Eng Int Syst (2025) 4: 385–395

Engineering
Intelligent Systems

Eye Tracking and Gesture Recognition System for Intelligent Human-Computer Interaction

Ke Wang[1,a] **and Weihua Zhu**[1,b*]

[1] *School of Software Engineering, Jilin Technology College of Electronic Information, Jilin 132000, Jilin, China*

In response to the issue of decreased accuracy of eye tracking and gesture recognition systems under different lighting conditions, user postures, and hand occlusions, the aim of this study is to enhance the human-computer interaction experience in smart home scenarios and construct an intelligent human-computer interaction system. The system captures eye movement images using infrared ray sources and cameras, processes iris edges using Canny edge detection, and captures changes in the user's posture using the Lucas-Kanade optical flow algorithm to improve the accuracy of eye tracking. Multi-scale convolutional neural network (CNN) layers are designed, and a self-attention mechanism is applied to enhance the precision of hand feature extraction. Long short-term memory (LSTM) networks are used to improve the accuracy of dynamic gesture recognition and achieve a more natural and intuitive interaction mode. The experimental results show that the accuracy of eye tracking reaches 95% at a light intensity of 1000 lux, and that of gesture recognition is the lowest at 89% among all light intensity conditions tested. In the case of partial occlusion, the highest success rate of gesture recognition is 95%. After interferences such as head movement and high-frequency blinking are added, the success rates of eye tracking are 95% and 94%, respectively. The experimental results demonstrate the efficiency of the system under good lighting conditions and its robustness in capturing complex gestures and occlusion situations, showing the accuracy of the optimized system for smart home control.

Keywords: Intelligent Human-Computer Interaction, Eye Tracking, Gesture Recognition, Convolutional Neural Networks, Long Short-Term Memory

1. INTRODUCTION

In the context of the rapid development of intelligent human-computer interaction systems, eye tracking and gesture recognition technology, as key interaction methods, have been widely applied in multiple fields such as virtual reality and smart homes. These technologies facilitate natural and intuitive interactions between users and computer systems. However, the accuracy of eye tracking may significantly decrease under different lighting conditions, user posture, and changes in eye position. In addition, factors such as the different shapes of the human eye and the reflection of glasses also affect the accuracy of eye tracking. Gesture recognition technology faces the problem of decreased recognition precision when it involves complex hand movements and shape changes, such as hand occlusion or complex backgrounds. The diversity of habitual gestures, hand sizes, and the postures of users also poses challenges for gesture recognition. These factors limit the application scope of eye tracking and gesture recognition technology; hence, the improvement of system performance has become an important research topic in this field.

This article proposes a comprehensive solution to the accuracy and robustness issues of eye tracking and gesture recognition technology in smart home scenarios. The proposed system can extract eye features and hand movements with high precision, effectively addressing challenges such as lighting changes, posture differences, and hand occlusion. The experimental results validate the effectiveness of the method, providing an efficient and intuitive means of achieving intelligent human-computer interaction.

*Corresponding author. [a]Email: 13843225515@163.com, [b]zhuweihua19760525@163.com.

This article first introduces the importance and challenges of eye tracking and gesture recognition technology in intelligent human-computer interaction systems. Then, the methods used to optimize the performance of eye tracking and gesture recognition technology are explained in detail. This is followed by a description of the experimental setup and evaluation process, including eye tracking precision testing under different lighting conditions and user postures, as well as the testing of the accuracy and robustness of the proposed gesture recognition system. Finally, the research findings are summarized and their potential value in applications such as smart home control is discussed.

Main contributions:

(1) The eye tracking method is optimized to improve the system's eye tracking precision under different lighting conditions and user postures.

(2) A gesture recognition system enhanced with multi-scale convolutional neural networks (CNNs) and self-attention mechanism is designed, significantly improving the accuracy of hand feature extraction, especially against complex backgrounds and in hand occlusion situations.

(3) The multimodal data fusion of eye tracking and gesture recognition is achieved, dynamically adjusting feature weights through shared convolutional layers and Bayesian optimization, and enhancing the naturalness and robustness of the interaction system.

