Mathematical Biosciences and Engineering

MBE, 20 (8): 13974–13988.
DOI: 10.3934/mbe.2023622
Received: 09 April 2023
Revised: 24 May 2023
Accepted: 08 June 2023
Published: 21 June 2023

http://www.aimspress.com/journal/MBE

Research article

An integrated neural network model for eye-tracking during human-computer interaction

Li Wang[1], Changyuan Wang[2,*], Yu Zhang[1] and Lina Gao[1]

[1] School of Optoelectronic Engineering, Xi'an Technological University, Xi'an 710000, China
[2] School of Computer Science, Xi'an Technological University, Xi'an 710000, China

*** Correspondence:** Email: cyw901@163.com.

Abstract: Improving the efficiency of human-computer interaction is one of the critical goals of intelligent aircraft cockpit research. The gaze interaction control method can vastly reduce the manual operation of operators and improve the intellectual level of human-computer interaction. Eye-tracking is the basis of sight interaction, so the performance of eye-tracking will directly affect the outcome of gaze interaction. This paper presents an eye-tracking method suitable for human-computer interaction in an aircraft cockpit, which can now estimate the gaze position of operators on multiple screens based on face images. We use a multi-camera system to capture facial images, so that operators are not limited by the angle of head rotation. To improve the accuracy of gaze estimation, we have constructed a hybrid network. One branch uses the transformer framework to extract the global features of the face images; the other branch uses a convolutional neural network structure to extract the local features of the face images. Finally, the extracted features of the two branches are fused for eye-tracking. The experimental results show that the proposed method not only solves the problem of limited head movement for operators but also improves the accuracy of gaze estimation. In addition, our method has a capture rate of more than 80% for targets of different sizes, which is better than the other compared models.

Keywords: human-computer interaction; eye-tracking; gaze estimation; vision transformer; feature pyramid network

1. Introduction

Human-computer interaction is the way that people exchange information with a system. The system can be a wide variety of machines, computer systems, and software [1]. Early human-computer interaction was mediated by machine language, and the interaction was accomplished through manually inputting machine language instructions to exchange information. With the development of computer and communication technology, there are more and more ways of human-computer interaction, including speech recognition, gesture recognition, and eye-tracking [2–4]. Human-computer interaction methods based on eye-tracking are wildly used in various fields because of the characteristics of real-time performance and flexibility [5–9]. Eye-tracking is the method that estimates the gaze point or direction of the eye by tracking the movement of the eye. In human-computer interaction, gaze control is a flexible method to enable communication with computers [10].

Eye-tracking has always been a hot topic in machine vision technology [11]. Gaze-tracking methods fall into two main categories: model-based methods and appearance-based methods [12–14]. Model-based methods generally use special equipment to collect images, detect eye features by image analysis, and then use these features to build models to estimate gaze. In the model-based approach, the popular sight features include the pupil, iris, canthus and corneal reflection points. The specific applications of these features include using the radius and center of the pupil to estimate gaze through a geometric model [15,16] and using corneal reflection points to estimate the gaze [17,18]. The pupil-canthus method is used to estimate the fixation point of users [19,20]. The model-based methods must ensure the quality of the acquired image to obtain an accurate gaze estimation. The accuracy of gaze estimation will be affected by image resolution, noise and illumination conditions. Therefore, to get an accurate and reliable gaze estimation model, the hardware must be equipped with high-quality cameras and special devices such as narrow-angle lenses and external lighting to extract adequately accurate and detailed edges or feature points. But in the wild, because of the influence of the head pose or light conditions, the method based on the model yields a high error rate [21]. In addition, it is necessary to analyze the prior knowledge of the eye model to establish a good line-of-sight estimation model. However, this method of establishing a good model based on prior knowledge is a challenging task [22]. In contrast to the model-based approach, the appearance-based approach directly estimates gaze by analyzing eye images. The specific process is as follows. First, collect the face or eye image of the tested person with the label. Then select the training sample as the input image data, fit the relationship between the human eye appearance and the fixation direction or fixation point through the training sample and finally input the test image sample to determine the gaze direction or fixation point of the corresponding area. This method uses a mass of statistical data to learn the invariance of appearance differences [23]. And it does not require the manual design of features, as it automatically extracts image features from the data, so it has good robustness.

Deep learning has aroused increasing research interest in recent years [24,25]. With the continuous development of deep learning theory, the gaze estimation methods based on appearance have been increasingly widely used [26–29]. In numerous approaches based on appearance, deep learning networks, especially convolutional neural networks (CNNS), exhibit good performance, to a certain extent, improving the accuracy of the gaze to estimate. In most of these studies, they used a front-facing camera to take an image of a human eye or face. To get a complete picture of a face or human eye, one must limit the movement of the head and narrow the field of vision. However, this approach is inapplicable to aircraft cockpit scenarios with multiple screens. This is because, during a

flight, the objects which need to be viewed are not concentrated on one screen but spread out across multiple screens. Therefore, to ensure that the flying personnel are not subject to the rotation angle of the head during gaze interaction, this paper proposes an eye-tracking method that uses multiple cameras to record images. This method can ensure that the complete frontal face image can be collected when the pilot turns their head to look at any target on the screen. Then, a CNN and transformer hybrid network model are applied to detect the fixation position of flight personnel in the process of human-computer interaction, using the frontal face image corresponding to each screen as input.

