IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 11, 2020, accepted April 23, 2020, date of publication April 27, 2020, date of current version May 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2990685

Learning a 3D Gaze Estimator With Adaptive Weighted Strategy

XIAOLONG ZHOU[1,2], (Member, IEEE), JIAQI JIANG[2], QIANQIAN LIU[2], JIANWEN FANG[1], SHENGYONG CHEN[2,3], (Senior Member, IEEE), AND HAIBIN CAI[4]

[1]College of Electrical and Information Engineering, Quzhou University, Quzhou 324000, China
[2]College of Computer Science Technology, Zhejiang University of Technology, Hangzhou 310023, China
[3]School of Computer Communication and Engineering, Tianjin University of Technology, Tianjin 300384, China
[4]Department of Computer Science, Loughborough University, Loughborough LE11 3TU, U.K.

Corresponding authors: Xiaolong Zhou (xlvision@hotmail.com) and Shengyong Chen (sy@ieee.org)

This work was supported in part by the National Natural Science Foundation of China under Grant U1709207, Grant 61876168, and Grant 61906168, in part by the National Key Research and Development Program of China under Grant 2018YFB1305200, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY18F030020 and Grant LY15F020041.

ABSTRACT As a method of predicting the target's attention distribution, gaze estimation plays an important role in human-computer interaction. In this paper, we learn a 3D gaze estimator with adaptive weighted strategy to get the mapping from the complete images to the gaze vector. We select the both eyes, the complete face and their fusion features as the input of the regression model of gaze estimator. Considering that the different areas of the face have different contributions on the results of gaze estimation under free head movement, we design a new learning strategy for the regression net. To improve the efficiency of the regression model to a great extent, we propose a weighted network that can adjust the learning strategy of the regression net adaptively. Experimental results conducted on the MPIIGaze and EyeDiap datasets demonstrate that our method can achieve superior performance compared with other state-of-the-art 3D gaze estimation methods.

INDEX TERMS Gaze estimation, weighted network, regression model, adaptive strategy.

I. INTRODUCTION

The gaze vector can be speculated from the pupil to the target's attention. It has been increasingly important as a non-verbal cue in many fields, including marketing and consumer research [1], [2], human-computer interaction [3]–[5], medical care [6]–[8], aviation and vehicle driving [9], and criminal investigation [10]–[12]. However, the existing gaze estimation systems often have the following defects: redundancy calibration process, low tolerance to head movement, limitation of lighting conditions and complex system settings, which limit the commercial promotion of gaze estimation.

In order to reduce the influence of the above-mentioned defects, there have been an increasing number of methods proposed for gaze estimation, which can be roughly classified into two major categories: model-based and appearance-based methods.

The model-based gaze estimation [13], [14] method uses the fitting model to estimate gaze direction relying on invariant facial features, such as pupil center [15], iris outline [16] and corneal infrared reflection [17]. Guestrin [18] established an eye model based on the pupil center, fixation point and eye center, and only one camera with two infrared light sources can complete the eye line estimation. Hennessey *et al.* [19] considered the influence of various head postures, relying on complex and detailed calibration steps to complete the gaze estimation of free head movement. To simplify the calibration procedure, Shih and Liu [20] proposed an improved Le Grand model and combined with the head attitude compensation model. By solving the linear equation to estimate the optical axis, this method can use two cameras to achieve the purpose of single point calibration and update the mapping function dynamically. Zhou *et al.* [21] developed a binocular model-based gaze tracking method, proposed an improved iris center localization method based on gradient characteristics, and simplified the individual calibration process requiring only one calibration point.

The appearance-based gaze estimation method extracts input features from the human eye appearance images and realizes gaze estimation by establishing a mapping

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

relationship between input features and gaze direction. Different from the model-based gaze estimation methods, the appearance-based methods usually only need a single camera to capture the user's eye images. Common input features include complete face image, human eye image, color opponency and histogram extracted from eyes. There are many kinds of mapping relationships, including k-Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), Gaussian Process (GP) and Artificial Neural Network (ANN). Zhang *et al.* [22] first extracted three low-dimensional features from the eye images, including the color opponency, gray scale intensities and direction information, and then used a KNN classifier with k = 13 to learn the mapping from image features to gaze direction. Wang *et al.* [23] added the depth feature to the traditional gaze estimation and applied the RF regression based on cluster-to-classify node splitting rules. Kacete *et al.* [24] used RF regression to estimate the gaze vector from the high dimensional data with the face information. The RF could do parallel processing as well as the training speed is relatively fast. Wu *et al.* [25] located the eye region by modifying the characteristics of active appearance model and used SVM to classify the five gaze directions.

Recently, with the development of machine learning and the support of massive data, extensive learning-based gaze estimation methods have been presented. These methods such as Convolutional Neural Network (CNN)-based methods, have great potential to handle the challenges faced by traditional methods, including redundancy calibration process, complex head postures, and limitation of lighting conditions. Zhang *et al.* [26] built a novel in-the-wild dataset and employed the CNN to learn the mapping from the head pose and eye images to gaze angles. Krafka *et al.* [27] introduced an eye tracking method for mobile devices, which used face image, individual eye and face grid as input. Zhang *et al.* [28] used a spatial weights CNN to encode face images and flexibly suppressed or enhanced the information of different face regions. Cheng *et al.* [29] proposed a concept of two-eye symmetry to predict the 3D gaze direction, and designed an evaluation network to adaptively adjust the regression network according to the performance of the eyes. Palmero *et al.* [30] used face, eye region and face landmarks as separate information flows in CNN to estimate gaze in static images. The learning features of all frames were input into a many-to-one recurrent module sequentially, and the 3D gaze vector of the last frame was predicted. Fischer *et al.* [31] recorded a new dataset of different head postures to improve the robustness of gaze estimation, applied semantic image inpainting to the area covered by glasses to eliminate the obtrusiveness of the glasses and built a bridge between training and test images. Yu *et al.* [32] introduced a constrained landmark-Gaze model to get the relation of eye landmark locations and gaze directions. Park *et al.* [33] used single eye image as input and simplified the task of 3D gaze estimation. They mapped the appearance of the eye to the intermediate pictorial representation, which was easier to learn the end-to-end model. In [34], [35], the authors introduced a hybrid-model that used CNN to map image to eye landmarks and then mapped eye landmarks to eye gaze. Wang *et al.* [36] proposed to combine adversarial learning and Bayesian inference into a unified framework. They also added an antagonistic component to traditional CNN-based gaze estimators so they could learn features that respond to the gaze.

In order to further utilize the powerful function of CNNs and improve the accuracy of gaze vector prediction, we propose an adaptive weighted 3D gaze estimation method. The main contributions of this paper are listed as following.

(1) We improve the Itracker model [27] to predict single-frame gaze. The face gird in the conventional Itracker model can be used to locate face position to supply the location information for 2D gaze estimation, while this branch is removed in the improved model because it is useless in our 3D gaze estimation. To further improve the performance of the model, we concatenate the facial stream, left eye stream and right eye stream to obtain their joint characteristics.