2. RELATED WORK

In previous research, many scholars have explored various aspects of eye tracking and gesture recognition. They explored the factors affecting the quality of eye tracking data, the diversity and research trends of eye tracking applications in virtual reality, and new research directions for gaze interaction and eye tracking in the field of augmented reality, solving the problems of data quality control, diversity assessment of virtual reality applications, and innovative gaze interaction in the field of extended reality in eye tracking research [1–3]. Kaduk et al. [4] addressed the performance evaluation issues of low-cost eye tracking technology in terms of precision and real-time performance by comparing a webcam-based eye tracking system with a laboratory-level EyeLink 1000 eye tracker. Their research results showed that although the precision of the webcam-based system was slightly lower than that of EyeLink, it performed well in terms of accuracy and precision in focusing on targets, and was comparable to the performance of existing mobile eye tracking devices. Dunn et al. [5] proposed a set of minimum reporting projects based on consensus and open invitations from the international eye tracking community. This report addressed the issue of inconsistent reporting standards in eye tracking research. Boerman et al. [6] solved the problem of Instagram users' recognition and understanding of influencer marketing advertisements through eye tracking and online experiments, and revealed users' preference for recognizing influencer marketing leads, as well as the impact of disclosure, brand presence, and influencer type on users' persuasive knowledge level. Qi et al. [7] analyzed the application of monocular cameras in gesture recognition. They explored the entire visual gesture recognition process from data collection to classificationand discussed algorithm progress and system improvements, solving the efficiency problem of gesture recognition in natural human-computer interaction. Li et al. [8] proposed a domain-independent real-time millimeter wave gesture recognition system, which solved the problems of poor adaptability, large collection of data, and insufficient real-time recognition performance of existing gesture recognition systems in new fields. By designing a data augmentation framework and spatiotemporal gesture segmentation algorithm, the recognition accuracy of the system for new users, new environments, and new locations was significantly improved, enhancing the robustness and effectiveness of the system. Lee et al. [9] solved the problem of repeated gesture detection in real-time dynamic gesture recognition. By defining and quantifying gesture progress sequences, the compression of gesture sequences was achieved to eliminate repeated changes; also, classification recognition was performed, improving the accuracy and efficiency of the system in recognizing repeated and non-repeated gestures in continuous gesture streams. Gao et al. [10] proposed a quality-oriented signal processing framework by segmenting gesture signal time series and establishing mathematical models to evaluate the quality of signal perception. They optimized signal processing strategies and solved the position dependency problem in wireless signal-based gesture recognition. Significant progress has been made in the research of eye tracking and gesture recognition technology. Scholars have conducted in-depth discussions on data quality, application diversity, and new research directions, proposing various solutions to improve system performance and user experience.

In intelligent human-computer interaction systems, gesture recognition and eye tracking are widely used. Scholars have made some improvements to address some issues with the entire human-computer interaction system. Researchers have developed machine learning algorithms and manual rules for gesture recognition and designed customized gesture datasets and algorithms to achieve dynamic gesture control, thereby promoting the development of natural and intelligent human-computer communication interfaces [11–13]. Li et al. [14] proposed a gesture recognition method based on a dual-channel region convolutional neural network, which solved the problems of low accuracy and poor real-time performance in visual gesture recognition in human-computer interaction systems, thus improving recognition precision and robustness and enhancing the interpretability and efficiency of human-computer interaction. Ma et al. [15] improved the robustness and user experience of the system by using blink-based action combinations as interactive object triggers, while optimizing the size and spacing of interactive objects and constructing an efficient and usable eye-controlled multimedia player model. Sevcenko et al. [16] proposed a cognitive load measurement method based on eye movement tracking. By analyzing the eye movement data of players in simulation tasks, the difficulty and performance of tasks were successfully predicted, solving the problem of real-time monitoring and adaptation of operator cognitive load in human-computer interaction systems. Wang et al. [17] proposed an eye tracking

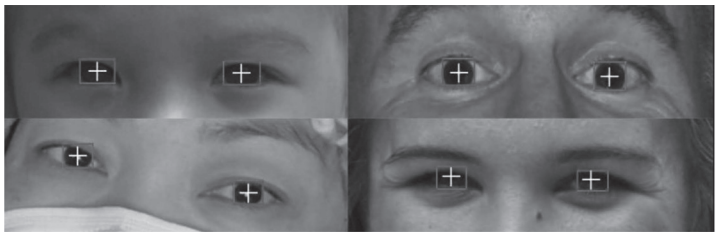

Figure 1 Iris positions and pupil positions.

