ones and remote ones. For head-mounted methods, gaze direction can be directly corresponding to pupil center position, without the need to determine the position and pose of the head. A binary feature selection approach is proposed to obtain robust pupil center in [11]. Different neural networks are adopted in order to achieve real-time pupil center [12, 13]. However, head-mounted devices are intrusive to people, then remote methods have emerged. A real-time remote pupil center detection approach is proposed in [14]. A high-speed remote eye tracker is developed to achieve an average gaze estimation error below 1 degree [15]. Because feature-based methods often rely on high-quality eye images, they are only suitable for short distance situations. For medium and long distance situations, it is often difficult to accurately estimate the gaze due to the degradation of image quality.

The appearance-based approaches are used to estimate the gaze by learning the mapping between the eye image and the gaze point or the gaze direction, including adaptive linear regression methods [16], Gaussian process regression ones [17], support vector machine ones [18], and convolutional neural network (CNN) ones. In the gaze tracking methods based on CNN, the eye image is taken as the input of the network, and the eye features are extracted by reasonably designing the network structure, then the gaze point or gaze direction is estimated. The methods are potentially suitable for low-quality images and are calibration-free, so they are extensively used in gaze tracking at present. Zhang et al. adopt different deep learning networks to estimate gaze [19-21]. In [20], AlexNet framework is adopted, in which spatial weight mechanism is introduced to estimate gaze, and the left-one cross-validation is performed in MPIIGaze[21] and EYEDIAP[22] databases respectively. Reference [19] and [21] are similar. The eye image is first normalized and then used as the input of the network, and the head pose is cascaded to the first full connection layer (FC), so as to estimate the gaze direction. The difference is that the LeNet-5 network is used in [19], while the VGG-16 network [23] is used in [21] for gaze estimation. The normalization of eye image facilitates cross-validation in different databases. Training was carried out on UT Multiview database [24], and cross-database cross-validation was performed on both MPIIGaze and EYEDIAP databases. In [25], the head image, left eye image and right eye image are used as input respectively, and three residual neural networks (resnet-18) [26] are used. Furthermore, the head pose is integrated into the final FC to estimate the gaze point on the screen of the mobile device. The GazeCapture database [27] is used for verification, and the accuracy obtained is higher than that of iTracker. As mentioned earlier, most current gaze tracking systems are principally for gaze points on the screens of computers or phones. It is more practical to track the gaze in natural scenes, but there are few researches on it. In [28], the saliency map in the natural scenes is modeled, and the position with the maximum value in the saliency map is regarded as the predicted gaze point. Because the saliency point is not necessarily the gaze point, the error for gaze estimation based on saliency map is large. In [29], the gaze point in natural scenes is estimated. Two channels, saliency map channel and gaze channel, are constructed. In the saliency map channel, the whole image is taken as the input, and the first five layers of Alex network are used. While in the gaze channel, the face image is taken as the input, and the first five layers of Alex network are also adopted, then the head position is integrated with it by the FC. Then the two channels are multiplied pixel-by-pixel, and the final gaze direction is obtained by shifting the grid. In [30], a multi-task method, in which gaze direction and gaze point are predicted simultaneously, is used to estimate the gaze in natural scenes. The head image and head position are taken as input. After the head image is processed by ResNet-50, it is integrated with the head position channel to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2021.3064280, IEEE Transactions on Cognitive and Developmental Systems

predict gaze direction. Then, the gaze direction is mapped into a multi-scale gaze direction field, which is cascaded with the original image and input into the feature pyramid network (FPN) to obtain the heat map about the gaze point, and the output of the network is supervised by the heat map.

As mentioned earlier, in view of the challenges faced in accurate gaze estimation in natural scenes, following the work in [29, 30], a method of gaze tracking in natural scenes based on attention mechanism and CNN with residual blocks is proposed, in order to further improve the accuracy of gaze tracking in real-time.

C. Contribution of the Paper

In order to effectively and accurately track the gaze in low-quality images, a method based on attention mechanism and CNN with residual blocks is proposed to track the gaze in natural scenes, and it is verified on the GazeFollow database [29] which contains six kinds of databases. The main contributions of this paper are summarized as follows.

1) The attention mechanism is introduced into the network to make the important features of attention more prominent, which contribute to improve the accuracy of gaze tracking.
2) In order to further improve the real-time performance of the network model, MobileNetV2 with good real-time performance is adopted. In the MobileNetV2, not only depthwise separable convolutions but also residual blocks are used, which makes the real-time performance improved with less loss of gaze tracking performance.

In addition, the face image is used as input for gaze tracking in the model, and the head pose information is implied in the face image, so there is no need for a special head pose estimator, which avoids the influence of the head pose estimation inaccuracy on gaze tracking.

The remaining sections are arranged as follows. In Section 2, the proposed method of gaze tracking is expounded, including the overview of the network model, three networks with residual blocks which are ResNet-50, FPN and MobileNetV2, attention connection mechanism, loss function and implementation details. In section 3, the database and the evaluation metrics are introduced. After that, the visual representation of the experimental results, the performance comparison of various networks based on residual blocks, the comparison with other methods of gaze tracking in natural scenes, and the real-time results are presented in section 4. The final conclusions and future work are demonstrated in Section 5.

II. Proposed Gaze Tracking Method

The main purpose of gaze tracking is to determine the position and direction of the gaze according to the eye position, the head position and the head pose. The gaze point studied here is located in the original image, and the gaze direction is from the head center to the gaze point. The information of eye position and head pose is implicit in the face image, and the gaze point is located in the original image, so the face image, head position and original image are taken as input, and the gaze position is taken as output to construct CNN model.

The proposed gaze tracking method is primarily based on CNN with residual blocks. At the same time, in order to improve the accuracy of gaze tracking, attention mechanism is integrated into the network model.

A. Overview of Gaze Tracking System

The overview of the proposed gaze tracking system is shown in Fig. 1.