2. Related work

With the rapid progress of artificial intelligence technology, the traditional human-computer interaction cannot adapt to the multimodal human-machine intelligent environment for the efficient transmission of information. Therefore, it is of great significance to study how to actualize intelligent human-computer interaction. Eye-tracking provides a feasible solution for intelligent interaction. For example, Zhang et al. [30] proposed a multi-device gaze estimation algorithm based on a CNN for specific users. In this algorithm, cameras are installed on five devices, such as mobile phones, tablet computers, and smart TVs to collect the face image dataset of user interaction with the device. When training the CNN, it uses the encoder of a specific device and the shared feature extraction layer to process the image and gives the gaze estimation of the decoder of each device. Li et al. [31] designed an eye-tracking method for gaze control for surgical robots. In this approach, the direction of movement of the surgical robot or area is decided by the by the user's point of gaze. They used images collected by a single camera training a CNN to get the user's gaze position. Finally, the user can control the surgical robot to move in nine directions according to the eye gaze information. Lorenz and Thomas [32] developed an eye-tracking system for detecting human interaction intentions. It uses two continuous cascaded convolutional networks to extract face features and estimate the head pose to determine eye fixation direction. Robots can judge human intentions based on line of sight. Kim et al. [33] developed an interactive system that can control devices through the user's gaze and simple gestures. The system's gaze estimation module uses a video stream recorded by a camera. It detects the user's facial image in the video to get information feature vectors, including the head pose. Then, these feature vectors are fed into the CNN to train the user's gaze estimation model. Luo et al. [34] developed a human-computer interaction control system for wheelchairs using eye-movement tracking and blink detection. It first extracts the pupil feature of the eye through binarization of the human eye image and then obtains the movement trajectory of the eye. Then, the eye movement tracker locates the eye's gaze direction. At the same time, the convolutional neural grid detects the open and closed states of human eyes to judge whether the user blinks. Finally, the system operates according to the user's gaze direction or blink movement to control the operation of the electric wheelchair.

All of the above gaze-based interaction methods show good performance for specific applications. However, the deflection of the user's head will result in a tendency to decrease the accuracy of gaze estimation because the image collected by only one camera cannot contain full-face or complete eye information. Therefore, we have constructed a multi-camera system to study the method of gaze interaction without restricting head movement. In this paper, we mainly design a hybrid network of a CNN and a transformer for gaze estimation to improve the reliability of gaze interaction, aiming at the problem of eye-tracking in the process of visual target acquisition by flight personnel in the cockpit scene should not be more than 4 levels. The fond of heading and

subheadings should be 12-point normal Times New Roman. The first letter of headings and subheadings should be capitalized.

3. The proposed method

3.1. Introduction of the method of this paper

The steps of this approach are shown in Figure 1. As the subjects look at the target on one of the three screens, the three cameras will get the initial set of images. Each group of images contains one front and two side images of subjects. First, select the image taken by the frontal camera from the initial set of images, and then detect the facial landmarks of the subjects by using the facial feature point localization network. Each group of images contains one front and two side images of subjects. According to the facial landmarks, set the face ROI and obtain the frontal face image. Then, the human face image is input into the hybrid network of the CNN and transformer built to track the number of screens watched by the pilot in the cockpit and the pilot's fixation position.

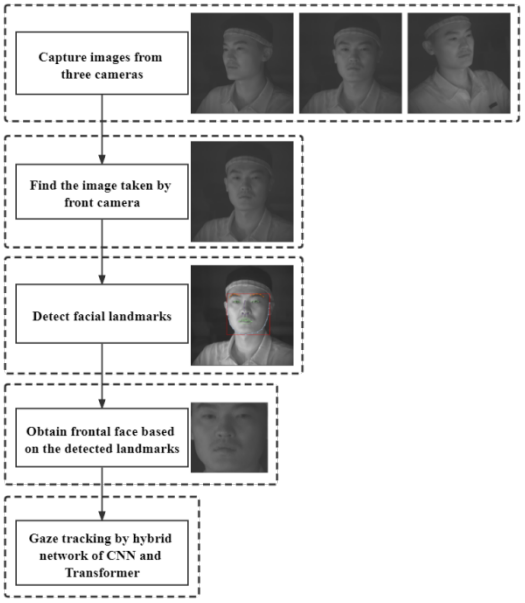


Figure 1. Steps of this method.

3.2. Image processing

Pick the image of the frontal camera from the images of the three cameras through the preprocessor, using a facial feature point localization network [35] to extract the face image. This facial feature point localization network, based on the hourglass network [36] architecture used for human posture estimation, replaces the original bottleneck block of the hourglass network with layered, parallel, and multi-scale blocks [37], and then it carries out landmark localization of the face. Obtain the corresponding ROI by using face contour facial landmarks and cutting out the face image. The size of all cropped face images is 224×224. Figure 1 already shows an example of the result of processing a set of images.

3.3. Eye-tracking based on a deep learning network

3.3.1. A network model for eye tracking

To realize the eye-tracking of the aircraft cockpit scene, we design a eye-tracking model based on a deep learning network. The model comprises a vision transformer (VIT), a feature pyramid network (FPN) and fully connected layers. Figure 2 shows the model framework.