(2) During the process of model training, the face, left eye and right eye images have different influence on the final result. We propose a new loss function for the improved regression model. Based on the traditional regression loss function, we add the weight function of the three regional images.

(3) We propose a weighted network to judge the contribution of face, left eye and right eye images on the results of gaze estimation. According to the errors between the predicted value and the real value, the corresponding weight will be obtained. The adaptive weighting is realized by adjusting the strategy of regression model by weight value.

II. PROPOSED GAZE ESTIMATION METHOD

In this section, we present an adaptive weighted gaze estimation method. Firstly, the regression function of 3D gaze estimation is introduced. Then, the steps of data preprocessing are stated. Finally, the network architecture and the steps of adaptive weighted implementation are detailed. The overall architecture is shown in Fig. 1.

A. 3D GAZE ESTIMATION

Based on the image of eye appearance, a regression function f is constructed to establish the mapping relationship between image I and 3D gaze vector g, where $g = f(I)$. At present,

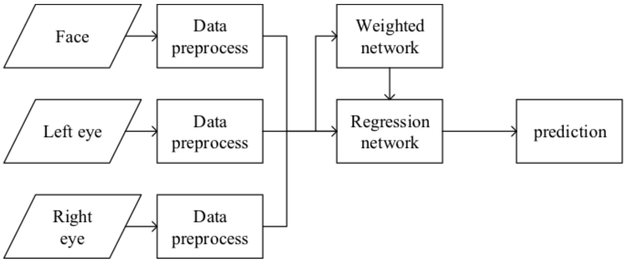


FIGURE 1. The overall architecture of proposed 3D gaze estimation method.

various regression models have been used in the gaze estimation methods, such as Neural Network, RF regression, GP regression, and SVM regression. We use the CNN to solve the problem because the regression of gaze estimation is usually highly nonlinear. With the development of deep neural networks, designing an efficient network architecture with large training dataset can solve this complex regression problem simply.

B. DATA PREPROCESSING

The results of gaze estimation are significantly affected by the head pose. Similar to [32], we normalize the image data to weaken the influence of this factor. The basic concept of data preprocessing is shown in Fig. 2.

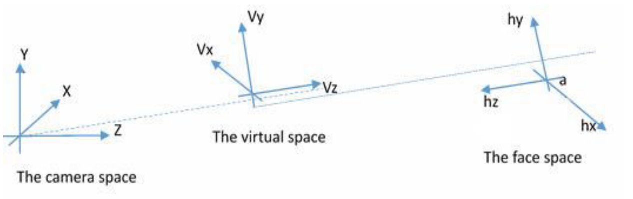


FIGURE 2. **The basic concept of data preprocessing.**

The data normalization method is to make a perspective transformation on the original image, so that the training model can be complete for gaze estimation in a specific virtual space. The method transforms the original image and the normalized image to satisfy the following three conditions. 1) The center of the face reference point is located at a fixed distance d from the center of the normalized image. 2) The horizontal direction of the head is parallel to the X-axis of the normalized image. 3) The face always has the same size in the normalized image.

We place the face reference point in the center of the image at a fixed distance from the camera. Assume that $a(a_x, a_y, a_z)$ is the face reference point under the camera space. The first condition is satisfied by setting the z-axis of the virtual space be $v_z = a_z/\|a_z\|$. To satisfy the second condition, the y-axis of the virtual space has to be defined as $v_y = v_z \times h_x$, where h_x is the rotation matrix of head pose in x-axis. The remaining x-axis of the visual space then can be computed by $v_x = v_z \times v_y$. Using these vectors, the rotation matrix can be defined as $R = \left[v_x/\|v_x\|, v_y/\left\|v_y\right\|, v_z/\|v_z\|\right]$. The transformation matrix is then defined as $M = SR$, where scaling matrix S to satisfy the third condition can be defined as S $= diag(1, 1, d/\|a\|_2)$.

We use the warp matrix W to transform the human face into an image plane of a specific camera space. Let $W = C_aMC_v^{-1}$, where C$_a$ is the internal parameter matrix of the original camera and C$_v$ is the internal parameter matrix of the virtual camera. In addition, the original gaze label also needs to be converted during the training stage by $g_v = Rg_a$,where g_v and g_a represent the normalized gaze label and initial gaze label respectively. In the test phase, $g_a = R^{-1}g_v$ is used to convert the virtual camera space to the original camera space for each prediction result.

The proposed data normalization method can cope with the influence of the difference of cameras in the real world on the prediction accuracy. This operation will not have any impact on the experimental process, but it should be noted that the accuracy of the internal parameter value of the camera is closely related to the presentation of the final sight vector result.

C. REGRESSION NETWORK ARCHITECTURE

In this paper, we propose an adaptive weighted regression model for the appearance-based gaze estimation. In practice, we observed that the left eye, right eye and face images have different contributions on the accuracy of regression in different scenes. The different image areas cannot achieve the same accuracy value. Therefore, when training a gaze regression model, it is better to rely on the high-quality images to train a more effective model. This model is composed of a main network and a sub-network. The main network performs the regression prediction from image to gaze vector, and the sub-network performs the adjustment of the Loss function of main network to achieve the purpose of adaptive adjustment.

The proposed network learns a regression model to predict the ground truth of gaze vector with left eye, right eye and face images. The overall structure is shown in Fig. 3.

In [27], the author used the face, both eyes and face grid separately into a branch of the network, and finally mapped the merged features which extracted from each branch to the ultimate 2D gaze point on the screen. Since the method in [27] predicts the gaze point on the screen, it not only needs to obtain the gaze vector, but also needs the face grid to provide the position information of the head position in the camera space. However, we mainly consider how to predict the gaze vector effectively. Therefore, we remove the face grid in our architecture. To realize the concept of adaptive weighted, the separate features and joint features of the face and the two eyes should be extracted and utilized.

As shown in Fig. 4, the regression network is a six-stream convolutional neural network. We use the reduced version of the convolution layers of a Lexnet as the basic network of each branch. Considering that when the eyeball rotates, many areas of the face will have big or small changes. In order to realize the self-adaptive adjustment of spatial weight, this paper adds three fused features to the basic features of face, left and right eyes. The fused features of face, left eye, and right eye are input as a single branch in the network. The first three streams are designed to extract the 64-dimensional deep features from the face, left eye and right eye respectively, and the last three streams are used to produce a joint 64-dimensional deep features. These six streams are then combined through a FC layer, and the dropout layer is used to prevent over-fitting problem. Finally, the corresponding gaze vectors are obtained through a 6-dimentiaonl FC layer.