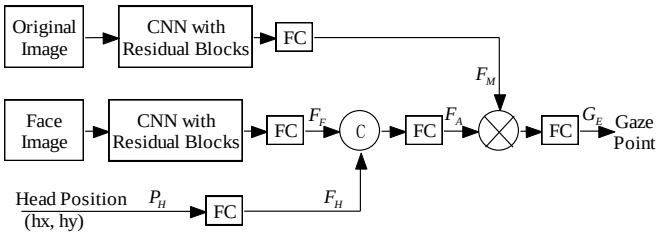

Fig. 1. The overview of the proposed gaze tracking system

The system is divided into two channels, one of which is the main channel with the original image as input, the other is the mask channel with the head image and the head position as input. The main channel is the upper one in Fig. 1. At first, the original image is processed by the CNN with residual blocks, and then after dimension reduction by a FC, the feature vector containing the location information of gaze point is produced. The mask channel is located at the bottom of Fig. 1. After the head image is processed by the CNN with residual blocks, it is cascaded and fused with the head position information to produce the weight vector with attention. After that, the main channel and mask weight channel are multiplied element-by-element. Then after dimension reduction by a FC, the position of the gaze point is obtained.

The CNN with residual blocks here includes ResNet-50, FPN based on ResNet-50, and MobileNetV2. In the network models using ResNet-50 and FPN, the structures of models are reasonably designed in order to improve the performance of gaze tracking. Furthermore, because the MobileNetV2 with residual blocks has good real-time performance, it is applied to the network model to improve the real-time of gaze tracking with a small loss of accuracy.

B. Networks with Residual Blocks

Because of the advantages of residual learning with high computational efficiency, the networks (ResNet-50, FPN and MobileNetV2) with residual blocks are used in our models. Each of them has its own advantages, which are described below.

1) ResNet-50: ResNet-50, proposed by He et al. is primarily composed of some bottleneck residual blocks. A typical bottleneck residual block is shown in Fig. 2.

It consists of the main branch from top to bottom and the Skip Connection on the right side. The core idea is to transmit the input information directly to the output by the Skip Connection, and only to learn the difference between the output and the input by the main branch, which simplifies the learning and thus improves the speed. For the main branch, the dimensionality is reduced by 1*1 convolution to generate 64 feature maps. Then,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2021.3064280, IEEE Transactions on Cognitive and Developmental Systems

the convolution operation of 3*3 is carried out to extract the main features. After that, the 1*1 convolution operation is performed to raise the dimension to generate 256 feature maps. Moreover, batch normalization (BN) operation is used in ResNet. By subtracting the mean value and dividing the variance of the same batch of data, the generalization ability of the model is enhanced, and the training speed is also improved. Furthermore, the Rectified Linear Units (ReLU) of nonlinear activation function is used to improve the nonlinear adaptability of the model. The bottleneck residual blocks adopted in the ResNet-50 make it have the following advantages.

(1) The problem of gradient vanishing is alleviated. For the deep neural network, with the increase of network depth, the gradients in the front layers are very small when the error is back propagated, which often leads to learning stagnation, that is, the problem of gradient vanishing. While in the ResNet network, BN operation and ReLU activation function are adopted to alleviate this problem.

(2) The degradation problem is addressed. For deep neural networks, with the increase of network depth, the increase of training parameters makes it difficult to optimize the algorithm, which easily leads to the increase of training error, that is to say, the problem of degradation occurs. In the ResNet, the residual blocks are used to solve the degradation problem, so that the network can reach deeper layers.

(3) The calculation efficiency is improved. In the bottleneck residual block, the convolution operation is carried out after dimension reduction, which saves the cost of calculation and promotes the efficiency.

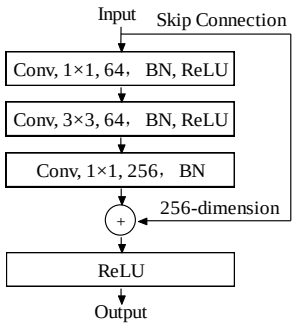

Fig. 2. Bottleneck Residual Block in ResNet

2) FPN: The FPN here is based on ResNet-50, and multi-scale features are fused [31]. The FPN network structure is shown in Fig. 3. The left side is a bottom-up process, which is primarily composed of the feature layers of ResNet-50 network. The image at the bottom of the Fig. is the input. Firstly, the size of the image is adjusted to 224×224, and then the image is normalized to 1×1. After that, it is processed by the module of conv1, which consists of a convolutional layer, a BN layer, a ReLU activation layer and a maximum pooling layer, and the output is C1. Then C1 is processed by conv2 module, which is composed of three bottleneck residual blocks, and the output is C2. Similarly, the C2 is processed by conv3, conv4 and conv5 in turn, which are composed of four, six and three bottleneck residual blocks respectively, and the outputs are C3, C4 and C5 respectively. The process of output C1 to C5 on the left is that of bottom-up pyramid feature extraction, in which multi-scale feature maps with different resolutions are generated. With the decrease of resolutions of feature maps, the semantic information is gradually enriched.

The right side of Fig. 3 is a top-down process. The C5 on the top level is dimensionally reduced by 1×1 convolution filter to obtain M5, which is filtered by 3 × 3 convolution to generate P5. And then, in order to integrate the feature map of the upper layer with that of the lower one, it is necessary to adjust feature maps of the two layers to have the same size. On the one hand, the feature map of the upper layer is up-sampled by twice the nearest neighbor; on the other hand, that of the lower layer is dimensionally transformed by a convolution of 1×1; and those of the two layers are fused by adding pixel-by-pixel. After feature fusion, M4, M3 and M2 are obtained respectively, as shown in Fig. 3. Finally, in order to reduce the aliasing effect caused by the up-sampling, the 3×3 convolution filtering is performed to obtain the outputs, which are P4, P3 and P2, respectively. The P2 is used as the final output here. In this way, by the FPN, the features with different scales are integrated.