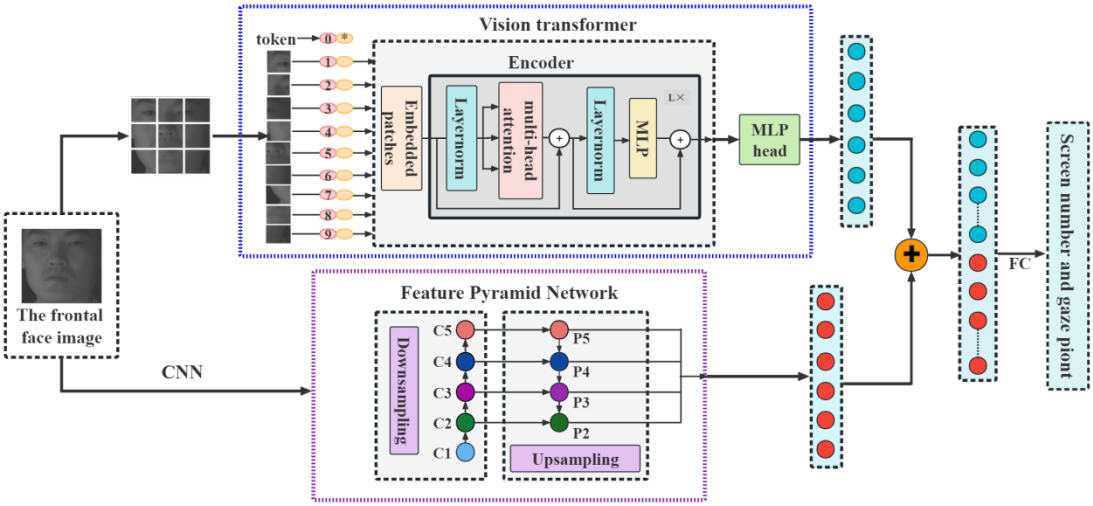


Figure 2. The overall framework of the eye-tracking model.

The cropped frontal face images were fed into the VIT and FPN, respectively, for feature extraction, and then the extracted features were fused through the fully connected layers. Finally, the screen number and the position of the fixation point on the screen were output. VIT extracts the global feature of the face image, and FPN extracts the local characteristics of the face image. Fusing global features and local features of faces can effectively improve the accuracy of gaze estimation.

The loss function of the model adopts the Minkowski distance, which is defined as

$$L_p(x_i, x_j) = (\sum_{m=1}^{n} |x_i^{(m)} - x_j^{(m)}|^p)^{\frac{1}{p}} \quad (1)$$

where $x_i, x_j \in X = R_n, x_i = (x_i(1), x_i(2), \ldots, x_i(n))^T, x_j = (x_j(1), x_j(2), \ldots, x_j(n))^T$. p is a variable parameter. The formula shows that the distance metric of Minkowski distance has tremendous flexibility. It can iterate over P to find the most suitable distance metric for practical applications. After several experimental trials, the value of P in this paper was calculated to be 4.

3.3.2. Vision transformer

A transformer is a new network model that uses the self-attention mechanism to extract intrinsic features [38]. Because the transformer advanced performance in natural language processing, Dosovitskiy et al. [39] attempted to use a standard transformer for image classification and called the network a vision transformer. VIT introduces the concept of an image patch to transform the image into sequence data that the transformer structure can process. Since the input to the standard transformer must be a one-dimensional token embedding sequence, VIT first segments the image

into fixed-size patches and generates a linear embedding sequence of these patches. Then the sequence can be used as the input to the transformer. This process is as follows.

Assum image $F \in R^{H\times W\times C}$ such that (*H*, *W*) is the resolution of an image and C is the number of channels. F is divided into *N* flattened 2D patches $X_p \in R^{N\times(p^2\cdot C)}$, where $N = HW/p^2$. We map each patch into a *D*-dimensional embedding vector via a learnable projection matrix E and add X_{token} before the *D*-dimensional embedding vector. X_{token} is also a *D*-dimensional learnable embedding vector, which can better represent global information. After that, add the location code E_{pos} which indicates the location information of the patch. We get the following patch embeddings

$$z_0 = [X_{token}; X_p^1 E; X_p^2 E; \ldots; X_P^N E] + E_{pos}, E \in R^{(p^2\cdot C)\times D}, E_{pos} \in R^{(N+1)\times D} \quad (2)$$

The patch embeddings are input to the encoder of VIT and are processed sequentially by LayerNorm (LN), multihead attention mechanism (MSA) and multilayer perceptron (MLP). The processing equations are (3) to (4).

$$Z_l' = MSA(LN(z_{l-1})) + z_{l-1} \quad l = 1, \ldots, L \quad (3)$$

$$Z_l = MLP(LN(Z_l')) + Z_l' \quad l = 1, \ldots, L \quad (4)$$

Apply Layernorm before the multi-headed attention mechanism module and the multi-layer perceptron module and apply residual connectivity after the multi-headed attention mechanism module and the multi-layer perceptron module.

3.3.3. Feature pyramid network

The FPN is a CNN for detecting multi-scale targets [40]. The FPN combines the fine-grained spatial information of shallow feature maps with the semantic information of deep feature maps. It dramatically improves the performance of target detection. The core structure of FPNs contains bottom-up pathways and top-down pathways.