The face and the both eyes can play different roles in the training network. If one of the areas is more likely to achieve

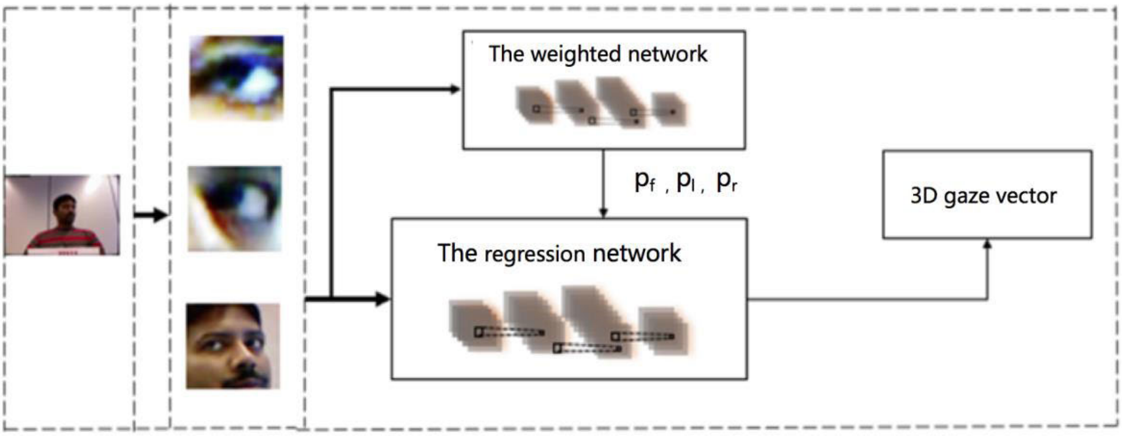


FIGURE 3. Overview of the proposed adaptive weighted regression model.

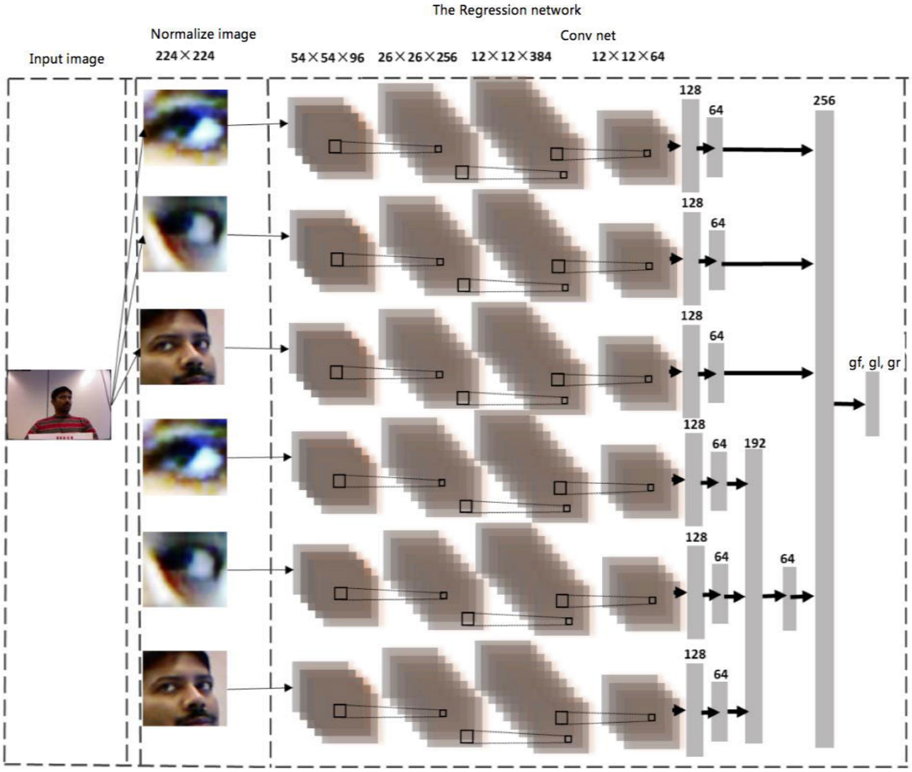


FIGURE 4. The regression network.

a smaller error, then we should expand its weight in the optimization of the network. Following this idea, we propose a new strategy to optimize the network.

We first calculate the angular error of the currently predicted 3D gaze direction of the face and the both eyes.

$$e_f = \arccos(\frac{g_f \cdot f(I_f)}{\|g_f\| \, \|f(I_f)\|}) \tag{1}$$

$$e_l = \arccos(\frac{g_l \cdot f(I_l)}{\|g_l\| \, \|f(I_l)\|}) \tag{2}$$

$$e_r = \arccos(\frac{g_r \cdot f(I_r)}{\|g_r\| \, \|f(I_r)\|}) \tag{3}$$

where $f(I)$ represents the predicted value of the gaze vector (the gaze regression), and g represents the ground truth of the gaze vector. We then calculate the weighted average error of the three errors.

$$e = \lambda_f \cdot e_f + \lambda_l \cdot e_l + \lambda_r \cdot e_r \tag{4}$$

where λ_f, λ_l, and λ_r determine the errors of the face, the left eye and the right eye, respectively. If the image of which

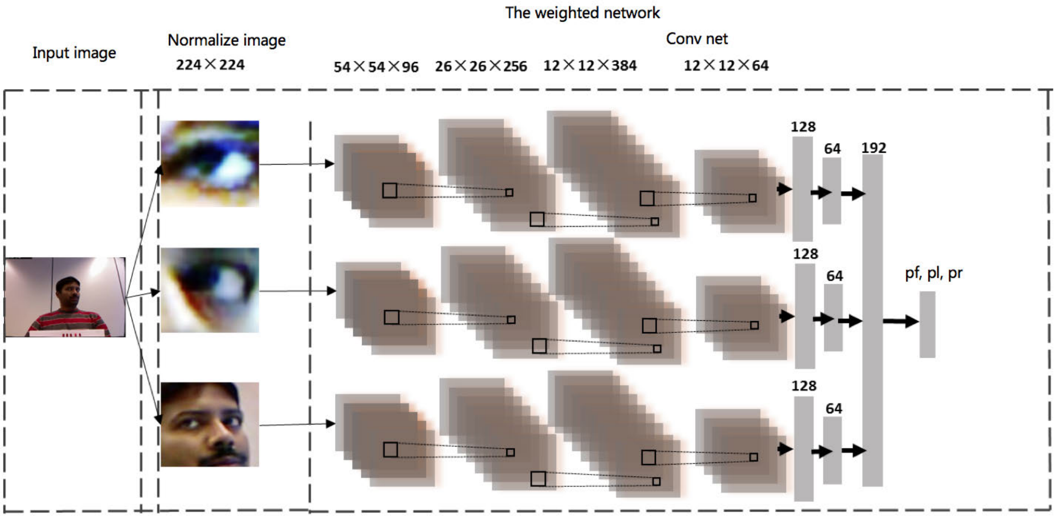


FIGURE 5. **The weighted network.**

region is more likely to produce smaller errors, the weight of the network should be increased when optimizing the network. With this concept in mind, we propose the following formula to set the weights.