method suitable for aircraft cockpits. The precision of gaze estimation was improved through a multi-camera system and a hybrid network structure, which solved the problem of limited head movement for operators and enhanced the capture rate of targets of different sizes, thereby improving the efficiency and intelligence level of human-computer interaction in intelligent aircraft cockpits. Wang et al. [18] presented a human-environment interaction system based on eye tracking and brain-machine interface. By detecting the attention and gaze points of patients with amyotrophic lateral sclerosis, voice control of household appliances was achieved, solving the daily self-care difficulties caused by the loss of motor ability in patients. Bharath et al. [19] proposed a human-computer interaction system based on facial expressions. Using a webcam to capture eye and mouth movements, facial features were recognized and processed through a Haar classifier to achieve hands-free control of virtual mice and keyboards, solving the problem of traditional input devices being difficult for paralyzed and physically disabled people. Gunawardena et al. [20] made a comprehensive review of 36 studies conducted between 2010 and 2020, analyzing in depth the progress of mobile device eye tracking technology in terms of algorithms, devices, and calibration methods and exploring its applications in fields such as healthcare and education. At the same time, the limitations of existing technologies were pointed out, and a real-time eye movement tracking solution based on edge computing was proposed. The literature indicates that improving the accuracy and real-time performance of eye tracking and gesture recognition, enhancing the robustness of the system, and optimizing the human-computer interaction interface are the focuses of current research.

3. IMPLEMENTATION OF SYSTEM

In recent years, with the rapid development of smart home technology, users' demand for more natural and intuitive human-computer interaction methods has been constantly increasing. Traditional remote controls and voice control methods are often complex to operate and inefficient in certain scenarios. Eye tracking technology captures the trajectory of users' eye movements, which allows them to directly manipulate smart home devices by staring at them or at an interface point, providing a contactless interaction experience. At the same time, gesture recognition technology achieves fast and flexible command input by detecting the users' hand movements. The combination of the two not only improves the precision of control; it also simplifies the operation process and facilitates an efficient and natural interaction. However, complex lighting environments, changes in user posture, and errors in eye-hand coordination often affect the precision and robustness of the system. The method proposed in this article optimizes eye tracking and gesture recognition technology to enhance the multimodal interaction experience in smart home scenarios.

3.1 Optimization of Eye Tracking

In eye tracking, infrared ray sources are used to illuminate the eyes, reducing the interference of ambient light, and cameras are used to capture eye movement images. Eye data obtained from images is processed, and eye features including the position of pupils and the corneal reflection are extracted. Then, the direction of the gaze is determined. Figure 1 shows the captured images of the iris and pupil positions.

The above images were obtained from the Ricis Nabati Computer Vision Project dataset. In Figure 1, the red boxes represent the iris positions, and the intersections of two white line segments represent the pupil position. After initial localization, the edges of the irises are processed using Canny edge detection [21–22]. Circles are fitted to find the center position of the irises. Corneal reflection points are bright spots formed by the light source shining on the surface of the corneas. The threshold segmentation is used to extract corneal reflection points from the background [23–24].

Figure 2 shows the positions of the pupil and corneal reflection points on a three-dimensional image of the eyeball. The coordinate (x_p, y_p) of the pupil and the coordinate (x_r, y_r) of corneal reflection point are obtained. The direction of gaze is approximated by calculating the vector between the corneal reflection point and the center of the pupil, and its direction vector $\overrightarrow{d}$ is represented as:

$$\overrightarrow{d} = (x_p - x_r, y_p - y_r) \tag{1}$$

This vector reflects the offset direction of the gaze relative to the corneal reflection point. The specific gaze angle is described through an eye rotation model. The horizontal rotation angle of the eyeball in the image is θ_x, and the vertical rotation angle is θ_y. The distance from the center of the eyeball to the cornea is measured and represented by r. The above three parameters satisfy the relationships:

$$\theta_x = \arctan\left(\frac{x_p - x_r}{r}\right) \tag{2}$$

$$\theta_y = \arctan\left(\frac{y_p - y_r}{r}\right) \tag{3}$$

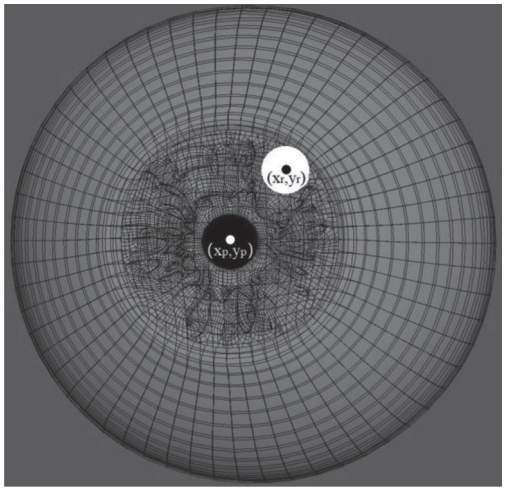

Figure 2 Three-dimensional image of the eyeball.

The calculation of gaze direction is based on the relative position between the pupil position and the corneal reflection point and is converted into the actual gaze direction through a geometric model. During this process, complex lighting and changes in user posture are the main issues. To improve the performance of the eye tracking system under complex lighting conditions, the CNN is optimized.

When performing feature extraction of CNN, the depth and width of the convolutional layer are increased, and the sensitivity to local features is enhanced. Feature extraction is performed through convolution kernel [25–26].

$$f_{i,j} = \sum_{m=1}^{3\times3}\sum_{n=1}^{3\times3} W_{m,n} \cdot I_{i+m-1,j+n-1} \tag{4}$$

where $f_{i,j}$ is the pixel value of position (i, j) in the output image. 3×3 represents the size of the convolution kernel. $W_{m,n}$ is the weight value of the convolution kernel. $I_{i+m-1,j+n-1}$ is the adjusted image position coordinate. Multi-layer convolution processing can capture subtle structures such as pupils and eyelids, strengthen resistance to lighting changes, and improve the stability of eye tracking.

The use of a large number of convolutional layers can easily lead to gradient vanishing problems, making model training difficult. Therefore, this study uses skip connections for processing:

$$y = F(x, W) + x \tag{5}$$

where y is a feature map that undergoes skip connections; $F(x, W)$ is the convolution operation in the residual block; and x is the input feature map of the residual block. In skip connections, the input feature map is directly added to the convolved feature map, allowing the model to learn the features of lighting changes more effectively.

When the head tilts or rotates, the relative position of the eyes undergoes significant changes, resulting in inaccurate eye tracking results and a decrease in system precision. Therefore, in this study, the Lucas-Kanade optical flow algorithm is used to capture real-time changes in user posture and dynamically adjust the eye tracking results [27–28]. Lucas-Kanade analyzes the pixel movement between consecutive frames and calculates the trajectory of the user's eye motion areas. In continuous image sequences, when the user's posture changes and causes eye displacement, the system can adjust the predicted results of eye position to maintain the stability of eye tracking.

Lucas-Kanade is used to calculate the eye motion vectors in each frame of the image, and the motion vectors are used to dynamically adjust the gaze direction prediction output by the CNN model. If the user's eyes move to the left, the system adjusts the CNN-predicted gaze direction based on this information to compensate for this change in eye positions. In this study, Kalman filters are applied to eliminate noise from different sources of information [29–30]. The filter combines CNN's gaze direction prediction with the motion vector of optical flow, and weighted averaging is used to obtain the final eye tracking results.

3.2 Optimization of Gesture Recognition

The indoor environment of a home is relatively complex, and when there are many indoor items, it is easy to encounter occlusion. Therefore, for gesture recognition, it is necessary to address the issues of hand feature recognition and occlusion.

Gesture recognition systems need to accurately extract spatial features from hand images. When processing hand image information, multi-scale CNN convolutional layers are set up for feature extraction. The CNN model uses a single-size convolution kernel for hand image recognition during image processing, which can easily overlook the details in the hand images and lead to inaccurate feature extraction. Therefore, in this study, convolution kernels of different sizes are designed to capture hand feature information. The scale of the large convolution kernel is set to 5×5 to extract the overall contour and shape information of the hands, and the scale of the small convolution kernel is set to 1×1 to capture small features such as finger joints and fingertips. The feature maps processed by convolution kernels of different sizes are merged in the subsequent layers of the network, allowing the model to simultaneously consider both the global contour and the local details of hands.

