Research on the Estimation of Gaze Location for Head-eye Coordination Movement

Qiyou Wu
School of Computer Science and Engineering Xi'an Technological University
Xi'an, China
E-mail: wu314650592@163.com

Changyuan Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: Cyw901@163.com

***Abstract*—Sight is the main source for humans to obtain information from the outside world. Due to the structure of the human eye [1], the range of human sight is limited. For this reason, people need to constantly move their line of sight when observing the surrounding environment and the target, and the movement of the sight is based on the coordinated movement of the head and the eye[2]. Therefore, the key issue for gaze research is how to correctly establish the relationship between head-eye movement and gaze movement. Taking the simulated flight environment as the research background, this paper collects a large number of head-eye images through the designed "three-camera and eight-light source" head-eye data acquisition platform, and proposes a gaze estimation method based on the combination of appearance and features, which effectively combines The relationship of head-eye coordination movement. Then, the ResNet-18 deep residual network structure and the traditional BP neural network structure are used to complete the effective fusion of the head pose and human eye features in the process of capturing the sight target, so as to realize the accurate estimation of the sight drop point, and its average accuracy up to 89.9%.**

Keywords-Head-Eye Coordination; Gaze Estimation Method；Experimental Platform Design; Deep Residual Networks

1. INTRODUCTION

Head-eye coordination is the process of coordinating and combining head and eye movements and synthesizing a unified action to complete the shift of sight to the target [3]. In the research of visual impact point estimation, it is necessary to establish the relationship between head movement, eye movement and gaze movement in order to obtain a more accurate gaze point location. Therefore, how to establish the head-eye-line of sight relationship is the key issue of research.

For the research on the relationship between head-eye-line of sight, the initial method is to limit the free movement of the human head and only track the eye movement [4], so that a slight movement of the head will cause a large systematic error, and it is not suitable for in practical application scenarios. For this reason, the mechanism of head-eye movement has become a hot research topic at that time. In 2008, Freedman [5] used physiological methods to study the relationship between eye movement and head-eye movement in rhesus monkeys, which are similar to human head-eye movement mechanisms. The experimental results show the relationship between the head-eye movement and the line of sight: when the target appears in a larger field of view, the eyes will first move towards the target before the head, and then the head starts to move in the same direction. Due to the fast movement of the eyes, the sight can quickly complete the target acquisition and stop moving. However, the head movement was relatively slow, and Ren did not stop moving in the target direction. At this time, under the action of the vestibular function, the tendency of the eyes to move in the opposite direction at a certain speed is used as a vestibulo-ocular reflex (VOR) [6] to compensate for the head movement, so as to ensure that the target exists stably in the line of sight. Inside. It can be seen that although the contribution of head movement to the target capture process is small, it also directly affects the direction of sight movement.

Gaze estimation is a study of the subject's current gaze direction or gaze location using existing detection technologies such as mechanical, electronic, and optical [7]. The early research on eye sight estimation benefited from the development of medicine and psychology, and researchers recorded the relevant information of eye movement by direct observation. Based on the different devices used for gaze estimation, gaze estimation research can be divided into wearable and non-wearable [8]. With the development of sight estimation technology, wearable sight estimation methods such as contact lens and electrooculography (EOG) have appeared. Although the influence of head movement is reduced, it is more disturbing to the subjects. , not suitable for long-term wear. With the development of image processing and computer vision technology, the advantages of video-based line-of-sight estimation methods are convenient and non-wearable. However, for gaze estimation research, how to efficiently integrate head motion data still needs further research.

In recent years, under the research upsurge of deep neural network, new progress has been made in the research of gaze estimation algorithm based on head-eye data fusion, which can be mainly divided into gaze target estimation, gaze location estimation and gaze direction estimation. In 2016, Recasens [9] et al. designed a deep neural network model composed of two branches, which were used to extract the head pose and gaze direction of human images respectively, and to estimate the gaze target by judging the saliency of the target; Kyle Krafka [10] et al. took mobile phones and tablet computers as the research objects, and designed a deep neural network composed of four branches, respectively inputting left and right eye images, face images and face positions, and realized a two-dimensional plane. line-of-sight estimate. The idea of sharing and processing the parameter weights of the left and right eye image branches in this study has been used for reference by many subsequent studies; in 2019, the Google [11] team further improved the above model and changed the line of sight estimation model to three The branch and the coordinate positions of the four corners of the eyes are used to replace the face image and the face position, and finally a good estimation effect of the line of sight is achieved; for the estimation of the line of sight, it is usually represented by a direction vector formed by two horizontal and vertical angles. , Zhang [12]'s research in 2015, spliced the head pose data of the input image with the eye features, and used a shallow network structure similar to LeNet [13] to estimate the line of sight direction. The way of data fusion has greatly inspired the follow-up research in this paper.

2. MATERIALS

1) Experiment preparation)

The data collection experiment in this paper recruited 8 male graduate student volunteers as subjects, aged 23-30 years old, in good health and with good eyesight. Before the experiment, each subject was familiar with the specific content and precautions of the experiment, and they all participated in the experiment voluntarily. Each subject signed a written commitment and informed letter to ensure the legitimacy of the experiment in this paper. In order to exclude external interference, after the preparation for the experiment, each experiment was completed by only one subject alone.

The experimental equipment includes a DELL computer, an inertial sensor (MTI-G-700), three industrial cameras, three high-definition displays (resolution 2560× 1440), and eight infrared point light sources. Among them, the inertial sensor was worn to about the position of the occipital bone behind the subject's head to measure the Euler angles (Pitch, Roll, Yaw) of the subject's head posture when capturing the target. Since the MTI-G-700 inertial sensor is only used to measure the change of the head posture in this paper, it will be referred to as the head posture instrument in the following. In this study, three Point Grey GS3-U3-41C6NIR-C industrial cameras were selected, and the resolution of the collected images was 2048×2048 and the chromaticity was near-infrared (NIR). Three cameras were installed above the three high-definition monitors, and were used to simultaneously capture the head motion images

and eye motion images captured by the subjects during the experiment.

For the data collection experiment of head-eye coordination movement, this study innovatively built a non-wearable sight-drop data collection platform of "three eyes and eight light sources". This platform not only expands the subject's head movement range, but also effectively reduces the impact of changes in lighting conditions. It is mainly composed of three industrial cameras mounted on three monitors and eight near-infrared light sources evenly distributed on the border of the monitors, which are used to record the head-eye images of the subjects when the target is captured. The schematic diagram of the platform deployment is shown in Figure 1.

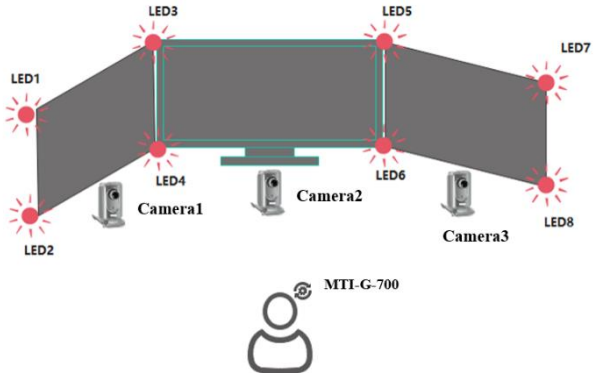

Figure 1. Deployment of the "Three Cameras and Eight Lights" platform

2) Maintaining the Integrity of the Specifications

The head-eye movement data collection experiment designed in this study mainly refers to the process of the subjects performing visual interaction with the randomly appearing objects on the three screens through coordinated head-eye movement. Before starting the experiment, the infrared light source, camera, head attitude meter and other equipment should be calibrated to ensure that each equipment is in normal operation. The subjects were required to wear the head posture meter, and adjust the horizontal distance between the sitting position and the middle screen to save about 60cm to ensure that they were within the best focal length of the three cameras. At the beginning of the experiment, the subjects' eyes need to face the center of the middle screen, and press the record button to calibrate the initial Euler angle of the head posture. After the calibration is successful, the target to be captured appears randomly on the three screens in the form of a red circle with a radius of 30 pixels. The subject uses the head-eye movement to aim at the target, and press the record button to complete the target capture process. . During the experiment, there are no other requirements for the subjects, and the head can move freely in a large range. After the experiment, the program will record the Euler angle of the head pose, head image, eye image and the coordinates of the center point of the target each time the subject captures the target, and set it as a set of data. The experimental process is shown in Figure 2.