$$\begin{cases} \lambda_f = \dfrac{1/e_f}{1/e_f + 1/e_l + 1/e_r} \\ \lambda_l = \dfrac{1/e_l}{1/e_f + 1/e_l + 1/e_r} \\ \lambda_r = \dfrac{1/e_r}{1/e_f + 1/e_l + 1/e_r} \end{cases} \tag{5}$$

Considering that the error between the predicted value and the actual target value in the images of the three regions will be different, we calculate the mean square error between the predicted value and the target value.

$$\begin{cases} predicted_t = [f(I_f), f(I_l), f(I_r)] \\ observed_t = [g_f, g_l, g_r] \\ MSE = \dfrac{1}{N}\sum_{t=1}^{N}(observed_t - predicted_t)^2 \end{cases} \tag{6}$$

By combining Eqs. (4), (5) and (6), we can get the final Loss function.

$$\mathrm{LR} = \mathrm{MSE} + 3\frac{e_f \cdot e_r \cdot e_l}{e_r \cdot e_f + e_l \cdot e_f + e_l \cdot e_r} \tag{7}$$

D. WEIGHTED NETWORK

As mentioned above, the regression network can predict the gaze vector by the high-quality face and eye images. We then design the weighted network to learn the selection of the regression network and show its dependence on different regional characteristics in the optimization process.

As shown in Fig. 5, the network is a three-stream convolutional neural network. Each stream extracts 64-dimensional deep features from the face, left eye and right eye respectively. A simplified version of the Alexnet [37] is the basic network of each branch followed by a 3-dimensional fully connected layer. Finally, the Softmax regressor is used to get the probability bias vector $[p_f, p_l, p_r]^T$ of the corresponding face and both eyes.

In order to train the weighted network to predict the choice of regression network, we set the following Loss function.

$$\begin{cases} predicted_t = [p_f, p_l, p_r] \\ observed_t = [p_{tf}, p_{tl}, p_{tr}] \\ p_{tf} + p_{tl} + p_{tr} = 1 \\ MSE = \dfrac{1}{N}\sum_{t=1}^{N}(observed_t - predicted_t)^2 \end{cases} \tag{8}$$

where p_f is the probability that the regression network depends on the face region in the prediction process. p_l and p_f are the probabilities that depend on the left eye and the right eye respectively.

During training, the ground truth of p is determined by the gaze vector error from regression network. Taking the P_{tf} as an example, P_{tf} is set to be 1 if $e_f < e_l$ and $e_f < e_r$, and P_{tf} is set to be 0 in other cases. In other words, when the error of the face in the regression network is the smallest, we should choose to maximize p_f to learn the fact to realize the adjustment of the regression network. Similarly, p_{tl} is set to be 1 when e_l is the minimum; otherwise p_{tl} is 0. When the value of e_r is the minimum, P_{tr} is set to be 1; otherwise P_{tr} is 0.

The aim of the weighted network is to adjust the regression network to improve the accuracy of gaze estimation. For this purpose, the Loss function of the regression network is adjusted to

$$\begin{aligned} \mathrm{LR}^* = \mathrm{MSE} & \\ & + w \cdot 3\frac{e_f \cdot e_r \cdot e_l}{e_r \cdot e_f + e_l \cdot e_f + e_l \cdot e_r} \\ & + (1-w) \cdot \beta \cdot \frac{e_f + e_r + e_l}{3} \end{aligned} \tag{9}$$

where w is to balance the weight between the learning of left eye, right eye and face. The gaze vector depends on the input images of the regression network. If the ground truth of gaze vector (g_f, g_l, g_r) are approximately the same, we should not increase the weight of any area in the learning of the regression network. When the gaze vector (g_f, g_l, g_r) are greatly different, we prefer to train a certain region with small error in the regression network. The adaptive adjustment is realized by determining the output (p_f, p_l, p_r) of the weighted network. In an ideal situation, p_f, p_l, p_r can present extreme values of 0 or 1, allowing the network to select areas that can generate small errors and have high image quality for training to improve the accuracy of the results. In the actual training process, p_f, p_l, p_r will only be a value between 0 and 1. The calculation is stated as follows.

$$\begin{aligned} w = (1 &+ (2a-1)\cdot p_f \\ &+ (2b-1)\cdot p_l \\ &+ (1-2a-2b)\cdot p_r)/2 \end{aligned} \tag{10}$$

where a = 1 if $e_f < e_l$ and $e_f < e_r$, otherwise a = 0; b = 1 if $e_r < e_l$ and $e_r < e_f$, otherwise b = 0. During the experiment, w is the decimal number between 0 and 1.

III. EXPERIMENTAL EVALUATION

To verify the effectiveness of the proposed 3D gaze estimation method, we evaluate it on two publicly available datasets: MPIIGaze [28] and EyeDiap [38].

Firstly, we cross-validate the method to demonstrate the performance of our algorithm. Then, we perform ablation experiments to evaluate the contributions of different regional images on the network. Next, we do experiments with different resolutions to show the robustness of the proposed network. Finally, we evaluate the effectiveness of the weighted network on the gaze vector. In this paper, we use the angle difference between the prediction vector and the ground truth vector to represent the accuracy of gaze estimation.

A. DATASETS

The MPIIGaze dataset consists of 213,659 images of 15 participants, including various illumination conditions, eye appearance and head posture. It's worth noting that we need to do normalization for the images and data of the MPIIGaze. We use the center of six facial markers provided in the dataset as the starting point of the gaze vector. The starting point of the gaze vector is also the facing point of the virtual camera in the normalization process. To reduce the influence of illumination difference, each input image is equalized by the adaptive histogram. To facilitate comparison with other state-of-the-art gaze estimation methods, we perform leave-one-person-out cross-validate on participants in the same way.

The EyeDiap dataset is a video data set of 16 participants, including various illuminations, scenes and head postures. We select an image per 15 frames from each video clip and filter out the frames that satisfy the following conditions: (1) participants do not look at the screen; (2) the annotations are not provided correctly; (3) the gaze angle violates the physical constraints (where: elevation angle theta (φ) $<=$ 40°, azimuth angle phi (θ) $<=$ 30°).

Similar to the MPIIGaze, the EyeDiap also needs to apply normalization firstly. We use the midpoint of two iris centers provided in the dataset as the origin of the gaze vector. We apply the adaptive histogram equalization to reduce illumination changes. The gaze targets on this data set are divided into two categories: screen targets and floating targets. To facilitate comparison, we use only screen targets for evaluation and divide 14 participants into four groups for leave-one-group-out cross-validation.

B. CROSS PERSON/GROUP EVALUATION

The proposed method is compared with the state-of-the-art 3D gaze estimation methods on the MPIIGaze and the Eyediap datasets. Tables 1 and 2 show the comparison results on both datasets, respectively. According to the comparison results, our method achieves superior performance both in the MPIIGaze and the Eyediap datasets. Fig. 6 and Fig. 7 show part of the prediction results of our method on the MPIIGaze and EyeDiap datasets, where the green and red lines represent the prediction results and ground truth of gaze vector respectively. Our method is robust that can maintain high prediction accuracy under the circumstance of various illumination difference and large head postures.