To further enhance the robustness of gesture recognition systems in complex environments, self-attention mechanisms

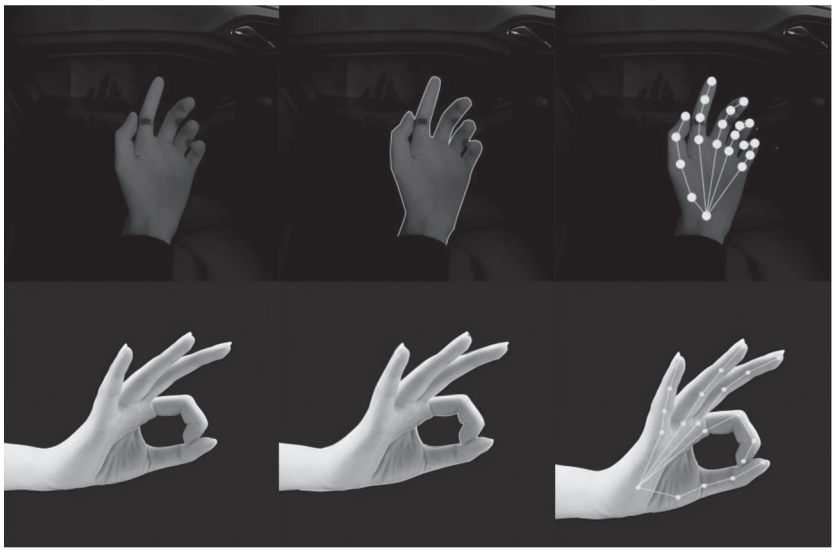

Figure 3 Hand feature extraction.

are used in this study [31–32] to enable CNN to ignore irrelevant background information in images, effectively reducing background interference. In hand images, the hand areas occupy only small parts of the images, while background information accounts for most of the pixels. The CNN model calculates the correlation between each pixel in the input images and other pixels and assigns an attention weight to each pixel based on these correlations. The hand areas have a higher correlation with other hand pixels, so the self-attention mechanism can automatically assign high weights to the hand areas, allowing the model to focus more on the extraction of hand features. Figure 3 shows the visualization results of hand feature extraction.

Each row in Figure 3 represents the feature extraction of a hand image, with the original image, contour extraction image, and hand feature point extraction are the images from left to right. Through the design of multi-scale convolution kernels, gesture recognition systems can simultaneously capture global and local features of hands, improving the recognition accuracy of the model in complex environments. Meanwhile, combined with the self-attention mechanism, the model can automatically ignore the interference of complex backgrounds and focus on feature extraction in the hand areas.

When users perform quick gestures, the system needs to capture precisely the time series changes of hand movements. The dynamic changes of gestures contain certain temporal information, and CNN is good at processing static images; hence, in this study, LSTM is integrated in CNN to process the temporal information of gestures. The hand spatial features extracted by CNN are passed as input sequences to the LSTM network, which dynamically tracks the features and generates a final temporal feature vector to describe the overall changes of gesture.

In the smart home scenario, a user's hands may be obstructed by the position of their body or by external objects or. To solve this problem, an occlusion processing module is designed. This module utilizes generative adversarial networks (GAN) to reconstruct features [33–34]. When the hands are obstructed, the system first detects and marks the obstructed area through CNN. Then GAN is used to generate a network for image completion of the area. The generative network of GAN takes partially-occluded hand images as input and outputs an area-completed image that is similar to the real hand structure. Through adversarial training, GAN gradually learns how to generate reasonable completed images that conform to the anatomical structure of the hand in partially-occluded situations [35–37]. The discriminative network is responsible for determining whether the generated completed images are realistic enough. The entire training process involves the generator and discriminator competing with each other to continuously improve the quality of the completed images, ensuring that the system can continue to recognize complete gestures. Figure 4 shows the visualization results of occlusion processing.

Each row in Figure 4 represents the feature extraction of a hand image, from left to right representing the original image, feature point extraction without occlusion processing module, and feature extraction with occlusion module. Due to occlusion, all feature points of the hands in the images cannot be extracted. The occlusion processing module completes the remaining feature points of the hands (red feature points).