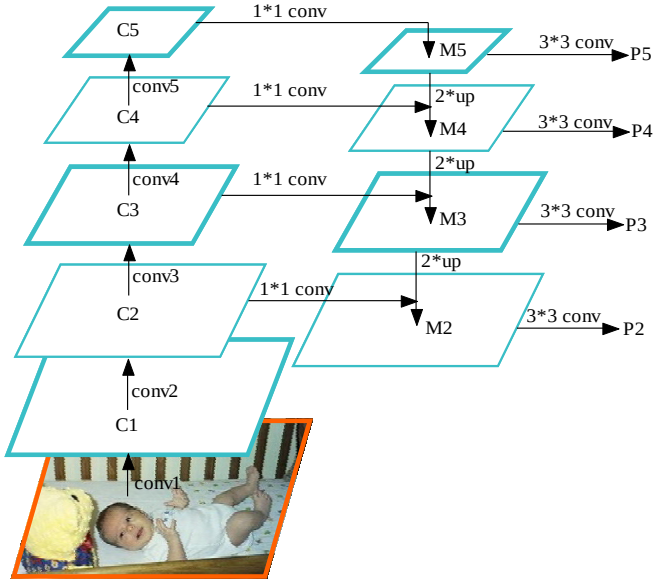

Fig. 3. FPN network structure

3) MobileNetV2: MobileNetV2 is a lightweight neural network model used for mobile phones proposed by Google [32, 33]. MobileNetV2 is basically composed of inverted residual blocks, as shown in Fig. 4, where Dwise is Depthwise. Similar to the bottleneck residual block in ResNet-50, in the inverted residual block of MobileNetV2, the Skip Connection reflecting residual learning is also adopted. Different from the ResNet-50, in MobileNetV2, the depth separable convolution is employed, that is, a standard convolution is decomposed into a depth one and a point-by-point one. The depth convolution is to convolve each input channel to carry out filtering operation. Point-by-point convolution is to use 1×1 convolution to linearly combine the outputs of all depth convolutions. The decomposition operation of first filtering and then combination effectively reduces the calculation time and model parameters

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2021.3064280, IEEE Transactions on Cognitive and Developmental Systems

without the loss of the accuracy. Moreover, unlike the bottleneck residual blocks in ResNet, in the MobileNetV2, expansion is followed by compression. After expansion, the 3*3 convolution operation is carried out in order to learn more features. And finally, compression is performed to screen out excellent features. In addition, ReLU6 of activation function is used in the MobileNet. Compared with ReLU, in ReLU6, the maximum output value is limited to 6, so that a good numerical resolution can be obtained for the mobile devices with low precision. For the final output, ReLU6 is removed, and the linear activation function is adopted instead, to prevent feature information from being destroyed due to the nonlinear activation of ReLU6.

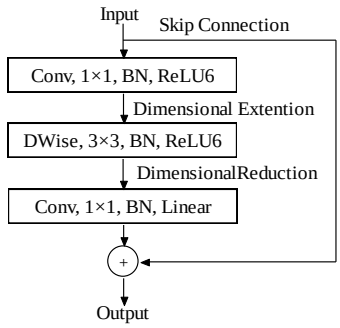

Fig. 4. Inverted residual blocks in MobileNetV2

C. *Attention Mechanism*

Attention mechanism is widely used in visual fields such as human pose estimation [34], saliency map detection [35] and expression estimation [36]. In the attention mechanism, an intermediate attention map is learned as a weight so as to select the important features. This is achieved by multiplying the source feature map with the attention map element-by-element, so as to select useful features from the source feature map by the weighting effect of the attention mechanism.

Inspired by the attention mechanism, the weight connection of attention mechanism is applied to gaze tracking in the paper. On the one hand, the main features are extracted from the original image by CNN with residual blocks, and then a FC is used to obtain the output F_M of the main channel, as shown in Fig. 1. On the other hand, the face image is first processed by CNN with residual blocks and then dimensionally adjusted by FC to get F_F. Meanwhile, the head position information is dimensionally adjusted by FC to product the output F_H. Then F_F and F_H are cascaded, and after a dimension reduction by FC, the output F_A of the mask channel with attention weight is obtained. Then, F_M and F_A are multiplied element-by-element in order to assign large weights to important features. After dimension reduction by the FC, the 2D gaze point is generated. Finally, the Sigmoid activation function is used to make the gaze point within 0 and 1, which is the final estimated gaze point G_E. This process can be expressed as:

$$G_E = Sigmoid\left[W \times \left(F_M \otimes F_A\right) + b\right] \quad (1)$$

$$F_A = F_F © F_H \quad (2)$$

where $\otimes$ represents element-by-element multiplication, Sigmoid is sigmoid activation function, and © denotes cascade. The important features are endowed with larger weights, so that the useful information in the image can be extracted, thus improving the accuracy of gaze tracking.

D. *Loss Function*

The whole network is trained, where total loss is composed of distance loss (L_{Dist}) of gaze point and angle loss (L_{Angle}) of gaze direction in the paper. The distance loss (L_{Dist}) is the Euclidean distance between the estimated gaze point (G_E) and the target gaze point (G_T), expressed by

$$L_{Dist} = \left\| G_E - G_T \right\|_2 . \quad (3)$$

The closer the estimated gaze point is to the target gaze point, the smaller the distance loss of gaze point will be. The angle loss is mainly represented by cosine similarity between estimated gaze direction (D_E) and target gaze direction (D_T), where the cosine similarity can be expressed as

$$CosSim = \frac{D_E \bullet D_T}{\max\left(\left\| D_E \right\|_2 \times \left\| D_T \right\|_2, \varepsilon\right)}, \quad (4)$$

where

$$D_E = G_E - P_H , \quad (5)$$

$$D_T = G_T - P_H , \quad (6)$$

and P_H is the head position. The cosine similarity is the dot product of the two vectors of D_E and D_T divided by their norms. In addition, to prevent the denominator from being 0, ε is used and set to a smaller value of 1e-08. The cosine similarity is the cosine of the angle between the estimated gaze direction and the target one. When the angle between them is smaller, the cosine similarity between them is larger, so the angle loss of gaze direction is designed as

$$L_{Angle} = 1 - CosSim . \quad (7)$$

It can be seen that the closer the estimated gaze direction is to the target one, the smaller the angle loss will be. Then, the total loss function can be written as

$$Loss = L_{Dist} + L_{Angle} . \quad (8)$$

According to the above loss function, the training samples are employed to train the model to reduce the total loss to get the optimal network parameters.

E. *Implementation Details*

The proposed method is implemented by Python programming language and based on the Pytorch framework. The Adam algorithm is used to optimize the network parameters. Moreover, the function of optim.lr_scheduler.StepLR in Pytorch is employed to adjust the learning rate, which can be expressed as

$$l_r = l_{ro} \times \gamma^n , \quad (9)$$

where, l_r is the learning rate, l_{ro} is the initial learning rate and its value is set to 0.0001, and γ is the multiplier of the decline in learning rate, which is set to 0.1 here. In formula, n is the result that the epoch is divided by step_size, which is given by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2021.3064280, IEEE Transactions on Cognitive and Developmental Systems

$$n = \frac{epoch}{step\_size}, \tag{10}$$

where the *step_size* is set to 5. It can be seen that every 5 epochs, the learning rate is updated once. With the increase of the epoch, the learning rate gradually decreased, so that the output of the model will not fluctuate too much in the later stage of training, and be closer to the optimal solution. Furthermore, in order to reduce the overfitting of the network model, the weight decay strategy, namely L2 regularization method, is adopted, where the weight decay factor is set to 0.0001.