TABLE 1. Comparison results with the state-of-the-art methods on the MPIIGaze dataset.

Methods	3D degrees error
Cheng et al.2018[29]	13.5°
Nie et al. 2018[39]	7.1°
Krafka et al.2016[27]	6.7°
Zhang et al. 2015[26]	6°
Zhang et al. 2017[40]	5.4°
Zhang et al. 2017[28]	4.8°
Zhou et al. 2019[41]	4.18°
Our method	**2.75°**

TABLE 2. Comparison results with the state-of-the-art methods on the Eyediap dataset.

Methods	3D degree error
Cheng et al. 2018[29]	8.8°
Krafka et al. 2016[27]	8.3°
Park et al. 2018[34]	7.4°
Zhang et al. 2017[28]	6.0°
Zhou et al. 2019[40]	5.84°
Our method	**4.42°**

C. NETWORK EVALUATION

To verify the role of each module in the network, the network is divided into monocular module, face module, face + monocular module, binocular module and

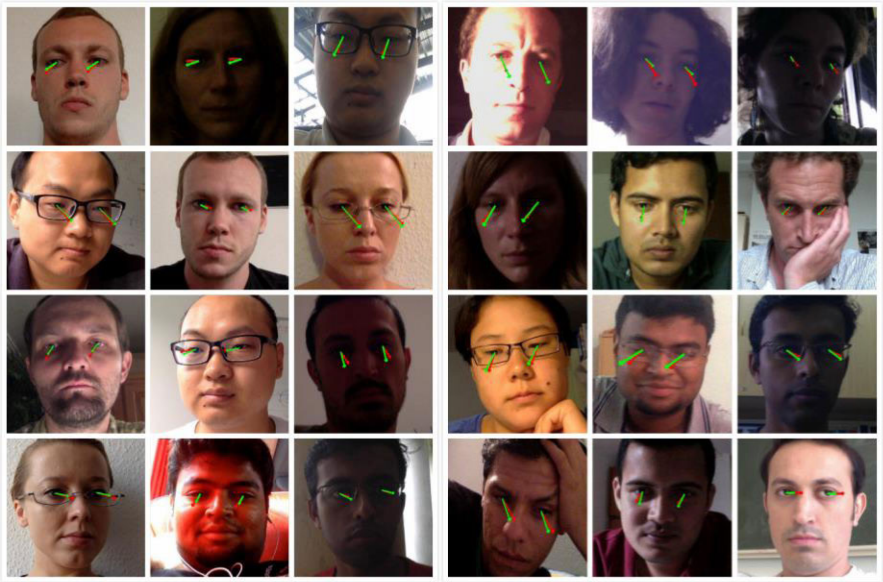


FIGURE 6. Some prediction results of our method on the MPIIGaze.

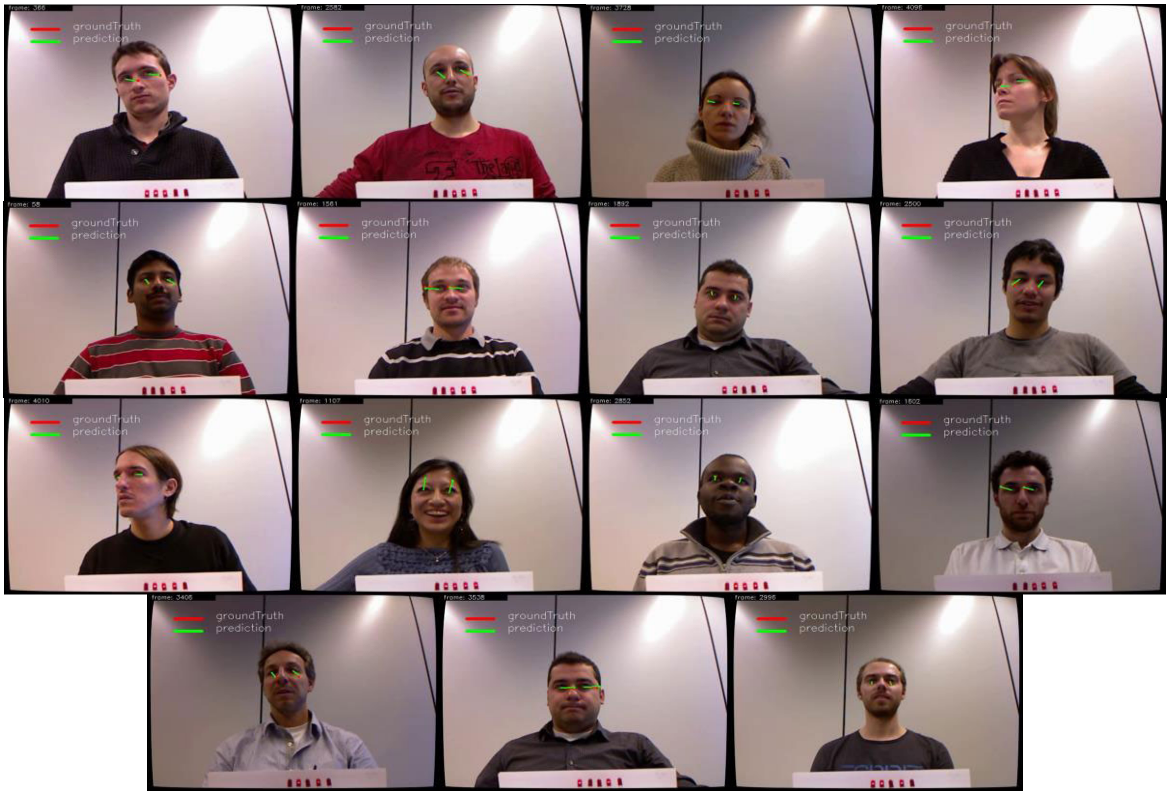
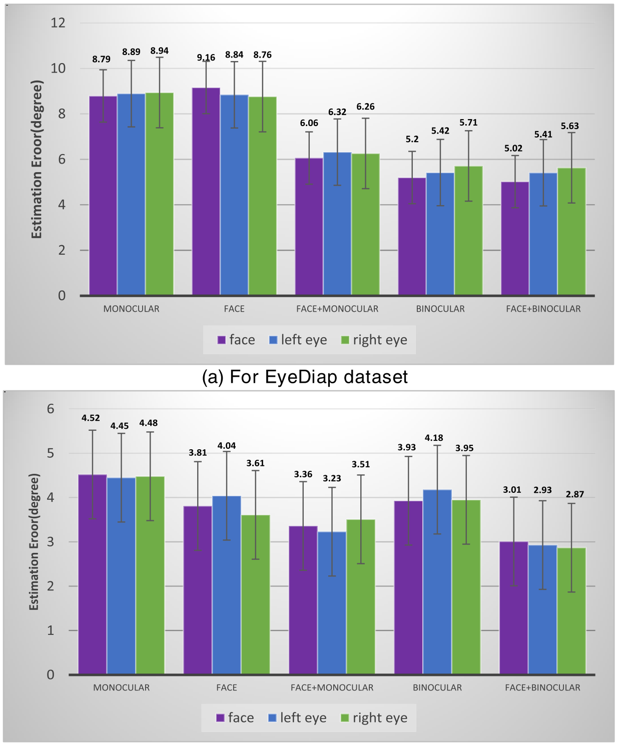


FIGURE 7. Some prediction results of our method on the EyeDiap.