3.3 Multimodal Data Fusion

In intelligent human-computer interaction, the data streams of eye tracking and gesture recognition need to be recognized simultaneously. In this study, the two types of data are combined to solve the problems of delay and asymmetry in the system.

To simultaneously process feature data for eye tracking and gesture recognition, a shared layer is set up in the convolutional layer. The eye tracking data comes mainly from images of eye areas, and the gesture recognition data comes from hand images. The traditional approach is to extract features separately from these modal data and then perform post fusion. This method can easily lead to a mismatch in the data feature space between the two modalities, ultimately affecting the fusion effect. Therefore, shared convolutional layers are used to share weights, enabling the system to capture

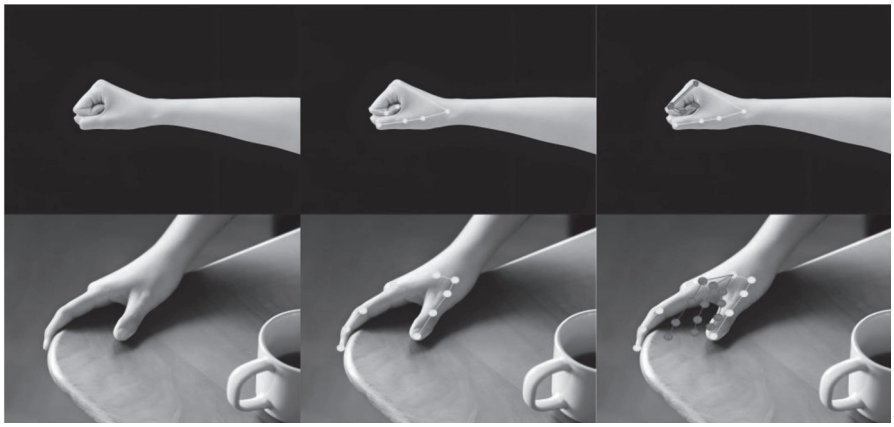

Figure 4 Visualization of occlusion processing.

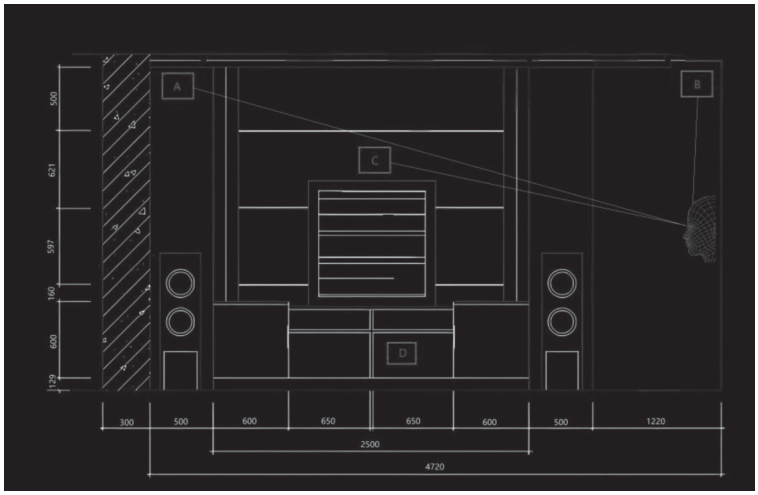

Figure 5 Plan of device layout.

the common underlying feature structure in both modalities when extracting gesture and eye data features, avoiding the fragmentation of the feature extraction process.

The feature data of eye tracking and gesture recognition have different temporal and spatial distribution characteristics. To solve this problem, this article adjusts the feature maps of eye tracking and gesture recognition to the same numerical range through batch normalization. On the time series, the two features also need to be aligned. The distance between two time series is calculated, and the alignment path is adjusted. Precise synchronous response between the two during the interaction process should be ensured.

After completing temporal matching, features from eye tracking and gesture recognition are integrated. The system performs fusion based on weighted features. The weights of features are dynamically adjusted according to different tasks. When users control home devices, their first step is to select the device. This is guided mainly by eye tracking, and the system prioritizes the results of this tracking. When operating the device, the system is dominated by gesture recognition results. Weight adjustment is carried out through Bayesian optimization, which automatically adjusts the weight distribution of each modal feature based on historical interaction data to ensure that the system's performance remains optimal in different application scenarios.