III. Database and Evaluation Metrics

The proposed network model based CNN with residual blocks is trained on the public gaze database in natural scenes. The performance of gaze tracking is evaluated by certain evaluation metrics and compared with other gaze tracking methods, to verify the effectiveness of the proposed method.

A. Database

In order to obtain the performance of the proposed method, it needs to be verified on public databases. The GazeFollow database in natural scenes is selected [29]. The database contains six different databases, namely SUN [37], PASCAL [38], Actions 40 [39], MS COCO [40], Places [41] and ImageNet [42]. The diverse database provides a large number of data samples for the methods based on deep learning, and also brings challenges for the accurate gaze tracking.

The training sample set in GazeFollow database [29] contains 119,125 images and 12,557 data labels. The number of data labels is larger than that of images, because some data labels provide the gaze information of different people in the same image. The test sample set in the database contains 4782 images corresponding to 4782 data labels, which can be used to test and verify network models. The data label consists of the head bounding box, head position, target gaze position, the storage path of the image, and the source database where the image is located. The algorithm is trained on the training sample set and verified on the test sample set.

B. Evaluation Metrics

The main evaluation metrics are distance error of gaze point, angle error of gaze direction, and AUC. The smaller the errors are and the larger the AUC is, the higher the gaze tracking accuracy is, and the better the network performance is.

1) Distance Error of Gaze Point: The distance error of gaze point has the same expression as the distance loss of gaze point, shown in Equation (3), which is the Euclidian distance between the estimated gaze point and the target gaze point.

2) Angle Error of Gaze Direction: Angle error of gaze direction is the angle between the estimated gaze direction D_E and the target gaze direction D_T. The cosine of the angle between D_E and D_T can be obtained by the cosine similarity of formula (4). Then, the angle between D_E and D_T can be acquired by inverse cosine. Finally, the angle error of gaze direction can be obtained by converting radian into angle, which can be expressed as

$$E_{Angle} = \frac{180}{\pi} \times \arccos(CosSim). \tag{11}$$

3)AUC: AUC which is the area under the Receiver Operating Characteristic (ROC) curve, is a performance metric to measure the quality of deep neural network models. When a test sample is input into a network model, the probability of a predicted output is often generated. The prediction probability is then compared with a threshold value. If the former is greater than the latter, the test sample is classified as a positive category; otherwise it is a negative one. According to prediction results of network models, ROC curve can be drawn by changing the threshold value from 0 to the maximum. Its horizontal axis is the proportion of prediction samples with positive category to all target samples with positive one. Its vertical axis is the proportion of prediction samples with positive category to all target samples with negative one. The closer the ROC curve is to the upper left corner, the greater the AUC value is, and the higher the accuracy of the network model is.

In order to evaluate the performance of gaze point localization by AUC metric, the heat map of gaze point is used [28]. The heat map of the target gaze demonstrates that the value of the pixel at the target gaze position is 1 (positive category), and those at other positions are 0 (negative category). The heat map of the predicted gaze is determined by a 2D Gaussian distribution function centered on the predicted gaze point. Each pixel in the image acts as a binary classifier. If the probability value of the pixel in the heat map of the predicted gaze point is greater than the threshold value, it is classified as gaze point (positive category), otherwise it is non-gaze point (negative category). By changing the threshold value, the ROC curve can be obtained, and then the AUC can be determined.

IV. Experimental Results and Performance Evaluation

The proposed method is tested on the GazeFollow database, the test results are visualized, and compared with other methods of gaze tracking in natural scenes, as well as with other network structures, to verify the effectiveness of the proposed method.

A. Visualization of Experimental Results

In order to observe the quality of gaze tracking results, they are visually represented in the image. Some image examples with good results and poor ones are listed, as shown in Fig. 5 and Fig. 6 respectively. The area of the head is marked with a white rectangle box, which center is highlighted with a black circle. Gaze direction is from the head center to the gaze point. The target gaze point is highlighted with a small green circle, and the target gaze direction is marked with a white line. The estimated gaze point is highlighted with a small blue circle, and the estimated gaze direction is marked with a green line. From the examples with good results in Fig. 5, it is obvious that the estimated gaze points are close to the target ones, and corresponding gaze directions are also very close. From the examples with poor results in Fig. 6, it is clear that the distance errors of gaze point and the angle errors of gaze direction are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2021.3064280, IEEE Transactions on Cognitive and Developmental Systems

both large.

Fig. 5. Image examples with better results

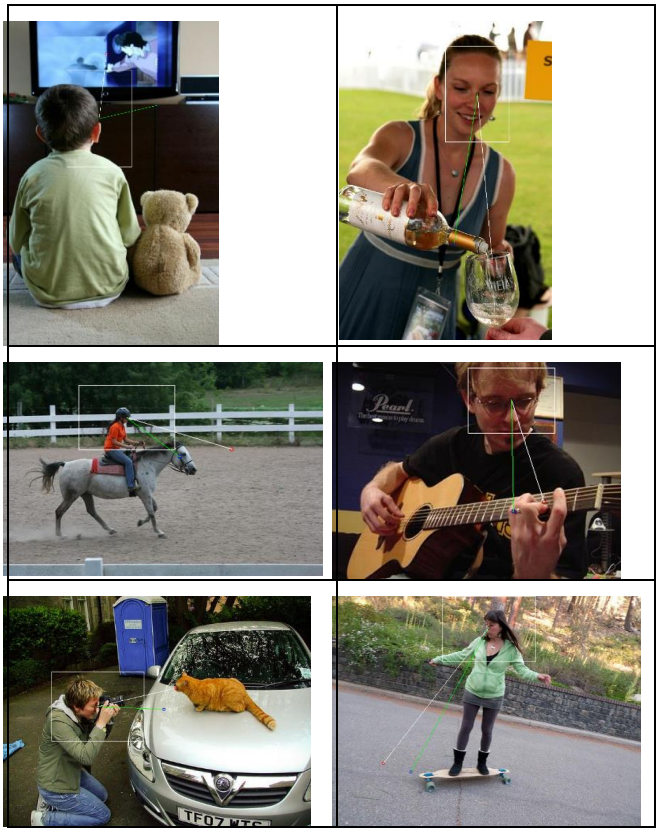

Fig. 6. Image examples with poor results

B. Performance Comparison with Different Network Structures

In order to verify the rationality of the network design, we try to use other network connection modes and make comparative experiments with the proposed one.