The bottom-up pathway is the forward process of the CNN. In the forward process, the size of the feature map changes after passing through some layers, while it does not change when passing through some other layers. The layers that do not modify the size of the feature map are grouped into one stage so that each extracted feature is the output of the last layer of each step, thus forming a feature pyramid. Specifically, it serves to output the features of the last residual structure in the five stages of the residual neural network. Then, the feature map is up-sampled by a top-down pathway so that the up-sampled feature map has the same size as the feature map of the next layer. The feature maps generated by the bottom-up way are C1, C2, C3, C4, and C5 in Figure 2. The feature maps generated by the top-down path are P2, P3, P4, and P5 in Figure 3.

4. Experimental design and results

4.1. Data acquisition equipment

Figure 3 shows the experimental environment of the flight simulation platform. The flight simulation platform comprises a six-axis full-motion platform, three displays, flight joysticks, data measurement instruments, and a mainframe. This study builds a system for capturing targets with

gaze during human-computer interaction based on a simulated flight platform. Figure 4 shows the structure of the system.

Figure 3. Simulation flight platform.

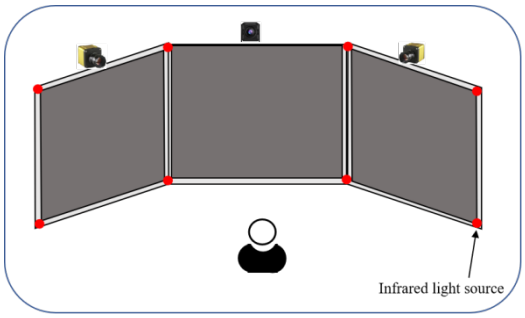


Figure 4. The structure of the experimental system.

The system comprises a head motion sensor, three industrial digital cameras, eight infrared light sources, and visual target calibration software. The head motion sensor measures the subject's head posture data, and three industrial black-and-white digital cameras acquire frontal and side images of the subjects. Infrared light sources ensure that the captured images are not affected by external ambient lighting. The function of the visual target calibration software is to record where the target appears during the simulated flight.

4.2. Experimental process

To collect head and eye movement data during human-computer interaction, 12 graduate students with normal vision, aged 21–25 years, were recruited as subjects. All subjects had no neurological or psychiatric disorders history and signed an informed consent form before the experiment. In addition, this study has passed the review of the ethics committee of the unit.

The equipment needed calibration before the experiment. Each subject completed 10 sets of experiments, each lasting 30 minutes. Figure 5 shows the experimental process. First, the subjects adjust their sitting posture and wear the head motion sensor. Then, the user opens the visual target calibration software, and a red circle will randomly pop up on the display screen of the simulated flight platform every 10 seconds. The red circle is the target that the subject needed to capture. When it appears, the subject looks at the center of the red circle and presses the space key to indicate that

the subject has obtained the target. At this time, three cameras will take an image of the subject while capturing the objective. The head motion sensor also saved the subject's head posture data. The visual target calibration software recorded the coordinates of the center point of the red circle target. Throughout the experiment, the subject's head was able to rotate and capture targets anywhere on the three displays.

Figure 5. Experimental process diagram.

4.3. Analysis of results

The face images captured by the pre-processed frontal camera were input into the proposed eye-tracking model. The features of face images were extracted using a VIT and FPN, respectively, then, these features were fused through a fully connected layer. The final output was the screen number and the coordinates of the gazing point on the screen. Figure 6 shows the structure of the model.