3.4 System Applications

The intelligent human-computer interaction system designed in this study provides users with a means of contactless home appliance control by integrating eye tracking and gesture recognition technologies. The system design focuses on the home environment and arranges relevant devices to achieve control over home devices such as lighting, air conditioning, and television.

Firstly, the sensing devices are arranged. Cameras are installed in the four corners close to the ceiling, and clear-view cameras are installed near the TV, air conditioner, and lighting fixtures for better control and improved overall recognition accuracy. A central controller is set up and connected to various household appliances via wireless network, responsible for receiving instructions from the eye tracking and gesture recognition system and sending control signals to household appliances. Each household appliance is equipped with a wireless receiving module that receives instructions from the central controller, such as switching and adjusting.

Figure 5 shows the placement of cameras and central controllers in the living room. In the figure, areas A and B are the high cameras in the living room; area C is a camera with a TV perspective; area D is the central controller. The camera

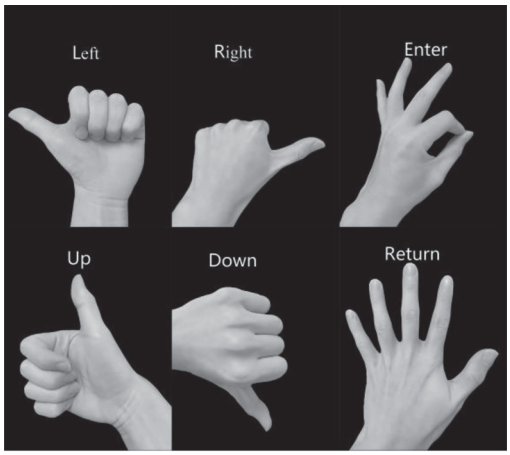

Figure 6 Basic functional gestures.

B in the figure cannot track the user's eyes due to angle issues, while A and C have good angles for eye tracking. The camera model is Intel RealSense D435. The central processing unit uses NVIDIA Jetson Xavier NX. The wireless receiving module of the home appliance uses ESP8266 module.

The main task of eye tracking is to help users select the household appliances they need to operate. When the system detects that the user's gaze stays on a specific household appliance for more than 3 seconds, the device is marked as "selected". The user's gestures are coordinated to confirm that it has been selected accurately. This feedback improves the naturalness of interaction and reduces the incidence of inaccurate operation.

After the eye tracking system completes device recognition, the gesture recognition system is responsible for performing specific operations on the devices. The system uses infrared cameras and depth cameras to capture the user's hand movements, and recognize the gestures. Figure 6 displays some basic functional gestures.

Figure 6 contains six basic functional gestures, namely left, right, enter, up, down, and return. On the basis of basic functional gestures, dynamic gestures such as waving and clenching fists are added to control household appliances. Taking TV control as an example, when controlling the TV, the user first rests their gaze on the TV and uses the enter gesture to select it. The left and right directional gestures are used to switch channels. The volume is adjusted by using the up and down gestures, with up signifying an increase of volume, and down meaning a decrease of it. When the user selects a program or input source, the enter gesture is used to confirm and prevent a wrong operation. The operable state can be returned to using the return gesture. If the TV needs to be turned off, the dynamic gesture of waving to the left or the right is used.

During the system application process, ensuring low latency and high concurrency can effectively improve the user experience. When arranging devices, the system performs several data processing tasks on the device side to reduce the transmission latency that relies on remote servers. In high-concurrency scenarios, the system optimizes the interaction process through asynchronous task scheduling to ensure smooth and real-time response when multiple users or devices interact simultaneously. The eye tracking and gesture recognition system involves a large amount of personal behavior data. Therefore, it is crucial to ensure data privacy. The system encrypts user data using advanced encryption standards during transmission and storage to prevent data theft or breach.