1) Comparison of network connection modes: In order to verify the connection mode, in addition to the multiplication connection mode with attention mechanism designed in the paper (as shown in Fig. 1), the cascading connection mode (as shown in Fig. 7) and the addition connection mode (as shown in Fig. 8) are also tested. The CNNs with residual blocks in both channels are designed as the ResNet-50. The numbers in Fig.7 and Fig. 8 denote the output size of feature layers.

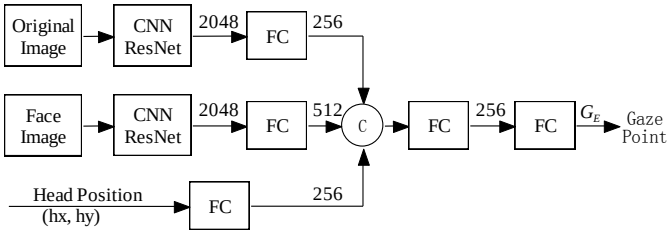

Fig. 7. Cascade connection network

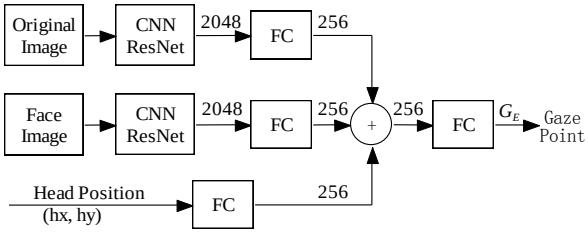

Fig. 8. Addition connection network

According to the evaluation metrics described in Section III-B, the above three connection modes are tested and compared, whose results are shown in Table I, where ⊗, ⓒ and ⊕ represent three different connection modes: multiplication, cascade and addition, respectively.

TABLE I
COMPARISON OF EXPERIMENTAL RESULTS OF DIFFERENT CONNECTION MODES

Connection Mode	AUC	Distance Error	Angle Error(°)
ResNet ⊗ ResNet	0.922	0.133	16.1
ResNet ⓒ ResNet	0.920	0.135	16.7
ResNet ⊕ ResNet	0.907	0.149	18.5

It is obvious that the performance of the addition connection is the worst, the cascade mode is in the middle, and the multiplication mode is the best. In the addition mode, the physical concept of the information fusion is not clear, which invisibly causes the destruction of the feature information. In the cascade mode, although some important information can be extracted, because the face image is a local amplified one of the original image, and the two channels with the face image and the original image as the input respectively are cascaded, which makes some repetitive feature information about the face image produced in the network model, and the information about the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2021.3064280, IEEE Transactions on Cognitive and Developmental Systems

gaze position in the original image invisibly reduced, gaze tracking performance is not very ideal. In the multiplication mode with attention mechanism, the original image in the main channel contains the gaze information, while the face image in the mask channel contains the information of eye position and head pose. According to the attention information implied in the face image, the gaze position in the original image is given a higher weight, which makes the important features representing the gaze position more prominent, so the performance of gaze tracking is the best. The result also verifies the effectiveness of the adopted attention mechanism.

2) Comparative Experiment of Different Network with FPN: The FPN is used to replace the ResNet in different channels and connection modes separately in order to further verify the rationality of the network connection mode and the selection of network modules. The comparison results are shown in Table II.

TABLE II
COMPARISON OF EXPERIMENTAL RESULTS OF DIFFERENT CONNECTION MODES WITH FPN

Different Network Module Connection	AUC	Distance Error	Angle Error(°)
FPN ⊗ ResNet	0.922	0.133	16.3
FPN ⓒ ResNet	0.919	0.136	16.7
FPN ⊕ ResNet	0.907	0.149	18.7
ResNet ⊗ FPN	0.919	0.136	17.1
ResNet ⓒ FPN	0.918	0.137	17.4
ResNet ⊕ FPN	0.905	0.151	19.2

In Table II, The first three rows indicate that FPN is used in the channel with the original image as input, and ResNet is used in the channel with the face image as input. While, in the last three rows, FPN and ResNet are interchanged. It can be seen from the first three rows that the performance of the network using the multiplication connection mode is the best, and the same conclusion can be drawn from the last three rows, which is the same as the conclusion in the previous part, and also fully verifies the effectiveness of the attention mechanism. Moreover, by comparing the performance of the network using the same connection mode in the first three rows and the last three rows respectively, it is obvious that the former is better than the latter. This indicates that regardless of connection mode, the network model, in which FPN is adopted in the channel with the original image as input, and ResNet is adopted in the channel with the face image as input, is better. Finally, by comparing the results of the first three rows in Table II with those in Table I using the same connection mode respectively, it is clear that the performance of network model with both channels using ResNet is the best. The reason is that in FPN, in the process of integrating high-level information with low-level information, in some sense, high-level semantic information is destroyed. Compared with FPN, in ResNet, only high-level semantic information is extracted, so its accuracy is higher. Furthermore, compared with FPN, there is no feature fusion process in ResNet, so the real-time performance is better. Therefore, the network model using ResNet in both channels is ideal.

3) Comparative Experiment of Different Networks with MobileNetV2: In MobileNetV2, not only residual blocks but also depthwise separable convolutions are adopted, which reduces network parameters and saves computing costs. Therefore, MobileNetV2 is used to replace ResNet in network model, so as to improve its real-time performance while ensuring its precision. The multiplication connection mode is adopted in the models. The results are shown in Table III. It is similar to the previous one, the former network is used in the main channel , while the latter network is used in the mask channel.

TABLE III
COMPARISON OF EXPERIMENTAL RESULTS OF DIFFERENT CONNECTION MODES WITH MOBILENET

Different Network Module Connection	AUC	Distance Error	Angle Error(°)
MobileNetV2 ⊗ ResNet	0.920	0.136	16.4
ResNet ⊗ MobileNetV2	0.913	0.142	18.2
MobileNetV2 ⊗ MobileNetV2	0.910	0.145	18.3

It can be seen from Table III that the results of the first row are significantly better than that of the other two rows, indicating that the adoption of ResNet in the mask channel shows obvious advantages, which is also consistent with the conclusion drawn from Table II. The face image, as a local magnified one of the original image, covers the information of head pose and eye position, which is the most important factor to determine the gaze direction. Therefore, the selection of the network module in the mask channel with face image as input directly affects the accuracy of gaze tracking. Because of the excellent advantages of ResNet, the model using ResNet in the mask channel achieves the best performance. In addition, although the results in the first row of Table III are slightly inferior to that of Table I, due to the higher real-time performance of MobileNetV2, the model, in which ResNet is replaced by MobileNetV2 in the main channel, is more suitable for situations with higher real-time requirement.