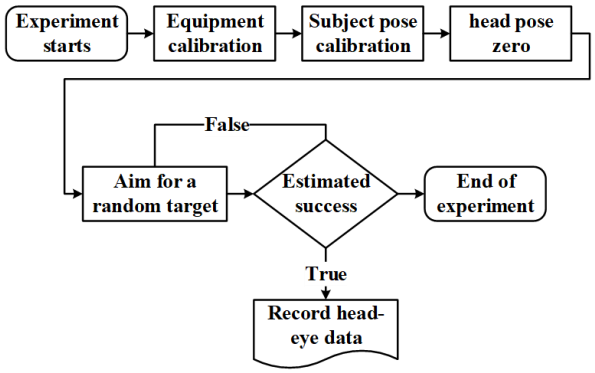

Figure 2. Flow chart of head-eye movement data collection

Considering that the experimental operation is relatively simple, in order to ensure the experimental status of the subjects and the quality of the experimental data, the duration of a single experiment is set to 20 minutes in this study. During the experiment, the equipment was deployed on a six-axis full-motion simulated flight platform, as shown in Figure 3. Due to the long-term use of the camera, the performance will be effectively degraded, and there may be cases of missed shots, so simple manual screening is required after each experiment. Finally, after screening unqualified samples, a total of 31507 groups of head-eye movement data were collected in this paper.

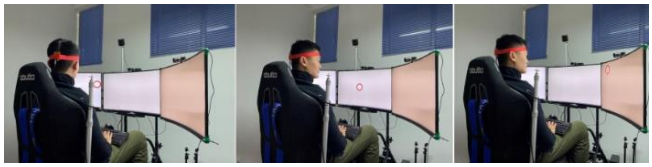

Figure 3. The experimental process of head-eye movement data collection

3. METHODS

In this study, the appearance-based line of sight estimation method is used to estimate the location of the subject's line of sight through head-eye images captured by a non-wearable multi-eye camera. Although the appearance-based line-of-sight estimation method has strong robustness, it also has some problems, such as a great restriction on the free movement of the subject's head and a great influence by the change of lighting conditions. The "three eyes and eight light sources" platform built in this paper can not only expand the subject's head movement range through the strapdown of three monitors; Feature points are added to the external image to reduce the influence of lighting. For this reason, this research uses the method of image processing and feature extraction, fuses head features and eye features, establishes a neural network model for line of sight estimation, and then realizes the research of line-of-sight drop estimation.

1) Head feature extraction

In this paper, the posture measuring instrument was worn on the back of the subject's head and used as the three-axis reference point for head movement. The head attitude data mainly includes: pitch angle (Pitch), yaw angle (Yaw), roll angle (Roll), namely looking up, shaking head and turning head, through these three Euler angles, the head position can be estimated more accurately. space pose. In this study, a right-handed Cartesian coordinate system is used, and the three-axis positions of X, Y, and Z in space and the corresponding Euler angles of the head posture are shown in Figure 4.

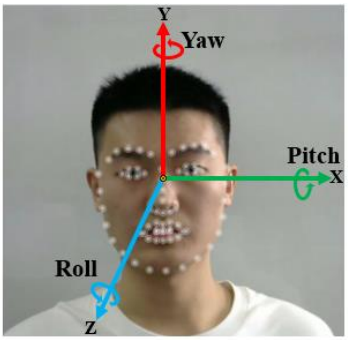

Figure 4. Euler angle of head posture

In order to facilitate the statistics and analysis of the head pose data, this study visualized the recorded data. The three-axis pose data of the head is shown in Figure 5. Through the three-axis Euler angle deflection angle, it can be intuitively observed that the head yaw angle (Yaw) and the pitch angle (Pitch) change greatly during the target acquisition process, while the roll angle (Roll) changes less.

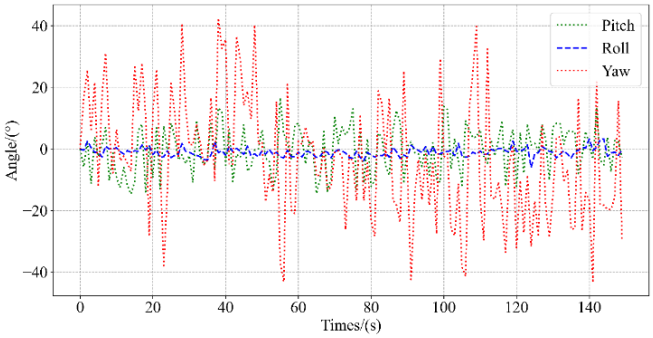

Figure 5. Three-axis attitude data record

2) Eye feature extraction

The human eye detection method used in this study mainly uses the AdaBoost cascade classifier combined with Haar-like features to first detect the face area, and then detects and intercepts the human eye area in the face area. Based on the pre-trained classifiers for faces, eyes, etc. included in OpenCV, this study carried out face detection and eye detection on the front view image collected by the camera corresponding to the target appearing screen. The recognition results are shown in Figure 6. According to the results of human eye detection, the monocular area of the subject is intercepted at a resolution of 64×64, and the left and right eye images are obtained as the input of the next convolutional neural network model.

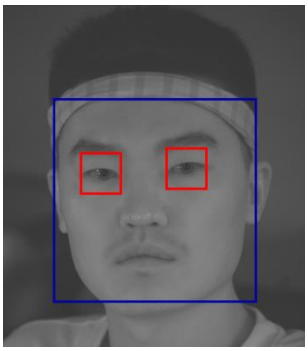

Figure 6. Face detection, Eye detection results

Aiming at the requirement of lighting conditions in appearance-based visual estimation method, this study combines the idea of feature-based visual estimation method, and places eight near-infrared point light sources equidistantly on the boundary of three screens. The Purkinje formed by the reflection are used as the feature points of the eye image. Since the camera is a near-infrared camera, the brightness of the Purkinje formed by the reflection of the infrared point light source through the corner of the eye is not affected by external light. To sum up the above assumptions, this study performed threshold processing on the intercepted left and right eye images before training the line of sight placement model, and obtained left and right eye images with more obvious Purkinje spots, which were used as the control group input by the convolutional neural network model. Its influence on the estimation result of the line-of-sight landing point.

In this study, the three monitors (resolution: 2560×1440) are numbered in the order of left (0), middle (1), and right (2). degree. This study takes the five target points on the left (0) screen as an example for analysis, and the specific positions are shown in Figure 7. Among them, the first two coordinates of each point are the position of the target center pixel on the screen, and the third coordinate is the screen number.

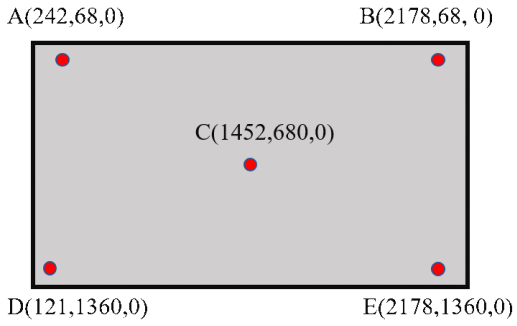

Figure 7. Location of the target center point

The eye images collected at the five points A, B, C, D, and E in the above figure are processed in the order of binarization thresholding, truncation thresholding, and super-thresholding zero. The detection effect is shown in Figure 8. It can be observed from the figure that for different fixation points, the number and positional relationship of Purkinje spots formed by the subjects' eyes are different. Treatment can detect 6-7 Purkinje. Among them, the Purkinje after the truncated thresholding process is more obvious, which can effectively eliminate the reflected light spots on the cornea of other external light sources, and retain the original eye image. Therefore, this paper will use the truncation thresholding method to process the cropped left and right eye images, and detect the left and right eye images with obvious Purkinje as the input of the next convolutional neural network model.