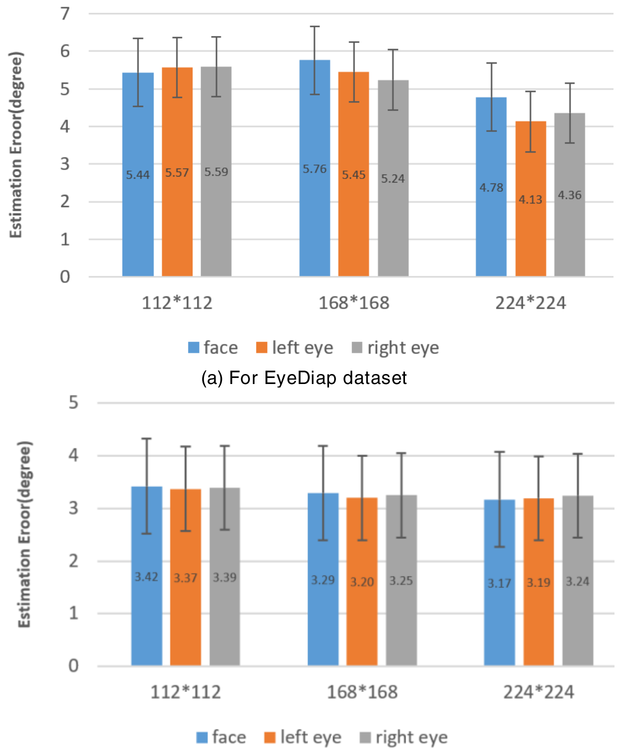
face + binocular module for evaluation. Fig. 8(a) shows the evaluation results of each module on the MPIIGaze dataset. It can be seen from the figure that the order of contributions on the final estimation accuracy from large to small is monocular, face, face + monocular, binocular, and face + binocular. The contribution of a single face branch is greater than that of a single eye branch on the final estimation results. However, the binocular module is more suitable for the learning strategy of regression + weighted network, and the final accuracy of the face + binocular module after adding the face



(a) For EyeDiap dataset

(b) For MPIIGaze dataset

FIGURE 8. Network Parts evaluation on the MPIIGaze and EyeDiap datasets.



(a) For EyeDiap dataset

(b) For MPIIGaze dataset

FIGURE 9. Ablation study on the MPIIGaze and EyeDiap datasets.

module is the best. Similarly, Fig. 8(b) shows the evaluation results of each module on the EyeDiap dataset. Face branch plays a better role in the prediction of line of sight than eye branch, but the function of face + binocular module is better than binocular module and face + monocular module, which is more conducive to the expression + weighted network.

D. RESOLUTION EVALUATION

Gaze estimation method often requires that the model can maintain high accuracy in a certain distance. Although our data are normalized before model training to reduce the resolution difference caused by images with different distances, the loss of some useful information cannot be avoided. This phenomenon may lead to a decline in our forecast results.

Therefore, we need to evaluate the influence of images with different resolutions on our method. In order to simulate this environment, images of 224 × 224 are downscaled to 168 × 168 and 112 × 112 respectively. In order to facilitate the comparison, the final input size of images with different resolutions needs to be consistent, so the two types of images are returned to 224 × 224 through upscaling. We conduct experiments on the MPIIGaze and EyeDiap datasets, and the results are shown in Fig. 9. Our method can maintain good accuracy even when the distance is twice as long as the original distance.

E. WEIGHTED NET EVALUATION

The proposed weighted network is the key technology of the proposed method. In this section, we evaluate the contribute of adding the weighted network to the regression network. We compare the experimental results before and after adding the weighted network, and the comparison experiment is completed on the MPIIGaze datasets. We perform leave-one-person-out cross-validation for the regression network and the adaptive weight adjustment network respectively. As shown in Table 3, gaze estimation results for all the 15 subjects in the MPIIGaze dataset are illustrated. The table shows the gaze vector error of each subject from the starting point of left eye, right eye and face to the target point under the RW-net and R-net. After joining the weight adjustment network, the prediction results generally have significant improvement. However, a few of them have not been improved and Figure 10 shows some evaluation results. Through the comparative analysis of the images, we can see that the negative impact of the weight adjustment network is mainly affected by the illumination. Due to the influence of illumination and other factors, the overall quality of the captured image is low, and the R-net cannot effectively evaluate the weight of regional image features. In the RW-net, it is more likely to use the average error of three regions to calculate the loss value rather than the weighted error of the three regions, which is in conflict with the idea of selecting the region with small error for training in the actual situation, so as to have a negative impact on the final accuracy.

TABLE 3. Gaze angular error comparison for the regression network and the adaptive weight adjustment network for each subject.

Method		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RW-net	*Face*	2.72	3.42	5.68	3.85	2.75	4.63	3.67	5.00	5.37	3.83	5.11	4.62	4.50	4.93	5.26
	Left eye	2.74	3.50	5.67	3.74	2.86	4.88	3.51	5.03	5.29	3.95	5.27	4.70	4.52	4.86	5.16
	Right eye	2.79	3.20	5.79	3.80	2.69	4.72	3.48	5.03	5.39	3.96	5.27	4.71	4.44	4.73	5.17
R-net	*Face*	2.83	3.75	5.92	4.42	3.89	5.02	3.72	4.92	5.75	3.89	6.10	4.17	4.35	4.66	6.18
	Left eye	2.74	3.95	5.91	4.27	3.53	5.10	3.55	4.89	5.81	3.87	5.79	4.31	4.31	4.75	6.19
	Right eye	2.80	3.93	5.81	4.30	3.67	4.95	3.63	4.82	5.79	3.91	5.78	4.41	4.30	4.77	6.30