C. Performance Comparison with Other Gaze Tracking Methods

According to the evaluation metrics described in Section III-B, the proposed method is compared with other gaze tracking methods in natural scenes, and the results are shown in Table IV.

TABLE IV
PERFORMANCE COMPARISON OF GAZE ESTIMATION METHODS IN NATURAL SCENES

Methods	AUC	Distance Error	Angle Error(°)
Judd et al. [28]	0.711	0.337	54.0
Random [29]	0.504	0.484	69.0
Center [29]	0.633	0.313	49.0
Fixed bias [29]	0.674	0.306	48.0
SVM + one grid [29]	0.758	0.276	43.0
SVM + shift grid [29]	0.788	0.268	40.0
Recasens et al. [29]	0.878	0.190	24.0
Pan et al.(SalGAN) [43]	0.848	0.238	36.7
Lian et al.[30]	0.906	0.145	17.6
FPN ⊗ ResNet	0.922	0.133	16.3
ResNet ⊗ ResNet	**0.922**	**0.133**	**16.1**
MobileNetV2 ⊗ ResNet	0.920	0.136	16.4

Here, the multiplication connection mode is adopted, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2021.3064280, IEEE Transactions on Cognitive and Developmental Systems

ResNet is used in the mask channel. The FPN, ResNet and MobileNet are adopted respectively in the main channel, and their results are shown in the last three rows of Table IV. Compared with the other advanced methods, the performance of the proposed three models is better, that is, the AUC is higher, the distance error and angle error are smaller, and the accuracy is higher. The results show that the performance of the proposed method outperforms that of the other state-of-the-art methods, which also verifies its effectiveness. Among the proposed three models, the one where the ResNet are adopted in the two channels is the best.

D. Real Time Performance

The real-time performance of the proposed three different networks (shown in the last three rows of Table IV) and that in [30] are test and compared. The program is run under the Ubuntu system with two GPUs. The running time of all 4782 test samples is counted, and the average running time (ACT) is obtained by dividing the total number of samples. The ACT for different batch size is counted, and their results are shown in Table V. Furthermore, the trend curves of the ACT are shown in Fig. 9, highlighted with different colors.

TABLE V
COMPARISON OF THE AVERAGE RUNNING TIME (ACT)

Batch size	ACT (ms) in [30]	ACT(ms) with FPN	ACT(ms) with ResNet	ACT(ms) with MobileNet
100	1.21	1.19	1.16	1.10
200	0.852	0.832	0.800	0.781
300	0.761	0.745	0.714	0.680
400	0.683	0.678	0.653	0.640
500	0.664	0.660	0.647	0.634
600	0.774	0.650	0.621	0.617
700	0.869	0.783	0.622	0.619

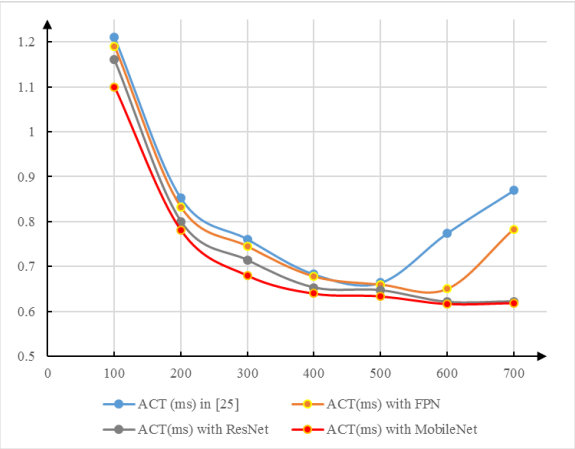

Fig. 9. Trend Curve of the average running time (ACT)

It can be seen from Table V that the running time will decrease with the increase of the number of batch samples. However, when the batch size increases to a certain value, the running speed will decrease, due to the limited memory resources. When the batch size continues to increase to another certain value, an error will be reported due to insufficient memory.

Moreover, the proposed network model where FPN is adopted in the main channel, is compared with that in [30]. The proposed model shows the advantages of reduced operation time and improved speed. In [30], a series network structure is adopted, and the multi-scale gaze field is computed, which adds extra calculation cost to the network. By contrast, the parallel network structure is adopted in the proposed method, and the model is relatively simple, so the running speed is higher. Furthermore, the real-time performance of the proposed three network models, where FPN, ResNet and MobileNet are adopted in the main channel respectively, is compared. It is clear that the operation time of the model with MobileNet is the shortest, followed by that with ResNet, and that with FPN is the longest. In view of the performance comparison of the proposed three models, the model with ResNet has the best gaze tracking performance and better real-time performance, so it is more suitable for practical applications. Although the gaze tracking performance of model with MobileNet is not as good as the other two proposed networks, its gaze tracking performance is better than other advanced methods listed in Table IV, and its real-time performance is the best, so it is more suitable for real-time applications.

V. GENERALIZATION ABILITY

In order to verify the generalization ability, we have evaluated our approach on another publicly available dataset named Daily Life Gaze following dataset (DL Gaze) [30]. Some result image examples are shown in Fig. 10. Comparison of experimental results of different connection modes is shown in Table VI. The results show that the model with multiplication connection mode has the best performance, which again confirms the effectiveness of attention mechanism.

Fig. 10. Result image examples on DL Gaze dataset

TABLE VI
COMPARISON OF EXPERIMENTAL RESULTS OF DIFFERENT CONNECTION MODES

Connection Mode	AUC	Distance Error	Angle Error(°)
ResNet ⊗ ResNet	0.910	0.141	16.9
ResNet ⓒ ResNet	0.905	0.145	17.1
ResNet ⊕ ResNet	0.898	0.154	18.3

TABLE VII
PERFORMANCE COMPARISON OF GAZE ESTIMATION METHODS ON DL GAZE

Methods	AUC	Distance Error	Angle Error(°)
Lian et al. [30]	-----	0.157	18.7
FPN ⊗ ResNet	0.904	0.146	17.6
ResNet ⊗ ResNet	0.910	0.141	16.9
MobileNetV2 ⊗ ResNet	0.904	0.146	17.6

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2021.3064280, IEEE Transactions on Cognitive and Developmental Systems

Moreover, the different models on DL Gaze dataset are compared, shown in Table VII. It is clear that the performance of the proposed three models is also better than that of [30], which verifies their good generalization ability. Furthermore, the model with lightweight MobileNetV2 is more suitable for real-time applications with less loss of accuracy.