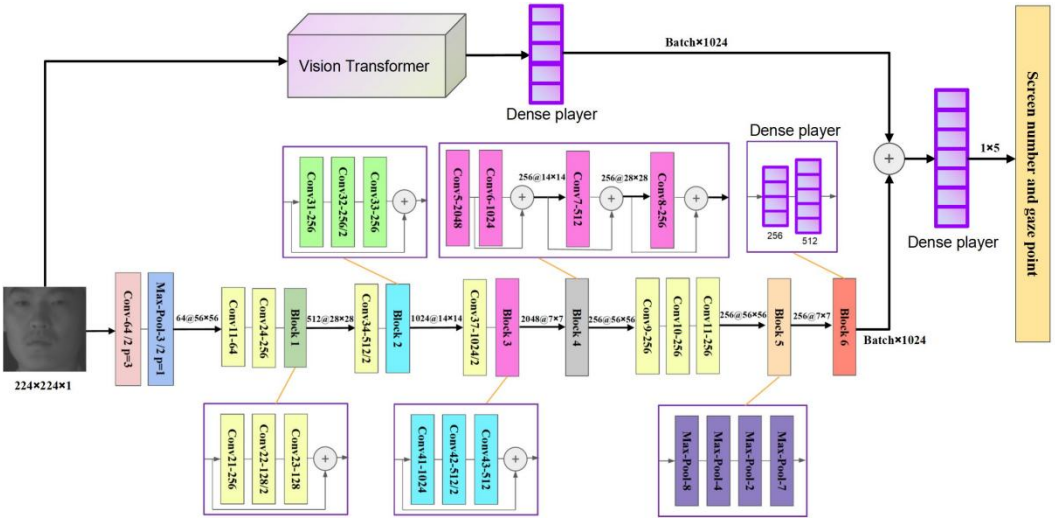


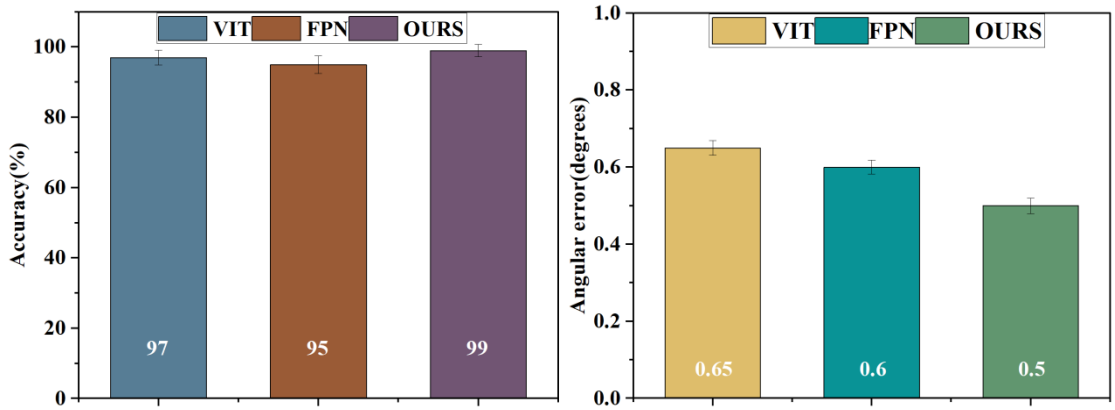
Figure 6. Structure of the model.

We used classification accuracy as a rubric for screen number prediction. We used the angular error between the true and the predicted gaze positions as the evaluation indicator for gaze estimation. We randomly selected 5000 groups of images from the collected dataset for analysis. The epochs for each experiment were 500. The learning rate for the first 270 epochs was 10^{-3}, while the learning rate for the last 230 epochs was 10^{-4}. The batch size for each training set was 16. We used simple cross-validation and 10-fold cross-validation to group the sample data for training when dividing the

training and test sets, respectively. In the simple cross-validation method, the first 80% of the data set was used as the training set, and the remaining data as the test set. The 10-fold cross-validation divided the dataset into 10 parts, with nine parts used as the training set and one as the testing set. We counted the test results of both methods, as shown in Table 1. Table 1 shows that the 10-fold cross-validation can improve the gaze estimation accuracy and outperforms the simple cross-validation method without considering the time consumed by the model training. Therefore, we chose 10-fold cross-validation for grouping the dataset in this paper.

Table 1. Comparison results for the two methods.

Method	Classification of screen number (%)	Angle error (°)	Average training time per session (s)
cross validation	0.985	0.5	33.4
10-fold cross-validation	0.997	0.4	49.8

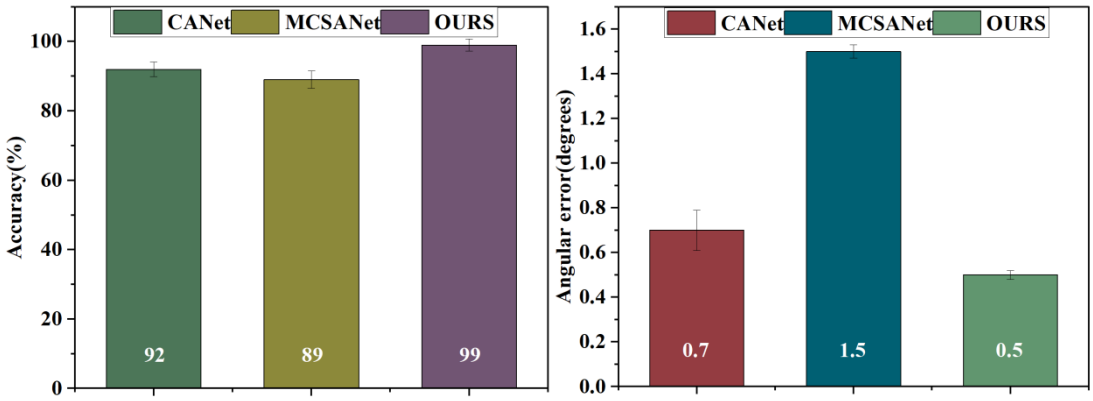


(a) Classification of screen number (b) Angular error

Figure 7. Comparison of gaze estimation results.

A single transformer network and a single FPN network were used as control groups for comparison with the proposed hybrid transformer and FPN parallel networks. The comparison results for the screen number prediction accuracy are shown in Figure 7(a). Figure 7(b) shows the comparison result for the gaze's angular error. In Figure 7(a), the classification accuracy of our proposed hybrid network is higher than that of other single networks. In Figure 7(b), the angular error of the proposed hybrid network is smaller than that of the single network. Therefore, Figure 7 shows that our constructed transformer and FPN hybrid parallel network outperform the single network.

To further compare the performance of the gaze estimation model in this paper, the CANet model [41] and the MCSANet model [42] were also used on the dataset of this paper. The prediction accuracy of the screen number and the error of gaze estimation for these three models are shown in Figure 8. The results in Figure 8 show that compared with the CANet and the MCSANet, the accuracy of the screen number prediction obtained by this method was the highest, and the angular error of gaze estimation was the lowest, which can better confirm the point of view during human-computer interaction.