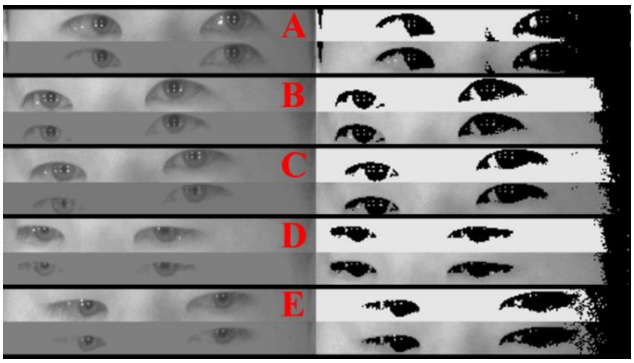

Figure 8. Purkinje detect results

For the deep convolutional neural network model, this paper refers to the deep residual network structure proposed by researchers such as He in 2015. At present, several commonly used ResNet networks mainly include: ResNet-18, ResNet-34, ResNet-50 and other variants. Although increasing the depth of the network can improve the accuracy of the model, the shallower residual network (ResNet-18) also has good accuracy in practical applications, and its model is small, which provides faster convergence speed and facilitates parameter optimization. Moreover, based on the short-circuit operation of the ResNet model, the combination of features of different resolutions can be realized, and it has a better feature extraction effect for the eye image input in this paper. Therefore, in this study, ResNet-18 is used as the estimation model of the human eye line of sight, and the network structure of ResNet-18 is shown in Table 1. By comparing different types of binocular image inputs (with and without thresholding to detect Purkinje), analyze the accuracy of the output on the screen where the sight falls. Since traditional residual neural networks are mostly used for classification tasks, this paper is inspired by the Google team's

research on line-of-sight drop estimation in 2019. Multiple fully-connected layers are connected after the hidden layer of the neural network to return the line-of-sight drop coordinates.

For the estimation model of human eye gaze point in this paper, the input is the grayscale images of the left and right eyes when the subject is facing the capture target in the middle of the screen. Due to the low resolution requirement of the eye image, this paper adopts the resolution size of 64×64 to capture the monocular image. Compared with the input size of the traditional ResNet-18 network RGB image (224×224×3), the image input in this paper is smaller (64×64×1), which improves the computational speed of the model. This model is divided into two branches with the same structure. The input layer is the cropped left and right eye images, and the main structure of the hidden layer of each branch is built according to the ResNet-18 network structure. It consists of 17 convolutional layers, 8 residual blocks and 2 pooling layers. Relu is used as the activation function of all convolutional layers to achieve feature extraction for left and right eye images. Finally, the left and right eye images are extracted through 4 fully connected modules. The feature maps of the eyes are fused and the estimated line-of-sight coordinates G(x, y, n) are output. Figure 9 shows the structure of the network model for the estimation of the human eye gaze point in this paper.

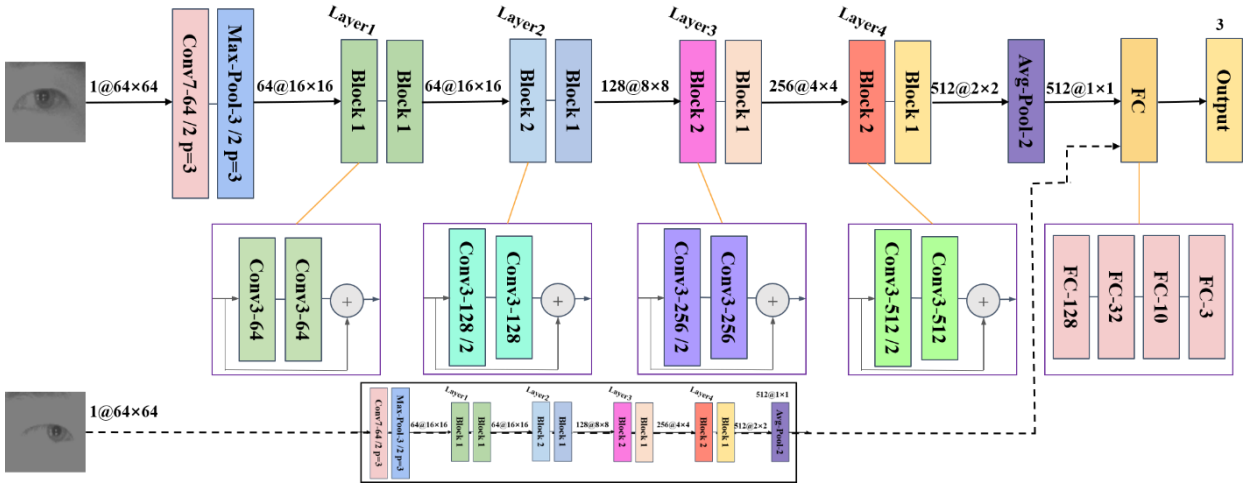

Figure 9. Eye gaze point estimation network structure

3) Eye feature extraction

In order to fuse the head pose data in the line-of-sight drop estimation, this study draws on the idea of using the traditional BP neural network, and uses multiple fully connected layers as the network branch of the head pose data feature extraction, mainly including: an input layer and a Hidden layer composition. Among them, the input layer is the subject's head posture Euler angles (Roll, Pitch, Yaw); the hidden layer is composed of three fully connected layers, the number of neuron nodes is 100, 16, 16 respectively, and Relu is used as the activation function uses the feature vector extracted by the last fully connected layer as the output.

Based on the principle of feature layer fusion, this research first preprocesses the head-eye data to complete feature extraction. Image feature extraction; for the head image, the Euler angle of the head pose is used to output the feature vector with the same dimension as the eye feature through the head feature extraction network to complete the dimension registration. The features of the two parts are fused through multiple fully connected layers to form a line-of-sight estimation model structure fused with head-eye movements. The network structure of the line of sight estimation model in this study includes three branches. The input layer inputs the left and right eye images and the Euler angle of the head pose when the subject is capturing the target. For the left and right eye branches, three fully connected

layers are used for feature extraction, and the number of neuron nodes is 128, 32, and 16 respectively; for the head pose branch, a head feature extraction network is used, and finally two neurons are used for feature extraction. The fully connected modules with the number of nodes are 16 and 3 to complete the feature fusion of the data of the three branches, and realize the regression of the landing point of the three screens, as shown in Figure 10.

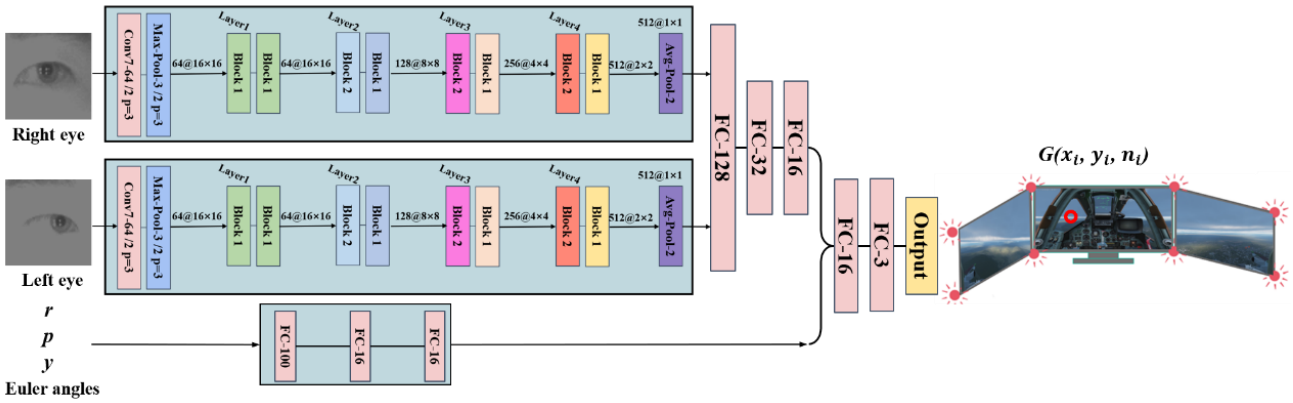

Figure 10. Gaze estimation model with head-eye movement fusion

4. RESULT

In order to improve the estimation performance of the human eye gaze point estimation model in the learning process, this paper uses the mean squared error (MSE) as the loss function. MSE represents the mean value of the sum of squares of the point errors corresponding to the predicted data and the original data:

$$MSE = \frac{1}{n}\sum_{i=1}^{m}\omega_i(y_i - \hat{y}_i)^2 \qquad (1)$$

where n is the number of samples, y_i is the original data, and $\hat{y}_i$ is the predicted data. When the MSE value is closer to 0, it indicates that the fitting ability of the model is stronger, and the estimation of the line of sight is more accurate. Based on the three-screen experimental platform in this paper, each screen is a two-dimensional plane, and the Euclidean distance is used to calculate the difference between the calibration point and the estimated point. The predicted receptive field is a circular area with the calibration point as the center and a radius of 30 pixels. In the training process of the model, this paper uses adaptive moment estimation (Adam) as the optimizer, which can adjust different learning rates for different parameters; the learning rate of the network is set to 10^{-3}, and the batch size is set is 32, and the training epoch is 100.