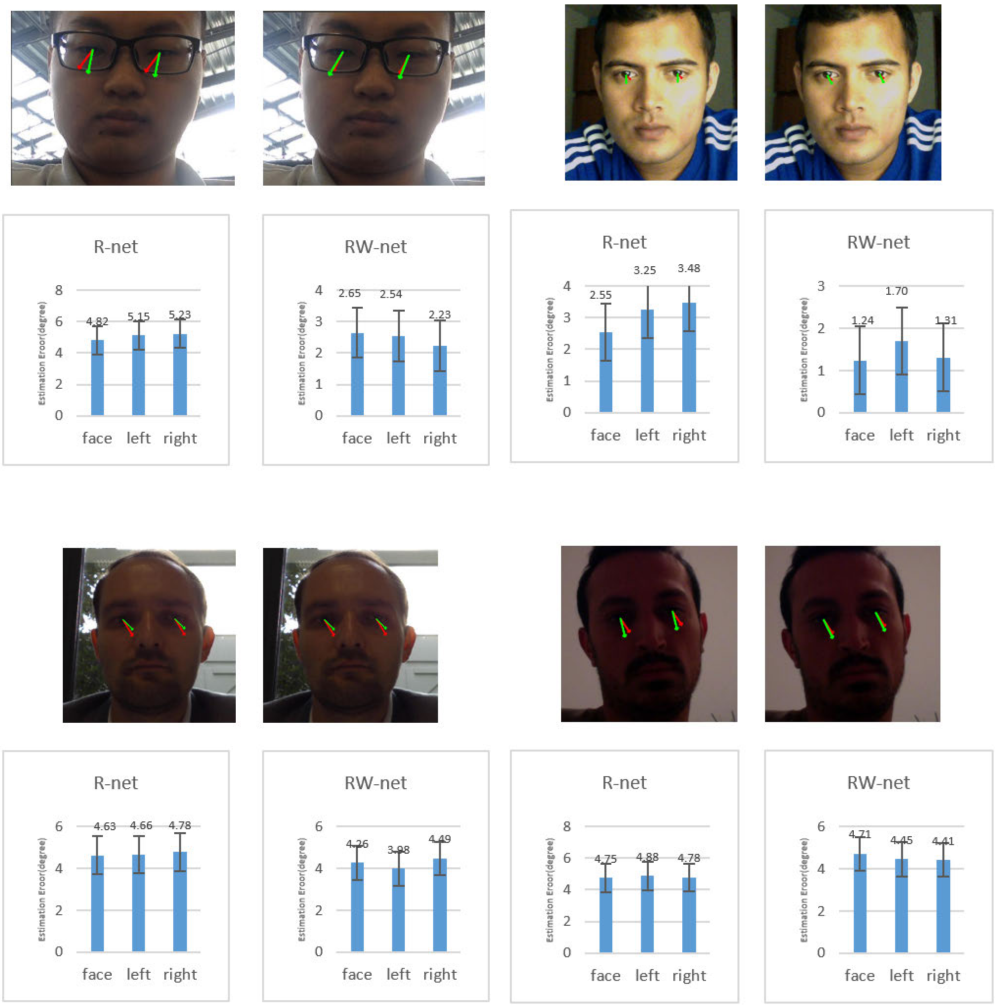


FIGURE 10. Comparison of face and both eyes' gaze errors.

IV. CONCLUSION

This paper has proposed an adaptive weighted 3D gaze estimation method based on deep learning. The method needs to maintain accuracy over a certain distance. We have evaluated the proposed model at different resolutions, and the results have showed that the proposed network has good robustness for images with different resolutions. In order to improve the prediction accuracy, this paper has proposed the weighted network to adjust the regression network. Based on the concept that the weight of which part is weighted based on the small error, the weight adjustment network can adapt the strategy well. Compared with the existing latest line-of-sight estimation methods, our method has significantly improved the accuracy. However, through experimental comparison, the training results of the weighted network under different lighting conditions are not good. Future work will consider how to improve the role of weighted network more effectively and consider using more advanced network structure to further improve its performance.

REFERENCES

[1] C. Pham, S. Rundle-Thiele, J. Parkinson, and S. Li, "Alcohol warning label awareness and attention: A multi-method study," *Alcohol Alcoholism*, vol. 53, no. 1, pp. 39–45, Jan. 2018.

[2] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16495–16519, Aug. 2017.

[3] Y. K. Meena, H. Cecotti, K. Wong-Lin, A. Dutta, and G. Prasad, "Toward optimization of gaze-controlled human–computer interaction: Application to hindi virtual keyboard for stroke patients," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 911–922, Apr. 2018.

[4] H. Cecotti, "A multimodal gaze-controlled virtual keyboard," *IEEE Trans. Human-Mach. Syst.*, vol. 46, no. 4, pp. 601–606, Aug. 2016.

[5] W. Pichitwong and K. Chamnongthai, "An eye-tracker-based 3D point-of-gaze estimation method using head movement," *IEEE Access*, vol. 7, pp. 99086–99098, Jul. 2019.

[6] D. X. Cifu, J. R. Wares, K. W. Hoke, P. A. Wetzel, G. Gitchel, and W. Carne, "Differential eye movements in mild traumatic brain injury versus normal controls," *J. Head Trauma Rehabil.*, vol. 30, no. 1, pp. 21–28, 2015.

[7] O. V. Komogortsev and C. D. Holland, "The application of eye movement biometrics in the automated detection of mild traumatic brain injury," in *Proc. Extended Abstr. 32nd Annu. ACM Conf. Hum. Factors Comput. Syst. (CHI EA)*, 2014, pp. 1711–1716.

[8] H. Cai, X. Zhou, H. Yu, and H. Liu, "Gaze estimation driven solution for interacting children with ASD," in *Proc. Int. Symp. Micro-NanoMechatronics Hum. Sci. (MHS)*, Nov. 2015, pp. 1–6.

[9] J.-S. Kim, S.-R. Kim, B.-C. Yu, and S.-W. Lee, "23-2: A novel low-power OLED driving method based on gaze tracking," in *SID Symp. Dig. Tech. Papers*, May 2018, vol. 49, no. 1, pp. 287–290.

[10] R. D. Watalingam, N. Richetelli, J. B. Pelz, and J. A. Speir, "Eye tracking to evaluate evidence recognition in crime scene investigations," *Forensic Sci. Int.*, vol. 280, pp. 64–80, Nov. 2017.

[11] J. Attard and M. Bindemann, "Establishing the duration of crimes: An individual differences and eye-tracking investigation into time estimation," *Appl. Cognit. Psychol.*, vol. 28, no. 2, pp. 215–225, Mar. 2014.

[12] J. Peth, J. S. C. Kim, and M. Gamer, "Fixations and eye-blinks allow for detecting concealed crime related memories," *Int. J. Psychophysiol.*, vol. 88, no. 1, pp. 96–103, Apr. 2013.

[13] K. Wang and Q. Ji, "Real time eye gaze tracking with 3D deformable eye-face model," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1003–1011.

[14] K. Wang and Q. Ji, "3D gaze estimation without explicit personal calibration," *Pattern Recognit.*, vol. 79, pp. 216–227, Jul. 2018.

[15] T. Pfeiffer, "Towards gaze interaction in immersive virtual reality: Evaluation of a monocular eye tracking set-up," in *Proc. Virtuelle und Erweiterte Realität-Fünfter Workshop der GI-Fachgruppe VR/AR*, 2008, pp. 81–92.