(a) Classification of screen number (b) Angular error

Figure 8. Performance comparisons of gaze estimation networks.

Figure 9. Interaction application scenarios.

The purpose of eye-tracking in this study was to evaluate the effectiveness of subjects' target capture by the proposed model for a flight cockpit scenario with multiple screens. The evaluation metric for target capture is the percentage of red-circled targets captured. Since the target is a circle, we specify that the subject captures the target if the error value of the gaze estimation is less than the radius of the circle target. Otherwise, it means that the subject did not acquire the target. Figure 9 shows the gaze interaction application scenario. The background of figure 9 is the cockpit of an aircraft in a flight simulation game. The display in the picture is the virtual integrated control panel (ICP). In the figure are red circles of different sizes of targets. Each red circle represents a button in the ICP. The user selected three of the red circle targets. These selected targets were numbered I, II, and III. The radius of Target I was 20 pixels. It was set to represent the button for the mode selection function. The radius of Target II was 40 pixels. It was selected to represent the button that implements the communication control function. The radius of Target III was 60 pixels. It was set to represent the button that completes the message input. Target I, Target II, and Target III were used as the objects to be captured by the subjects. The results of using different models for target capture are shown in Figure 10.

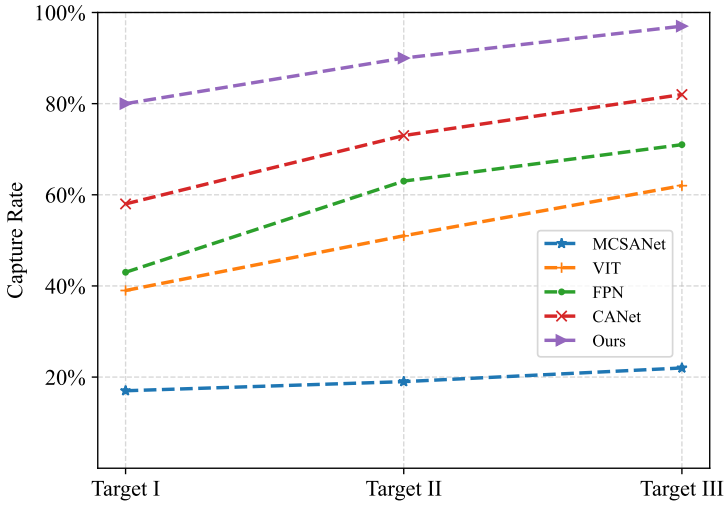


Figure 10. Comparison of target capturing results.

Figure 10 shows the capture rates of the five models for three targets. The success rates of all five models on target tracking tended to increase with increasing target size. Compared to other models, the eye-tracking model constructed in this paper can capture all types of targets effectively, with capture rates above 80%. For Target I,I.e. the smallest size, the capture rate of MCSANet was only 17%. Moreover, only the CANet model and our model had a capture rate of more than 50%. In particular, our model had the highest success rate of 80% for capturing Target I. These results indicate that the model proposed in this paper has low eye-tracking errors and can obtain good results when capturing small targets. For Target III with the largest size,the FPN, CANet, and our method each had a success rate of over 60%. However, only the eye-tracking model we built had the Target III capture rate exceeding 90%, showing the optimal performance. In conclusion, the comparison results show that the eye-tracking model established in this paper is more stable for the acquisition of different targets, which is better than other models.

5. Discussion and conclusion

Aiming at the target capture function in the human-computer interaction process in the flight cockpit scene, this paper presented a hybrid network combining a CNN and transformer for eye tracking. To improve the gaze estimation accuracy, cameras were installed on three display screens in the simulated flight cockpit to capture images containing the subjects' faces. First, all images captured by the frontal camera were selected and cropped to obtain the subject's face image. The advantage of using three cameras is that it removes the limitation of the subject's head rotation angle and expands the subject's field of view.

Then, inspired by previous studies using frontal face images for gaze estimation, we input the cropped frontal face images into the proposed eye-tracking model to predict the gaze position of the subjects. To test the model performance presented in this paper, we compared it with various models. We concluded that the transformer and FPN hybrid parallel network could improve gaze estimation accuracy.

Finally, we applied both the present model and other models to the target capture task in a

simulated flight cockpit scenario and found that the performance of our model is superior.

The experiments and models designed in this study achieved excellent results on the target capture task for human-computer interaction and achieved the desired goals. However, there are still some problems in the experiment, such as insufficient population distribution of subjects and insufficient ability of real-time target acquisition. Subsequent research will focus on the two main requirements of the extensiveness of the tested population and the real-time nature of target capture. At the same time, we will continue to optimize the neural network model and reduce its complexity.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