In this study, through the screening of the original data, after removing the images of the subjects with eyes closed, 30,000 images were selected from a total of 30,000 images and cropped to a size suitable for the model input, of which 70% were used for model training and 30% were used for model training. Performance Testing. At the same time, two different line-of-sight estimation models were constructed and used as a control experiment to compare and analyze the final prediction accuracy and other performances based on whether or not Purkinje detection was performed before inputting the original data.

The experimental results show that when only the original eye image is used as the model input, the model (Eye) needs to extract fewer features and the convergence speed is faster. At about 200 epochs, the model basically converges, and its average accuracy can reach 85.6%; when the input is the eye image after Purkinje detection, although the model (Eye & Purkinje) has a slower convergence speed, it is basically at about 400 epochs. Convergence, but its average accuracy can reach 87.7%.

By comparing the performances of the two models, the (Eye & Purkinje) model makes the Purkinje patch features more obvious through

thresholding before input, which increases the complexity of feature extraction in the hidden layer, thereby increasing the convergence time of the model, but it is relatively slow compared to the Eye model. The average accuracy of line-of-sight location estimation is increased by 2.1%, and the loss curve is relatively stable, and the stability of the model is better. Therefore, in this study, the (Eye & Purkinje) model is selected as the model for estimating the human eye gaze, and it is verified that the input of eye images with significant Purkinje spots can add feature points to the image, reduce the influence of lighting conditions, and improve the estimation accuracy of the model. Accuracy. The accuracy and loss curves of the training of the two models are shown in Figure 11.

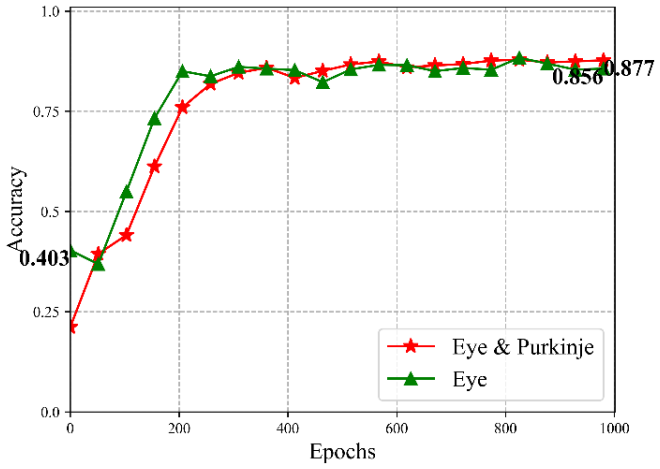

Figure 11. Comparison of the performance of different models

In order to evaluate the performance of the gaze point estimation model fused with head-eye motion, this paper compares it with two models that only use eye images. As shown in Figure 12, in the first 200 epochs, the performance of the model after adding head pose is better than that of the (Eye & Purkinje) model, but due to the need to fuse head-eye features, parameter optimization takes a long time, and the accuracy is not as good as the Eye model. From the analysis of the convergence speed of the model, although the (Eye & Purkinje & Head) model converges slowly, the model tends to be stable after 600 epochs, and the accuracy of the model is high. Enter the model, which compresses and correlates head-eye coordination motion data by fusing multi-dimensional head-eye data, and its accuracy is improved by up to 4.3%.

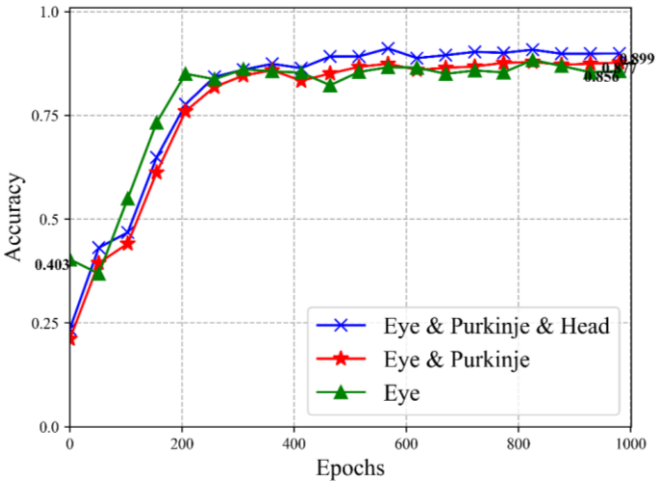

Figure 12. Comparison of the performance of three gaze estimation models

5. Conclusion

For the research on gaze point estimation, this paper uses three models for comparative analysis, namely, the gaze point estimation model using only eye images, the gaze point estimation model using eye features, and the fusion head-eye motion feature. Line-of-sight estimation model. By comparison, the line of sight estimation model (EPH) that fuses head motion and eye feature point images has a high test accuracy for the prediction results of the test set, and the estimated line of sight is basically within the prediction receptive field of the target to be captured. And there is no over-fitting phenomenon, and the average accuracy of the line of sight estimation can reach 89.9%. However, this paper also finds that the accuracy of the general estimation of these three models is not high. This problem is a common problem in the estimation method of line-of-sight placement based on appearance, which needs to be further studied and prospected.

To sum up, this paper is based on a line-of-sight estimation method that combines appearance and features, which effectively integrates the head motion and eye motion when the line of sight moves. Accurate estimation of the line-of-sight placement in a two-dimensional screen is achieved. A practical and effective research method is put forward for the estimation of sight drop point.

REFERENCES

[1] Atchison D A, Smith G, Smith G. Optics of the human eye[M]. Oxford: Butterworth-Heinemann, 2000.

[2] Wang Changyuan, Li Jingjing, Jia Hongbo, et al. Research methods and progress of head-eye movement[J]. Journal of Xi'an University of Technology, 2012, 32(3): 173-182.

[3] Mao Xiaobo. Research on Modeling and Control of Bionic Robot Eye Movement System[D]. Zhengzhou: Zhengzhou University, 2011.

[4] Lei Zhihui, Yu Qifeng. A new method to determine eye movement translation[J]. Experimental Mechanics, 2003, 18(4): 564-568.

[5] Freedman E G. Coordination of the eyes and head during visual orienting[J]. Experimental brain research, 2008, 190(4): 369-387.

[6] Mao Xiaobo, Chen Tiejun. A bionic model of head-eye coordination motion control[J]. Journal of Biomedical Engineering, 2011, 28(5): 895-900.

[7] Liu Jiahui, Chi Jiannan, Yin Yixin. Review of feature-based gaze tracking methods [J]. Journal of Automation, 2021, 47(2): 252-277.

[8] Zhang C, Chi J N, Zhang Z H, et al. Gaze estimation in a gaze tracking system[J]. Science China Information Sciences, 2011, 54(11): 2295-2306.

[9] Recasens ,A R C. Where are they looking?[D]. Massachusetts Institute of Technology, 2016.

[10] Krafka K, Khosla A, Kellnhofer P, et al. Eye tracking for everyone[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2176-2184.

[11] He J, Pham K, Valliappan N, et al. On-device few-shot personalization for real-time gaze estimation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019: 0-0.