[16] N. Markuš, M. Frljak, I. S. Pandžić, J. Ahlberg, and R. Forchheimer, "Eye pupil localization with an ensemble of randomized trees," *Pattern Recognit.*, vol. 47, no. 2, pp. 578–587, Feb. 2014.

[17] Z. Zhu and Q. Ji, "Novel eye gaze tracking techniques under natural head movement," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 12, pp. 2246–2260, Dec. 2007.

[18] E. D. Guestrin and M. Eizenma, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 1124–1133, Jun. 2006.

[19] C. Hennessey, B. Noureddin, and P. Lawrence, "A single camera eye-gaze tracking system with free head motion," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*, 2006, pp. 27–29.

[20] S.-W. Shih and J. Liu, "A novel approach to 3-D gaze tracking using stereo cameras," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 34, no. 1, pp. 234–245, Feb. 2004.

[21] X. Zhou, H. Cai, Y. Li, and H. Liu, "Two-eye model-based gaze estimation from a kinect sensor," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1646–1653.

[22] Y. Zhang, A. Bulling, and H. Gellersen, "Discrimination of gaze directions using low-level eye image features," in *Proc. 1st Int. workshop Pervasive Eye Tracking Mobile Eye-Based Interact. (PETMEI)*, 2011, pp. 9–14.

[23] Y. Wang, T. Shen, G. Yuan, J. Bian, and X. Fu, "Appearance-based gaze estimation using deep features and random forest regression," *Knowl.-Based Syst.*, vol. 110, pp. 293–301, Oct. 2016.

[24] A. Kacete, R. Séguier, J. Royan, and M. Collobert, "Unconstrained Gaze Estimation Using Random Forest Regression Voting," in *Proc. ACCV*, 2016, pp. 419–432.

[25] 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 Appl.*, vol. 70, no. 3, pp. 2037–2062, Jun. 2014.

[26] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4511–4520.

[27] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2176–2184.

[28] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2299–2308.

[29] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *Proc. ECCV*, 2018, pp. 105–121.

[30] C. Palmero, J. Selva, M. Ali Bagheri, and S. Escalera, "Recurrent CNN for 3D gaze estimation using appearance and shape cues," 2018, *arXiv:1805.03064*. [Online]. Available: http://arxiv.org/abs/1805.03064

[31] T. Fischer, H. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," in *Proc. ECCV*, 2018, pp. 339–357.

[32] Y. Yu, G. Liu, and J.-M. Odobez, "Deep multitask gaze estimation with a constrained landmark- gaze model," in *Proc. ECCV*, 2018, pp. 456–474.

[33] S. Park, A. Spurr, and O. Hilliges, "Deep pictorial gaze estimation," 2018, *arXiv:1807.10002*. [Online]. Available: http://arxiv.org/abs/1807.10002

[34] S. Park, X. Zhang, A. Bulling, and O. Hilliges, "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings," in *Proc. ACM Symp. Eye Tracking Res. Appl. (ETRA)*, 2018, pp. 21–30.

[35] K. Wang, R. Zhao, and Q. Ji, "A hierarchical generative model for eye image synthesis and eye gaze estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 440–448.

[36] K. Wang, R. Zhao, H. Su, and Q. Ji, "Generalizing eye tracking with Bayesian adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11907–11916.

[37] J. Deng, W. Dong, R. Socher, L. J. Li, and F. F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.

[38] K. Alberto, 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," in *Proc. Symp. Eye Tracking Res. Appl.*, 2014, pp. 255–258.

[39] S. Nie, M. Zheng, and Q. Ji, "The deep regression Bayesian network and its applications: Probabilistic deep learning for computer vision," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 101–111, Jan. 2018.

[40] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Mpiigaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 49, no. 1, pp. 162–175, Jan. 2017.

[41] X. Zhou, J. Lin, J. Jiang, and S. Chen, "Learning a 3D gaze estimator with improved Itracker combined with bidirectional LSTM," in *Proc. ICME*, 2019, pp. 850–855.

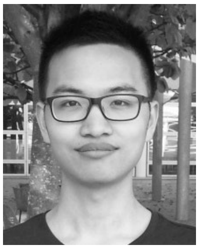
XIAOLONG ZHOU (Member, IEEE) received the Ph.D. degree in mechanical engineering from the Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Hong Kong, in 2013. From 2014 to 2019, he was an Associate Professor with the Zhejiang University of Technology, Zhejiang, China. From 2015 to 2016, he was a Senior Research Fellow with the School of Computing, University of Portsmouth, Portsmouth, U.K. He joined Quzhou University, Quzhou, China, in 2020. He has authored over 80 peer-reviewed articles in international journals and conference papers. His research interests include visual tracking, gaze estimation, 3-D reconstruction, and their applications in various fields. He received the Best Paper Awards at ROBIO2012 and PRCV2018 and the ICRA2016 CEB Award for Best Reviewers.

JIAQI JIANG was born in Zhejiang, China, in 1995. She received the B.S. degree in information management and systems from the Zhijiang College of Zhejiang University of Technology, Shaoxing, Zhejiang, China, in 2017, and the M.S. degree in software engineering from the Zhejiang University of Technology, Hangzhou, Zhejiang, in 2020. Her current research direction is gaze estimation.

QIANQIAN LIU was born in Henan, China, in 1996. She received the B.S. degree in computer science and technology from Xinyang Normal University, Xinyang, Henan, China, in 2018. She is currently pursuing the M.S. degree in computer science and technology with the Zhejiang University of Technology, Hangzhou, Zhejiang, China. Her current research direction is event-based object tracking.

JIANWEN FANG received the M.Eng. and Ph.D. degrees in computer science from Zhejiang University, China, in 2007 and 2013, respectively. He is currently working as a Professor with the School of Electrical and Information Engineering, Quzhou University, China. His research interests include intelligent information processing, intelligent image processing, computer graphics, and computer animation.

SHENGYONG CHEN (Senior Member, IEEE) received the Ph.D. degree in computer vision from the City University of Hong Kong, Hong Kong, in 2003. He is currently a Professor with the Tianjin University of Technology and the Zhejiang University of Technology, China. He received a fellowship from the Alexander von Humboldt Foundation of Germany and worked at the University of Hamburg, from 2006 to 2007. His research interests include computer vision, robotics, and image analysis. He has published over 100 scientific articles in international journals. He is a Fellow of IET and a Senior Member of CCF. He received the National Outstanding Youth Foundation Award of China, in 2013.



HAIBIN CAI received the M.S. degree in computer science from the Zhejiang University of Technology, Hangzhou, China, in 2015, and the Ph.D. degree in computer science from the University of Portsmouth, Portsmouth, U.K., in 2018. He is currently a Lecturer with the Department of Computer Science, Loughborough University. His research interests include gaze estimation, motion recognition, facial expression recognition, object tracking, image processing, and machine learning.