VI. Conclusion and Future Work

In view of the fact that most of the current gaze estimation methods aim at the screen of a computer or a mobile phone, while the gaze tracking methods in natural scenes are less studied, and the accuracy is also low, a gaze tracking method in natural scenes is proposed. The proposed method is based on CNN with residual blocks, and the attention mechanism is integrated into the network model to improve its accuracy. It is compared with other advanced gaze tracking methods in natural scenes, and the results show that the performance of the proposed method outperforms that of the other state-of-the-art methods, which also verifies its effectiveness. At the same time, the proposed network model with multiplication connection mode is compared with those with the other two connection modes, and the results show that the proposed model is the best, which also confirms the advantages of the attention mechanism. Furthermore, the models based on three CNNs with residual blocks (ResNet-50, FPN and MobileNetV2), are compared and tested. It follows that because the face image covers the significant gaze-related information, and the accuracy of ResNet is high, the model using ResNet in the channel with the face image as input is better than others. At the same time, the performance of the model with ResNet in the main channel is better than others. Therefore, the model using ResNet in the two channels is the best, which is suitable the practical applications. In addition, due to the better real-time performance of MobileNetV2, the ResNet-50 in the main channel is replaced into MobileNetV2. The result shows that the real-time performance of the model after the replacement is improved with less loss of gaze tracking performance. Therefore, the network model using MobileNetV2 is more suitable for real-time applications.

In this paper, the target gaze point is limited in the image, excluding the situation that it is outside the image or it is 3D space point, which will be studied in the future. Furthermore, due to the variety of human pose in natural scenes, the image taken may be a person's side face or back, in which the eyes may be closed or blocked. Therefore, it is difficult to locate the eyes, which invisibly brings a huge challenge to the accurate gaze tracking. Thus, the accuracy of gaze tracking is not very high, and it needs to be further improved. In addition, gaze estimation for a single image is studied in the paper, in which gaze tracking for video frame image is not performed, so real-time gaze tracking for video frame image is also the future work.

References

[1] L. Dai, J. Liu, Z. Ju, and Y. Gao, "Iris center localization using energy map synthesis based on gradient and isophote," *Journal of Intelligent and Fuzzy Systems*, vol. 38, no. 4, pp. 1-13, 2020.

[2] F. Vicente, Z. Huang, X. Xiong, F. D. l. Torre, W. Zhang, and D. Levi, "Driver Gaze Tracking and Eyes Off the Road Detection System," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2014-2027, 2015.

[3] S. Guasconi, M. Porta, C. Resta, and C. Rottenbacher, "A low-cost implementation of an eye tracking system for driver's gaze analysis," in *2017 10th International Conference on Human System Interactions (HSI)*, 2017, pp. 264-269.

[4] Z. H. Chen, H. Fu, W. L. Lo, and Z. R. Chi, "Strabismus Recognition Using Eye-Tracking Data and Convolutional Neural Networks," *Journal of Healthcare Engineering*, 2018, Art. no. 7692198.

[5] E. Hernandez, S. Hernandez, D. Molina, R. Acebron, and C. E. Garcia Cena, "OSCANN: Technical Characterization of a Novel Gaze Tracking Analyzer," *Sensors (Basel)*, vol. 18, no. 2, Feb 9 2018, Art. no. 522.

[6] S. M. Anzalone, J. Xavier, S. Boucenna, L. Billeci, A. Narzisi, F. Muratori *et al.* Quantifying patterns of joint attention during human-robot interactions: An application for autism spectrum disorder assessment [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865518300758

[7] K. Fujii, G. Gras, A. Salerno, and G.-Z. Yang, "Gaze gesture based human robot interaction for laparoscopic surgery," *Medical Image Analysis*, vol. 44, pp. 196-214, Feb. 2018.

[8] M.-Y. Wang, A. A. Kogkas, A. Darzi, and G. P. Mylonas. (2018). *Free-View, 3D Gaze-Guided, Assistive Robotic System for Activities of Daily Living*. Available: https://arxiv.org/abs/1807.05452v2

[9] S. He, "Research on Gaze Tracking Algorithm Based on Binocular Stereo Vision Technology," Chongqing University, Chongqing, China, 2017.

[10] X. Zhou, H. Cai, Y. Li, and H. Liu, "Two-eye model-based gaze estimation from a Kinect sensor," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1646-1653.

[11] W. Fuhl, D. Geisler, T. Santini, T. Appel, W. Rosenstiel, and E. Kasneci, "CBF: circular binary features for robust and real-time pupil center detection," presented at the Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Warsaw, Poland, 2018. Available: https://doi.org/10.1145/3204493.3204559

[12] W. Fuhl, H. Gao, and E. Kasneci, "Tiny convolution, decision tree, and binary neuronal networks for robust and real time pupil outline estimation," presented at the ACM Symposium on Eye Tracking Research and Applications, Stuttgart, Germany, 2020. Available: https://doi.org/10.1145/3379156.3391347

[13] W. Fuhl, H. Gao, and E. Kasneci, "Neural networks for optical vector and eye ball parameter estimation," presented at the ACM Symposium on Eye Tracking Research and Applications, Stuttgart, Germany, 2020. Available: https://doi.org/10.1145/3379156.3391346

[14] W. Fuhl, S. Eivazi, B. Hosp, A. Eivazi, W. Rosenstiel, and E. Kasneci, "BORE: boosted-oriented edge optimization for robust, real time remote pupil center detection," presented at the Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Warsaw, Poland, 2018. Available: https://doi.org/10.1145/3204493.3204558

[15] B. Hosp, S. Eivazi, M. Maurer, W. Fuhl, D. Geisler, and E. Kasneci, "RemoteEye: An open-source high-speed remote eye tracker," *Behavior Research Methods*, vol. 52, no. 3, pp. 1387-1401, Jun. 2020.

[16] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive Linear Regression for Appearance-Based Gaze Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 2033-2046, 2014.

[17] J. Choi, B. Ahn, J. Park, and I. S. Kweon, "GMM-based saliency aggregation for calibration-free gaze estimation," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 1096-1099.