[12] Zhang X, Sugano Y, Fritz M, et al. Appearance-based gaze estimation in the wild[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4511-4520.

[13] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

基于头眼协调运动的视线落点估计研究

吴其右
西安工业大学计算机科学与工程学院
中国，西安
E-mail: wu314650592@163.com

王长元
西安工业大学计算机科学与工程学院
中国，西安
E-mail: Cyw901@163.com

***摘要：*视线是人类获取外界信息的主要来源。由于人眼的构造[1]，人的视线范围是有限的。为此，人在观察周围环境与目标时需要不断地移动视线，而视线的移动又是基于头眼协调运动同时构成的，即当人大范围转移视线时，需要依靠头部的运动以扩大视野范围[2]。因此，对于视线研究的关键问题是如何正确地建立头眼运动与视线移动的关系。本文以模拟飞行环境为研究背景，通过所设计的“三目八光源”头眼数据采集平台采集大量的头眼图像，并提出一种基于外观与特征相结合的视线估计方法，有效地结合了头眼协调运动的关系。随后，利用 ResNet-18 深度残差网络结构与传统的 BP 神经网络结构完成对视线目标捕获过程中的头部姿态与人眼特征的有效融合，实现对视线落点的精确估计，其平均准确度可达 89.9%。**

关键字：头眼协调；实验平台设计；深度残差网络；视线估计方法

1. 介绍

头眼协调运动是协调组合头部、眼部运动并合成统一动作从而完成视线转移向目标的过程[3]。在视觉落点估计的研究中，需要建立头部运动、眼部运动和视线移动的关系，才能得到较为准确的视线落点位置。因此，如何建立头眼-视线关系是研究的关键问题。

对于头眼-视线关系的研究，最初的方式是通过限制人的头部自由运动，仅对眼部运动进行跟踪[4]，使得头部稍有移动便造成很大的系统误差，且不适应于实际应用场景。为此，头眼运动机制成为了当时研究的热点。2008 年，Freedman E G[5]使用生理学方法对与人类头眼运动机制相似的猕猴作为研究对象，测得视线移动与头眼运动的关系。其实验结果表明了头眼运动与视线的关系：当目标出现于较大视野范围中时，双眼会率先于头部朝目标进行移动，随后头部开始同向移动。由于双眼运动速度较快，视线可快速地完成目标捕获并停止运动。但头部运动较为缓慢，任未停止向目标方向移动。此时，在前庭功能的作用下，趋势双眼以一定的速度反向移动，作为补偿头部运动的前庭动眼反射(vestibulo-ocular reflex, VOR)[6]，保证目标稳定地存在于视线范围内。由此可得，虽然头部运动对目标捕获过程的贡献较小，但其也直接影响了视线移动的方向。

而视线估计是利用机械、电子、光学等现有检测技术对受试者当前视线方向或视线落点的研究[7]。早期的视线估计研究得益于医学与心理学的发展，研究者多以直接观察的方式记录眼睛运动的相关信息。基于视线估计所使用设备的不同，可将视线估计研究划分为穿戴式与非穿戴式[8]。随着视线估计技术的发展，出现了如接触镜(Contact Lens)和眼电图(EOG)等穿戴式的视线估计方法，虽然减小了头部运动的影响，但对受试者干扰较大，不适于较长时间的佩戴。伴随图像处理与计算机视觉技术的发展，使得基于视频的视线估计方法便捷、非穿戴式的优势逐渐显现出来，并在医疗诊断、辅助驾驶和人机交互等多个领域得到普及与应用。但对于视线估计研究而言，如何高效地融入头部运动数据仍然有待进一步的研究。

近些年，在深度神经网络的研究热潮下，基于头眼数据融合的视线估计算法研究有了新的进展，主要可分为对注视目标的估计、视线落点估计与视线方向估计。2016 年，Recasens[9]等人设计了一个由两支路组成的深度神经网络模型，分别用于提取人物图像的头部姿态与注视方向，通过目标显著性判断实现对注视目标

的估计；Kyle Krafka[10]等人以手机和平板电脑等设备为研究对象，设计了由四个支路构成的深度神经网络，分别输入左右眼图、人脸图像和人脸位置，实现了于二维平面的视线落点估计。此研究对左右眼图像支路的参数权值进行共享处理的思想，受到此后很多研究的借鉴；2019 年，Google[11]团队对上述模型做了进一步改进，将视线落点估计模型改为三支路并由四个眼角的坐标位置代替人脸图像与人脸位置，最终取得了不错的视线落点估计效果；对于视线方向的估计，通常由水平和竖直两个角度所构成方向向量表示，Zhang[12]于 2015 年的研究中，将输入图像的头部姿态数据与眼部特征进行拼接，使用类似于 LeNet[13]的浅层网络结构实现对视线方向的估计，此研究对头眼数据融合的方式对本文后续研究起到了很大启发。

2. 实验

1) 实验准备

本文的数据采集实验招募了 8 名男性研究生志愿者作为被试，年龄在 23-30 岁之间，身体健康，视力良好。在实验前，每位受试者已熟悉实验的具体内容和注意事项，且均为自愿参与实验。每位被试都签署了书面的承诺与知情书，保证本文实验的合法性。为排除外界干扰，在做好实验准备后，每次实验仅由一名被试单独完成。

实验设备包括一台 DELL 计算机、一部惯性传感器(MTI-G-700)、三部工业相机、三台高清显示屏(分辨率 2560×1440)和八个红外点光源。其中，惯性传感器佩戴至被试头部后方大约枕骨位置，用于测量被试在进行目标捕获时的头部姿态欧拉角(Pitch、Roll、Yaw)。由于在本文中 MTI-G-700 惯性传感器仅用于测量头部姿态的变化，下文将其简称为头部姿态仪。本研究选用了三部 Point Grey GS3-U3-41C6NIR-C 工业相机，其采集图像的分辨率为 2048×2048，色度为近红外光谱(NIR)。三部相机分别安装在三台高清显示器的上方，用于同时拍摄被试在实验过程中进行目标捕获的头部运动图像和眼部运动图像。

对于头眼协调运动的数据采集实验，本研究创新性地搭建了“三目八光源”非穿戴式的视线落点数据采集平台。此平台不仅扩大了被试的头部运动范围，还能有效减小光照条件变化的影响。其主要由三部挂载在三台显示器上的工业相机和八个均匀分布于显示器边界的近红外光源组成，用于记录被试在目标捕获时的头眼图像，平台部署示意图如图 1 所示。

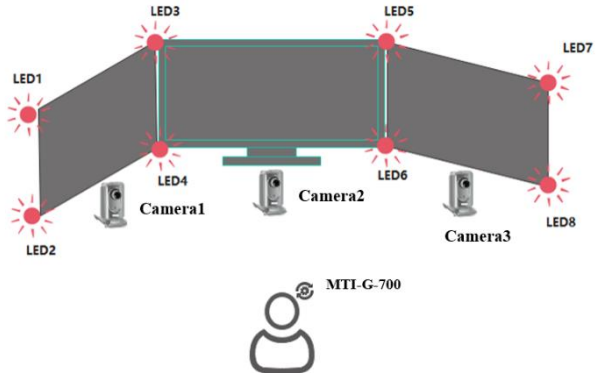

图 1 “三目八光源”平台部署示意图"

2) 实验设计

本研究设计的头眼运动数据采集实验主要是指：被试对三块屏幕上随机出现的目标通过头眼协调运动进行视觉交互的过程。开始实验前，要对红外光源、相机、头部姿态仪等设备进行校准，确保各设备出于正常运行状态。被试需佩戴好将头部姿态仪，调整坐位与中间屏幕的水平距离保存 60cm 左右，以确保自身出于三台相机的最佳焦距范围内。实验开始时，被试双眼需要正视中间屏幕的中心位置，按下记录键来标定头部姿态的初始欧拉角。标定成功后，待捕获目标以半径为 30 像素的红圈形式随机出现于三块屏幕上，被试通过头眼运动将视线落点瞄准在目标上，并按下记录键完成一次目标的捕获流程。实验过程中对被试无其他的要求，头部可以较大范围的自由移动。实验结束后，程序将记录每次被试在捕获目标时的头部姿态欧拉角、头部图像、眼部图像和目标中心点的坐标，并设置为一组数据。实验流程如图 2 所示。