We thank the anonymous reviewers for their constructive comments and gratefully acknowledge the support of the National Natural Science Foundation of China (No. 52072293) and the National Defense Science and Technology Innovation Zone (No. 2020-JCJQ-JJ-430).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Y. Shi, Z. Zhang, K. Huang, W. Ma, S. Tu, Human-computer interaction based on face feature localization, *J. Visual Commun. Image Represent.*, **70** (2020), 102740. https://doi.org/10.1016/j.jvcir.2019.102740
2. Q. Wang, P. Lu, Research on application of artificial intelligence in computer network technology, *Int. J. Pattern Recogn. Artif. Intell.*, **33** (2019), 1–12. https://doi.org/10.1142/S0218001419590158
3. B. Han, X. Yang, Z. Sun, J. Huang, J. Su, OverWatch: A cross-plane DDoS attack defense framework with collaborative intelligence in SDN, *Secur. Commun. Networks*, **2018** (2018), 1–15. https://doi.org/10.1155/2018/9649643
4. S. Andrist，X. Z. Tan，M. Gleicher, B. Mutlu, Conversational Gaze Aversion for Humanlike Robots, in: *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (2014), 25–32. https://doi.org/10.1145/2559636.2559666
5. H. Zhu, S. E. Salcudean, R. N. Rohling, A novel gaze-supported multimodal human-computer interaction for ultrasound machines, *Int. J. Computer Assisted Radiol. Surgery*, **12** (2019), 1–9. https://doi.org/10.1007/s11548-019-01964-8
6. R. Wang, Y. Xu, L. Chen, GazeMotive: A Gaze-Based Motivation-Aware E-Learning Tool for Students with Learning Difficulties, in: *Human-computer Interaction-INTERACT 2019*, (2019), 544–548. https://doi.org/10.1007/ 978-3-030- 29390-1_34
7. K. B. N. Pavan, A. Balasubramanyam, A. K. Patil, B. Chethana, Y. H. Chai, GazeGuide: An eye-gaze-guided active immersive UAV camera, *Appl. Sci.*, **10** (2020), 1668. https://doi.org/10.3390/app10051668
8. C. Creed, M. Frutos-Pascual, I. Williams, Multimodal gaze interaction for creative design, in:

Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, **8** (2020), 1–13. https://doi.org/10.1145/3313831.3376196

9. X. Yan, W. Hou, X. Xu, obstacle judgment model of in-vehicle voice interaction system based on eye-tracking, in: *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, **7** (2021), 569–574. https://doi.org/10.1109/CSCWD49262.2021.9437635
10. W. Pichitwong, K. Chamnongthai, An eye-tracker-based 3D point-of-gaze estimation method using head movement, *IEEE Access*, **7** (2019), 99086–99098. https://doi.org/10.1109/ACCESS.2019.2929195
11. P. Li，X. Hou，X. Duan，H. Yip，G. Song，Y. Liu, Appearance-based gaze estimator for natural interaction control of surgical robots, *IEEE Access*, **7** (2019), 25095–25110. https://doi.org/10.1109/ACCESS.2019.2900424
12. E. Lindén, J. Sjstrand, A. Proutiere, Learning to personalize in appearance-based gaze tracking, in: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, (2019), 1140–1148. https://doi.org/10.1109/ICCVW.2019.00145
13. S. Gu, L. Wang, L. He, X. He, J. Wang, Gaze estimation via a differential eyes, in: *Appearances Network with a Reference Grid*, *Engineering*, **7** (2021), 777–786. https://doi.org/10.1016/j.eng.2020.08.027
14. X. B, J. A，Z. Zhuo, Z. A, S. C, H. D, Improved it racker combined with bidirectional long short-term memory for 3D gaze estimation using appearance cues, *Neurocomputing*, **390** (2020), 217–225. https://doi.org/10.1016/j.neucom.2019.04.099
15. K. Mora, J. M. Odobez, Geometric generative gaze estimation (G3E) for remote RGB-D cameras, in: *IEEE Conference on Computer Vision & Pattern Recognition*, (2014), 1773–1780. https://doi.org/10.1109/CVPR.2014.229
16. C. Jen, Y. Chen, Y. Lin, C. Lee, M. T. Li, Vi-sion based wearable eye-gaze tracking system, in: *2016 IEEE International Conference on Consumer Electronics (ICCE)*, (2016), 202–203. https://doi.org/10.1109/ICCE.2016.7430580
17. J. Sigut, S. A. Sidha, Iris center corneal reflection method for gaze tracking using visible light, *IEEE Trans. Biomed. Eng.*, **58** (2011), 411–419. https://doi.org/ 10.1109/ TBME.2010.2087330
18. Y. Ebisawa, K. Fukumoto, Head-free, remote eye-gaze detection system based on pupil-corneal reflection method with easy calibration using two stereo-calibrated video cameras, *IEEE Trans. Biomed. Eng.*, **60** (2013), 2952–2960. https://doi.org/ 10.1109/TBME.2013.2266478
19. M. Yu, Y. Lin, X. Tang, D. Schmidt, Y. Guo, An easy iris center detection method for eye gaze tracking system, *J. Eye Movement Res.*, **8** (2015), 1–20. https://doi.org/10.16910/ jemr.8.3.5
20. L. Sesma, A. Villanueva, R. Cabeza, Evaluation of pupil center-eye corner vector for gaze estimation using a web cam, in: *Eye Tracking Research & Application (ACM)*, (2012), 217–220. https://doi.org/10.1145/2168556.2168598
21. Y. Cheng, S. Huang, F. Wang, C. Qian, F Lu, A coarse-to-fine adaptive network for appearance-based gaze estimation, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, (2020), 10–15. https://doi.org/10.1609/aaai.v34i07.6636
22. W. Lu, Y. Li, Y. Cheng, D. Meng, B. Liang, P. Zhou, Early fault detection approach with deep architectures, *IEEE Trans. Instrum. Meas.*, **67** (2018), 1–11. https://doi.org/10.1109/TIM.2018.2800978
23. E. Lindén, J. Sjöstrand, A. Proutiere, Learning to personalize in appearance-based gaze tracking,

in: *Proceedings of the IEEE/CVF international conference on computer vision workshops*, (2019), 1140–1148. https://doi.org/10.1109/ICCVW.2019.00145