[18] Y. L. Wu, C. T. Yeh, W. C. Hung, and C. Y. Tang, "Gaze direction estimation using support vector machine with active appearance model," *Multimedia Tools and Applications*, vol. 70, no. 3, pp. 2037-2062, 2014.

[19] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, and Ieee, "Appearance-Based Gaze Estimation in the Wild," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4511-4520.

[20] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation," 2017.

[21] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 162-175, 2018.

[22] K. A. F. Mora, F. Monay, and J.-M. Odobez, "EYEDIAP: a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras," presented at the Proceedings of the Symposium on Eye Tracking Research and Applications, Safety Harbor, Florida, 2014.
[23] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," presented at the ICLR, 2015.
[24] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1821-1828.
[25] E. T. Wong, S. Yean, Q. Hu, B. S. Lee, J. Liu, and R. Deepu, "Gaze Estimation Using Residual Neural Network," in *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2019, pp. 411-414.
[26] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, New York, 2016, pp. 770-778.
[27] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik *et al.*, "Eye Tracking for Everyone," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2176-2184.
[28] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 2106-2113.
[29] A. Continente, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 199-207.
[30] D. Lian, Z. Yu, and S. Gao, "Believe It or Not, We Know What You Are Looking At!," in *Computer Vision – ACCV 2018*, Cham, 2019, pp. 35-50: Springer International Publishing.
[31] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," vol. 1, no. 4, p. arXiv:1612.03144v2Available: https://arxiv.org/abs/1612.03144
[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *arXiv e-prints*, p. arXiv:1801.04381Accessed on: January 01, 2018Available: https://ui.adsabs.harvard.edu/abs/2018arXiv180104381S
[33] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan *et al.*, "Searching for MobileNetV3," *arXiv e-prints*, p. arXiv:1905.02244Accessed on: May 01, 2019Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190502244H
[34] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context Attention for Human Pose Estimation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5669-5678.
[35] T. Zhao and X. Wu, "Pyramid Feature Attention Network for Saliency detection," p. arXiv:1903.00179v2Available: https://arxiv.org/abs/1903.00179?context=cs.CV
[36] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention Mechanism-based CNN for Facial Expression Recognition," *Neurocomputing*, Jun. 2020.
[37] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485-3492.
[38] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303-338, Jun. 2010.
[39] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *2011 International Conference on Computer Vision*, 2011, pp. 1331-1338.
[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, Cham, 2014, pp. 740-755: Springer International Publishing.
[41] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, Montreal, Canada, 2014, pp. 487–495: MIT Press.
[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, Dec 2015.
[43] J. Pan, C. Canton Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol *et al.*, "SalGAN: Visual Saliency Prediction with Generative Adversarial Networks," *arXiv e-prints*, p. arXiv:1701.01081Accessed on: January 01, 2017Available: https://ui.adsabs.harvard.edu/abs/2017arXiv170101081P

LIHONG DAI received the B.S. degree in automatic control and M.S. degree in control theory and control engineering from University of Science and Technology Liaoning, China, in 2000 and 2004 respectively. She is currently pursuing the Ph.D. degree in Pattern Recognition and Intelligent System at Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang City, China.

Since 2004, she has been a teacher in school of Electronic and Information Engineering, University of Science and Technology Liaoning, China. She is currently a Senior Lecturer. Her research interests include gaze tracking, computer vision, machine learning, pattern recognition, and their applications on human-robot interaction and collaboration.

JINGUO LIU (M'07-SM'18) received the Ph.D. degree in mechatronics from Shenyang Institute of Automation (SIA), Chinese Academy of Sciences (CAS), in 2007, where he has been a Full Professor since January 2011. He has been the Assistant Director of the State Key Laboratory of Robotics since 2008, and has also been the Associate Director of the Center for Space Automation Technologies and Systems since 2015. His research interests include bio-inspired robotics and space robot. He has authored or coauthored five books, over 100 articles and holds 50 patents in above areas.

He is a Member of the IEEE Technical Committee on Safety, Security, and Rescue Robotics, a Member of the IEEE Technical Committee on Marine Robotics, and the Senior Member of the Chinese Mechanical Engineering Society. He was a recipient of the T. J. TARN Best Paper Award in Robotics from the 2005 IEEE International Conference on Robotics and Biomimetics, the Best Paper Award of the Chinese Mechanical Engineering Society, in 2007, the Best Paper Nomination Award from the 2008 International Symposium on Intelligent Unmanned Systems, the Best Paper Award from the 2016 China Manned Space Academic Conference, the Outstanding Paper Award from the 2017 International Conference on Intelligent Robotics and Applications, and the Best Paper Award from the 2018 International Conference on Electrical Machines and Systems. He services as the Associate Editor or Technical Editor of several journals such as IEEE/ASME Transactions on Mechatronics, Journal of Field Robotics, Mechanical Sciences, Science China Technological Sciences, Chinese Journal of Mechanical Engineering, and Chinese Journal of Aeronautics.

ZHAOJIE JU (M'08-SM'16) received the B.S. degree in automatic control and the M.S. degree in intelligent robotics from the Huazhong University of Science and Technology, China, and the Ph.D. degree in intelligent robotics from the University of Portsmouth, U.K. He held research appointments at University College London, London, U.K., before he started his independent academic position at the University of Portsmouth, in 2012. He has authored or coauthored over 200 publications in journals, book chapters, and conference proceedings and received five Best Paper Awards and one Best AE Award in ICRA2018. His research interests include machine intelligence, pattern recognition and their applications on human motion analysis, multi-fingered robotic hand control, human–robot interaction and collaboration, and robot skill learning.

Dr. Ju is an Associate Editor of several journals, such as IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS and Neurocomputing.

YANG GAO (S'00-M'03-SM'09) received the B.Eng. degree in electrical and electronics engineering and the Ph.D. degree in artificial intelligence, control and instrumentation from Nanyang Technological University (NTU), Singapore, in 2000 and 2003, respectively. She is currently the Professor of Space Autonomous Systems and Head of the STAR LAB at the Surrey Space Centre, University of Surrey, Guildford, UK. She specializes in robotic vision, machine learning, and biomimetic mechanism for industrial applications. She brings 20 years of R&D experience in solving robotic system problems, and is actively involved in development of real-world space missions, such as ESA's ExoMars, Proba3, VMMO (Lunar Ice Mapper), UK MoonLITE/Moonraker, and CNSA Chang'E3.