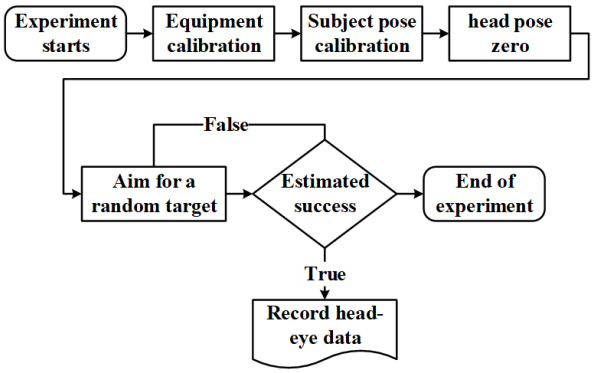

图 2 头眼运动数据采集流程图

考虑到实验操作较为单一，为确保被试的实验状态，保证实验数据质量，本研究设置单次实验时长为 20 分钟。实验时设备部署于六轴全动模拟飞行平台上，如图 3 所示。由于长时间使用相机性能会有效下降，存在多拍漏拍的情况，每次实验后需进行简单人工筛选。最终，通过对不合格样本的筛选后，本文共采集了 31507 组头眼运动数据。

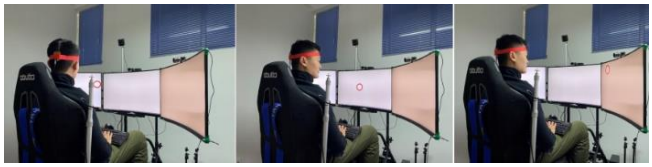

图 3 头眼运动数据采集实验过程

3. 方法

本研究采用基于外观的视线估计方法，通过非穿戴式多目相机所拍摄的头眼图像，实现对被试的视线落点估计。虽然基于外观的视线估计方法有较强的鲁棒性，但也存在着对被试头部自由运动限制较大、受光照条件变化影响较大等问题。而本文搭建的“三目八光源”平台通过三部显示器的捷联，不仅可以扩大被试的头部运动范围；并通过红外光源在人眼角膜上的反射光斑（普尔钦斑），为眼部图像增加了特征点，减小光照的影响。为此，本研究通过图像处理与特征提取的方法，融合头部特征与眼部特征，建立视线估计神经网络模型，进而实现视线落点估计的研究。

1) 头部特征提取

本文将姿态测量仪佩戴至被试的后脑勺位置，并将其作为头部运动的三轴基准点。头部姿态数据主要包括：俯仰角(Pitch)、偏航角(Yaw)、滚转角(Roll)，即抬头、摇头和转头，通过这三个欧拉角可以较为准确地估计出头部的空间姿态。本研究采用右手笛卡尔坐标系，其空间 X、Y、Z 三轴位置与对应的头部姿态欧拉角如图 4 所示。

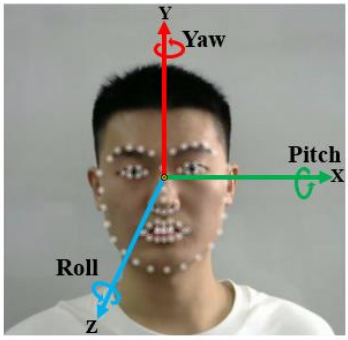

图 4 头部姿态欧拉角

为了便于对头部姿态数据的统计和分析，本研究对所记录数据进行了可视化处理，头部的三轴姿态数据如图 5 所示。通过三轴欧拉角偏转角度，可直观地观察出头部偏航角(Yaw)和俯仰角(Pitch)在目标捕获过程中的变化幅度较大，而滚动角(Roll)的变化幅度较小。

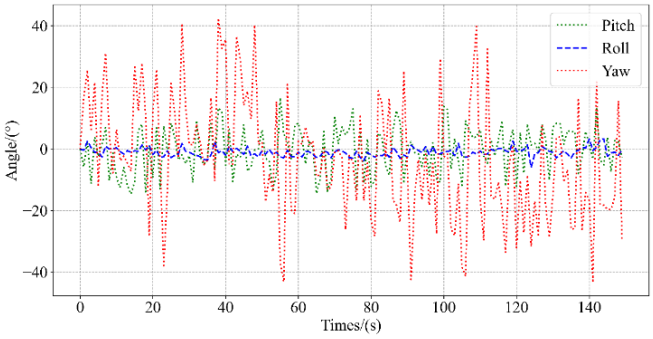

图 5 三轴姿态数据记录

2) 眼部特征提取

本研究使用的人眼检测方法主要是利用 AdaBoost 级联分类器结合 Haar-like 特征先检测出人脸区域，再在人脸区域中检测出人眼区域并进行截取。基于 OpenCV 所包含针对面部、眼部等进行过预训练的分类器，本研究对

目标出现屏幕对应相机所采集的正视图像进行了人脸检测、人眼检测，识别结果如图 6 所示。通过人眼检测结果，对被试的单眼区域按 64×64 的分辨率进行截取，得到左右眼部图像作为下一步卷积神经网络模型的输入。

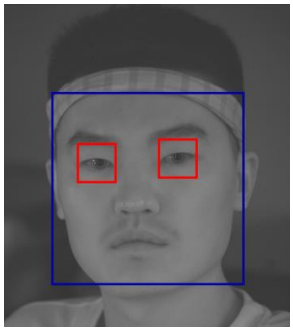

图 6 人脸检测、人眼检测结果

针对基于外观的视觉估计方法所存在对光照条件的要求，本研究结合基于特征的视觉估计方法的思想，在三个屏幕的边界等距地安放了八个近红外点光源，通过其在角膜的反射所形成的普尔钦斑作为眼部图片的特征点。由于相机属于近红外相机，因此红外点光源经眼角反射所形成普尔钦斑的亮度不受外界光照影响。综上述设想，本研究在对视线落点模型训练前，对所截取的左右眼图像进行阈值处理，得到含有较为明显普尔钦斑的左右眼图像并作为卷积神经网络模型输入的对照组，对比其对视线落点估计结果的影响。

本研究对三台显示器(分辨率：2560×1440)按左(0)、中(1)、右(2)的顺序编号，为了更加直观地对比几种阈值处理的效果与普尔钦斑的明显程度。本研究以左(0)屏的五个目标点位为例进行分析，具体位置如图 7 所示。其中，每个点前两个坐标为目标中心像素点于屏幕上的位置，第三个坐标为屏幕号。

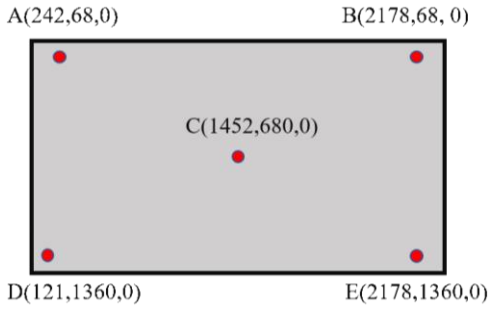

图 7 目标中心点位置

对上图 A、B、C、D、E 五点的所采集的眼部图像按照二值化阈值、截断阈值化和超阈值化零的顺序进行处理，对比几种阈值处理方法对普尔钦斑的检测效果，其结果如图 8 所示。由图中可观察出，对于不同的注视点，被试眼部所形成的普尔钦斑的数量与位置关系都有所不同，当注视左屏(0)屏幕中心位置时(C)，通过阈值处理可检测出 6-7 个普尔钦斑。其中，截断阈值化处理后的普尔钦斑较为明显，能有效地消除了其他外界光源的在角膜的反射光斑，保留了原始眼部图像。因此，本文将采用截断阈值化处理法对裁剪后的左右眼图像进行处理，检测含有明显普尔钦斑的左右眼图像作为下一步卷积神经网络模型的输入。