24. S. Cheng, J. Chen, C. Anastasiou, P. Angeli, O. Matar, Y. Guo, Generalised latent assimilation in heterogeneous reduced spaces with machine learning surrogate models, *J. Sci.comput.*, **94** (2023), 11. https://doi.org/ 10.1007/s10915-022-02059-4
25. S. Cheng, I. C. Prentice, Y. Huang, Y. Jin, Y. Guo, R. Arcucci, Data-driven surrogate model with latent data assimilation: Application to wildfire forecasting, *J. Comput. Phys.*, **464** (2022), 111302. https://doi.org/ 10.1016/j.jcp.2022.111302
26. J. Jiang, X. Zhou, S. Chan, S. Chen, Appearance-based gaze tracking: A brief review, in: *International Conference on Intelligent Robotics and Applications*, (2019), 629–640. https://doi.org/ 10.1007/978-3-030-27529-7_53
27. X. Zhang, Y. Sugano, M. Fritz, A. Bulling, MPIIGaze: real-world dataset and deep appearance-based gaze estimation, *IEEE Trans. Pattern. Anal. Mach. Intell.*, **1** (2019), 162–175. https://doi.org/10.1109/TPAMI.2017.2778103
28. B. Mahanama, Y. Jayawardana, S. Jayarathna, Gaze-net: appearance-based gaze estimation using capsule networks, in: *The Augmented Human International Conference*, (2020), 1–4. https://doi.org/10.1145/3396339.3396393
29. Y. Zhuang, Y. Zhang, H. Zhao, Appearance-based gaze estimation using separable convolution neural networks, in: *Electronic and Automation Control Conference (IAEAC)*, (2021), 609–612. https://doi.org/10.1109/IAEAC50856.2021.9390807
30. X. Zhang, M. Huang, Y. Sugano, A. Bulling, Training person-specific gaze estimators from user interactions with multiple devices, in: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, (2018), 1–12. https://doi.org/ 10.1145/3173574.3174198
31. P. Li, X. Hou, L. Wei, G. Song, X. Duan, Efficient and low-cost deep-learning based gaze estimator for surgical robot control, in: *2018 IEEE International Conference on Real-time Computing and Robotics (RCAR) IEEE*, (2019), 58–63. https://doi.org/ 10.1109/RCAR.2018.8621810
32. O. Lorenz, U. Thomas, Real time eye gaze tracking system using cnn-based facial features for human attention measurement, in: *Proceedings of the 14th International Joint Conference on Computer Vision*, (2019), 598–606. https://doi.org/10.5220/0007565305980606
33. J. H. Kim, S. J. Choi, J. W. Jeong, Watch & do: a smart iot interaction system with object detection and gaze estimation, *IEEE Trans. Broadcast Telev. Receivers*, **65** (2019), 195–204. https://doi.org/10.1109/TCE.2019.2897758
34. W. Luo, J. Cao, K. Ishikawa, D. Ju, A Human-Computer Control System Based on Intelligent Recognition of Eye Movements and Its Application in Wheelchair Driving, *Multi. Technol. Inter.*, **5** (2021), 50. https://doi.org/10.3390/mti5090050
35. A. Bulat, G. Tzimiropoulos, How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230 000 3D facial landmarks), in: *IEEE Computer Society*, (2017), 1021–1030. https://doi.org/10.1109/ICCV.2017.116
36. A. Newell, K. Yang, J. Deng, Stacked hourglass net-works for human pose estimation, in: *Computer Vision–ECCV 2016: 14th European Conference*, (2016), 11–14. https://doi.org/ 10.1007/978-3-319-46484-8_29
37. A. Bulat, G. Tzimiropoulos, Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources, in: *2017 IEEE International Conference*

on Computer Vision (ICCV), (2017), 3726–3734. https://doi.org/ 10.1109/ICCV.2017.400
38. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Attention is all you need, in: *Advances in neural information processing systems*, **30** (2017), 5998–6008. https://doi.org/10.48550/arXiv.1706.03762
39. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al, An image is worth 16x16 words: Transformers for image recognition at scale, in: *International Conference on Learning Representations*, (2021). https://doi.org/10.48550/arXiv.2010.11929
40. TY. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 936–944. https://doi.org/10.1109/CVPR.2017.106
41. Y. Cheng，S. Huang，F. Wang，C. Qian，F. Lu, A coarse-to-fine adaptive network for appearance-based gaze estimation, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 10623–10630. https://doi.org/10.1609/aaai.v34i07.6636
42. S. Liu, D. Liu, H. Wu, Gaze estimation with multi-scale channel and spatial attention, in: *The International Conference on Computing and Pattern Recognition*, (2023), 303–309. https://doi.org/10.1145/3436369.3437438

AIMS Press