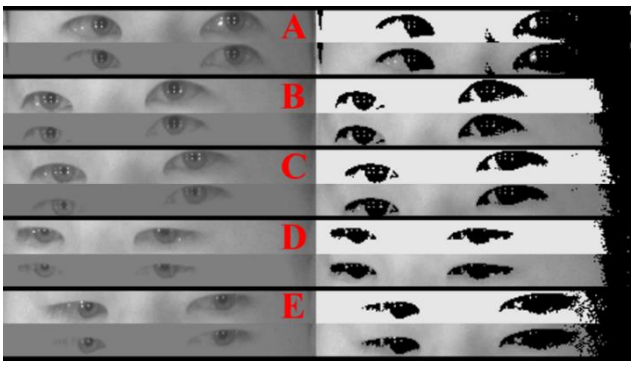

图 8 普尔钦斑检测结果

对于深度卷积神经网络模型，本文参考了 2015 年由何凯明等研究者提出的深度残差网络结构。目前，常用的几种 ResNet 网络主要包括：ResNet-18、ResNet-34、ResNet-50 以及其他变种，虽然增加网络的深度可以提升模型的准确率，但较为浅层的残差网络(ResNet-18)在实际应用中同样有良好的准确性，同时其模型较小，提供了更快的收敛速度，便于参数的优化。而且基于 ResNet 模型短接的操作，可实现对不同分辨率特征的组合，对于本文输入的眼部图像有较好的特征提取效果。因此，本研究以 ResNet-18 作为人眼视线落点估计模型，ResNet-18 网络结构如表 1 所示。通过对比不同类型的双眼图像输入（有无通过阈值处理检测普尔钦斑），分析其输出位于屏幕上视线落点的精度。由于传统的残差神经网络多用于分类任务，本文受 2019 年 Google 团队对视线落

点估计的研究启发，在神经网络的隐含层后连接多个全连接层，用于回归出视线落点坐标。

本文的人眼视线落点估计模型，输入为被试正视屏幕中带捕获目标时的左、右眼部灰度图像。由于眼部图像的分辨率要求较低，本文采用 64×64 的分辨率大小截取单眼图像。相对于传统 ResNet-18 网络 RGB 图片的输入大小(224×224×3)，本文的图像输入较小(64×64×1)，使得模型的计算速度提高。而本模型分为两条结构相同的支路，其输入层为裁剪后的左、右眼图图像，而每条支路隐含层的主要结构是按着 ResNet-18 网络结构搭建的，主要由 17 层卷积层、8 个残差块构成和 2 个池化层构成，使用 Relu 作为所有卷积层的激活函数，实现对左右眼图像的特征提取，最后通过 4 个全连接模块将左右眼的特征图进行融合并输出所估计的视线落点坐标 G(P_x, P_y, S_n)，本文的人眼视线落点估计网络模型结构如图 9 所示。

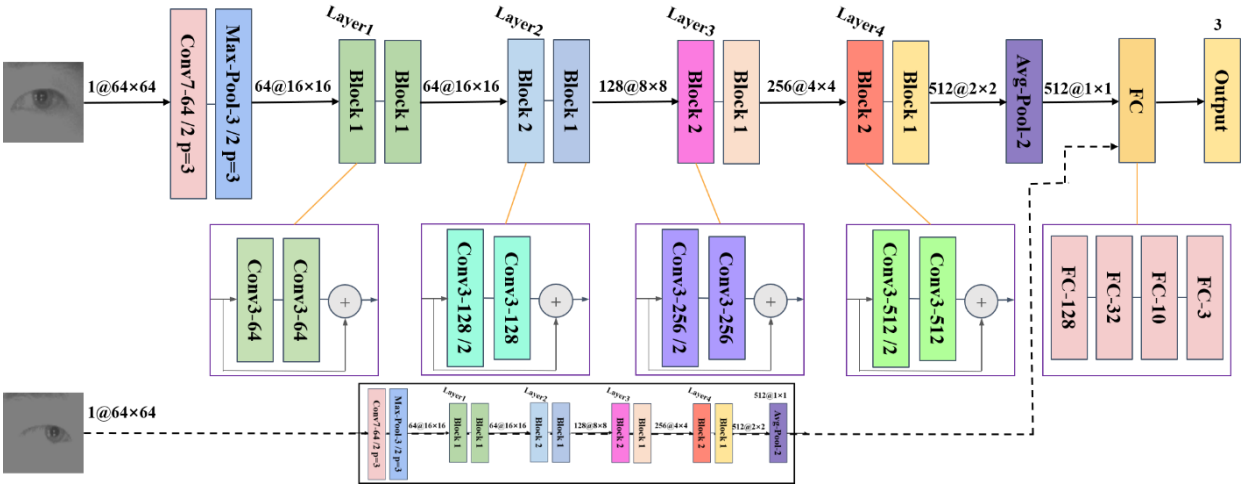

图 9 人眼视线落点估计网络结构

3) 融合头眼的视线落点估计

为了在视线落点估计中融合头部姿态数据，本研究借鉴采用传统的 BP 神经网络的思想，通过多个全连接层作为头部姿态数据特征提取的网络分支，主要包括：一个输入层和一个隐含层构成。其中，输入层为被试的头部姿态欧拉角(Roll, Pitch, Yaw)；隐含层由三个全连接层组成，神经元节点个数分别为 100，16，16，并使用 Relu 作为激活函数，将最后一层全连接层所提取的特征向量作为输出。

基于特征层融合的原理，本研究先对头眼数据进行预处理以完成特征提取，对于眼部图像由上文中所提及的眼部视线落点估计模型的输入层和隐含层实现左右眼部图像特征的提取；对于头部图像则是由头部姿态欧拉角，通过头部特征提取网络输出与眼部特征维度相同的特征向量，完成维度配准。两部分特征通过多个全连接层实现特征融合，构成融合头眼运动的视线落点估计模型结构。本研究的视线落点估计模型的网络结构共包括三个支路，输入层分别输入被试在捕获目标时的左右眼图像与头部姿态欧拉角。对于左右眼支路，使用三个全连接层进行特征提取，神经元节点个数分别为 128、32、16；对于头部姿态支路，采用头部特征提取网络，最后通过由两个神经元节点个数为 16 和 3 的全连接模块完成对三条支路数据的特征融合，实现对三屏实现落点的回归，如图 10 所示。

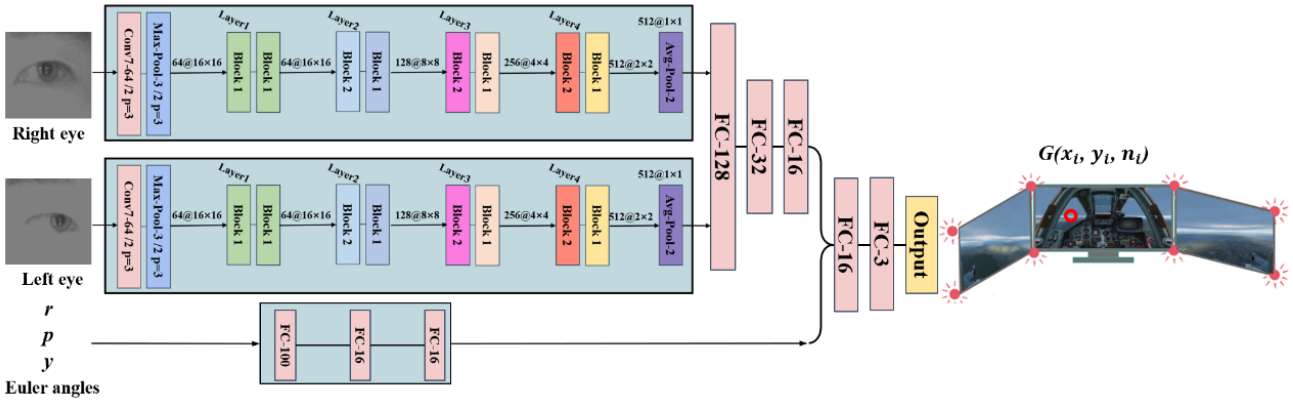

图 10 融合头眼运动的视线落点估计模型

4. 结果

为了使人眼视线落点估计模型在学习中朝向更精确的估计性能改进，本文使用均方方差(MSE)作为损失函数，MSE 表示预测数据与原始数据所对应点误差的平方和均值：

$$MSE = \frac{1}{n}\sum_{i=1}^{m}\omega_i(y_i - \hat{y}_i)^2 \qquad (1)$$

其中，n是样本的个数，y_i是原始数据，$\hat{y}_i$是预测数据。当 MSE 值越接近 0 时，说明模型的拟合能力越强，其视线落点估计也越准确。基于本文的三屏实验平台，每个屏幕为一个二维平面，采用欧式距离计算标定点与估计点间的差值，预测接受域为以标定点为圆心，半径 30 像素的圆形区域。在模型的训练过程中，本文使用自适应矩估计(Adam)作为优化器，其能对不同的参数调整不同的学习率；网络的学习率设定为10^{-3}，批处理量设置为 32，训练周期为 100。

本研究通过对原始数据的筛选，去除被试闭眼的图像后，从中共选取了 30000 张图像并将它们裁剪为适合模型输入的尺寸，其中 70%用于模型训练，30%用于模型的性能测试。同时，以输入原始数据前是否进行普尔钦斑检测的两种情况，构成两种不同的视线落点估计模型并作为对照实验，对比分析最终的预测准确率和其他性能。

实验结果表明，当仅使用原始眼部图像作为模型输入时，模型(Eye)所需提取的特征较少，收敛速度较快。在 200 个 epoch 左右模型基本收敛，其平均准确度可达到 85.6%；而当输入为普尔钦斑检测后的眼部图像时，虽然模型(Eye&Purkinje)的收敛速度较慢，在 400 个 epoch 左右基本收敛，但是其平均准确率可达到 87.7%。通过对比两种模型的性能，Eye&Purkinje 模型在输入前通过阈值处理使普尔钦斑特征更加明显，增加了隐含层特征提取的复杂度，从而增加了模型的收敛时间，但其相对于 Eye 模型的视线落点估计平均准确率提升了 2.1%，且损失曲线较为平稳，模型的稳定性更好。因此，本研究的选用 Eye&Purkinje 模型作为人眼视线落点估计的模型，并验证了输入含有显著普尔钦斑的眼部图像，可为图像增加特征点，减小光照条件的影响，提高模型估计的准确率。两种模型训练的准确率和损失变化曲线，如图 11 所示。

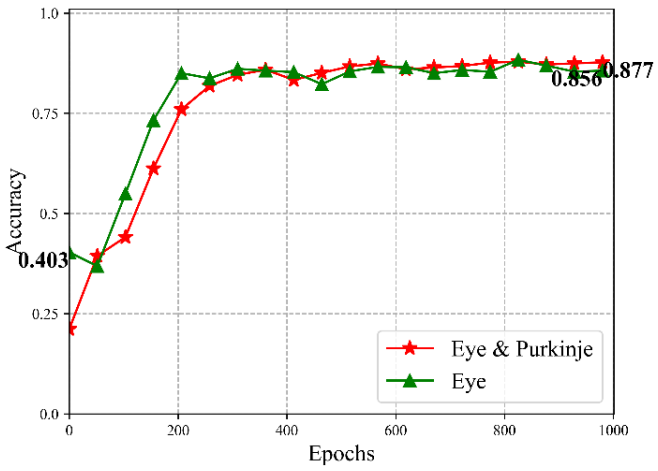

图 11 Eye 和 Eye&Purkinje 模型性能对比分析图

为了评价融合头眼运动的视线落点估计模型的性能，本文将其与仅使用眼部图像的两种模型进行了对比分析。如图 12 所示，在前 200 个 epoch 中，加入头部姿态后的模型性能已优于 Eye&Purkinje 模型，但由于需要融合头眼特征，参数优化所需时间较长，准确率不如 Eye 模型。从模型的收敛速度分析，虽然 Eye&Purkinje&Head 模型收敛较慢，但在 600 个 epoch 后模型趋于平稳，模型的精度较高，其平均准确率可达 89.9%，相对于仅使用单维度的眼部图像输入模型，该模型通过融合头眼多维度的数据，实现对头眼协调运动数据的压缩与关联，其准确率最多提升了 4.3%。

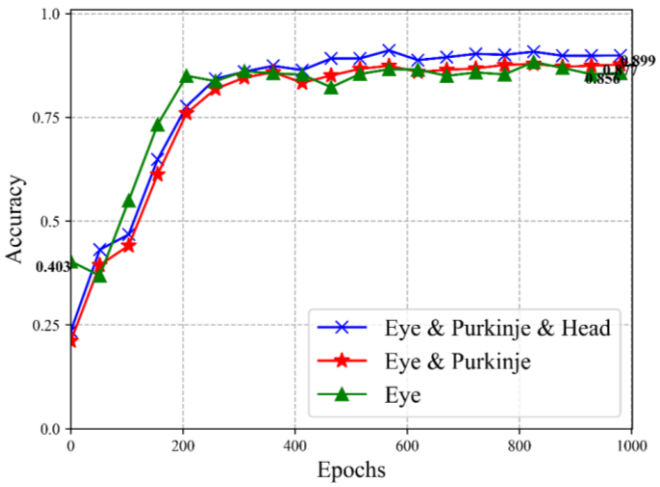

图 12 三种视线落点估计模型性能对比分析图

5. 总结

对于视线落点估计的研究，本文分别采用了三种模型进行了对比分析，即仅使用眼部图像的视线落点估计模型、使用眼部特征的视线落点估计模型与融合头眼运动特征的视线落点估计模型。通过对比，融合头部运动与眼部特征点图像的视线落点估计模型(EPH)对测试集的预测结果有较高的测试准确度，所估计的视线落地基本在待捕获目标的预测接受域内且未出现过拟合的现象，其对视线落点估计的平均准确率可达 89.9%。但本文也发现这三个模型普遍估计的精度不高，该问题是基于外观的视线落点估计方法存在的普遍问题，有待后续进一步的研究与展望。

综上，本文基于一种外观与特征相结合的视线估计方法，有效地融合了视线移动时的头部运动与眼部运动，将头部姿态数据与眼部特征通过深度卷积神经网络结构最终实现了对二维屏幕中视线落点的精确估计。为视线落点估计研究提出了一种切实有效的研究方法。

参考文献

[14] Atchison D A, Smith G, Smith G. Optics of the human eye[M]. Oxford: Butterworth-Heinemann, 2000.

[15] Wang Changyuan, Li Jingjing, Jia Hongbo, et al. Research methods and progress of head-eye movement[J]. Journal of Xi'an University of Technology, 2012, 32(3): 173-182.

[16] Mao Xiaobo. Research on Modeling and Control of Bionic Robot Eye Movement System[D]. Zhengzhou: Zhengzhou University, 2011.

[17] Lei Zhihui, Yu Qifeng. A new method to determine eye movement translation[J]. Experimental Mechanics, 2003, 18(4): 564-568.

[18] Freedman E G. Coordination of the eyes and head during visual orienting[J]. Experimental brain research, 2008, 190(4): 369-387.

[19] Mao Xiaobo, Chen Tiejun. A bionic model of head-eye coordination motion control[J]. Journal of Biomedical Engineering, 2011, 28(5): 895-900.

[20] Liu Jiahui, Chi Jiannan, Yin Yixin. Review of feature-based gaze tracking methods [J]. Journal of Automation, 2021, 47(2): 252-277.

[21] Zhang C, Chi J N, Zhang Z H, et al. Gaze estimation in a gaze tracking system[J]. Science China Information Sciences, 2011, 54(11): 2295-2306.

[22] Recasens ,A R C. Where are they looking?[D]. Massachusetts Institute of Technology, 2016.

[23] Krafka K, Khosla A, Kellnhofer P, et al. Eye tracking for everyone[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2176-2184.

[24] He J, Pham K, Valliappan N, et al. On-device few-shot personalization for real-time gaze estimation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019: 0-0.

[25] Zhang X, Sugano Y, Fritz M, et al. Appearance-based gaze estimation in the wild[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4511-4520.

[26] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.