4. SYSTEM PERFORMANCE EVALUATION

4.1 Precision Evaluation Experiment

Firstly, the accuracy of eye tracking is tested under different lighting conditions and user postures. An indoor space that meets the requirements of ordinary home furnishings is selected as the testing site, and sensing devices are installed. The experimental site is equipped with adjustable lighting with five levels of light intensity: 100 lux, 300 lux, 500 lux, 700 lux, and 1000 lux. Tests are conducted with the user in front, side, bent down, and head up postures, with the test content being the user gazing at a target point within a specified area. The system records the focal position of eye tracking, which is compared with manually annotated real data, and the accuracy, error rate, and missed detection rate of the system are calculated. The test is conducted 100 times under each light intensity. The results are shown in Table 1:

As shown in Table 1, the experiment tests three indicators of eye tracking for the intelligent human-computer interaction system under different light intensities and user postures. Overall, as the light intensity increases, the accuracy of eye tracking gradually improves. When the light intensity is 100 lux, the accuracy of front eye tracking is 83%. When the light intensity rises to 1000 lux, the accuracy increases to 95%. There are also differences between various postures under the same light intensity. The system performs best in the front posture, while there is still a certain performance gap in the postures of side, head up, and bent down due to changes in gaze angles and lighting conditions. The light intensity has a significant impact on the accuracy of the eye tracking system, indicating the importance of optimizing system performance under different lighting conditions.

Table 1 Test results for eye tracking under different light intensities.

Light Intensity (lux)	User Posture	Eye Tracking Accuracy (%)	Error Rate (%)	Missed Detection Rate (%)
100	Front	83	12	5
	Side	77	15	8
	Bent down	81	11	8
	Head up	79	13	8
300	Front	86	9	5
	Side	80	11	9
	Bent down	81	12	7
	Head up	83	10	7
500	Front	89	8	3
	Side	84	10	6
	Bent down	87	9	4
	Head up	83	11	6
700	Front	92	6	2
	Side	86	9	5
	Bent down	89	8	3
	Head up	88	8	4
1000	Front	95	5	0
	Side	91	6	3
	Bent down	92	7	1
	Head up	90	7	3

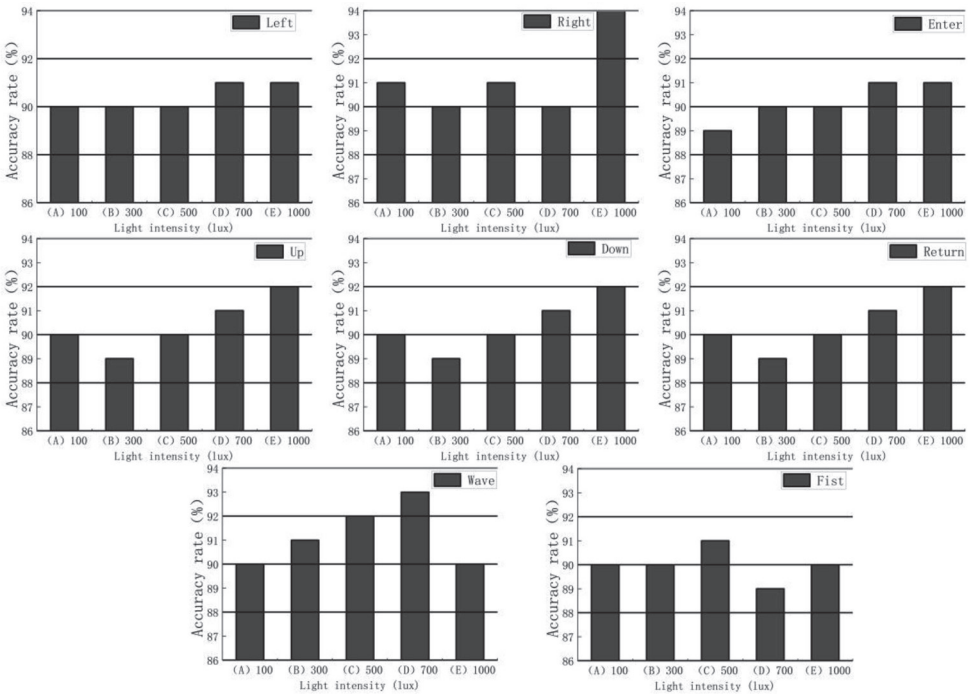

Figure 7 Test results for gesture recognition under different light intensities.

The accuracy of gesture recognition is tested by the system under different light intensities, recognizing different hand postures 100 times each. The results obtained are shown in Figure 7:

Figure 7 shows the recognition results of 8 hand postures, including left, right, enter, up, down, return, wave, and fist, under different light intensities. Compared with eye tracking, gesture recognition changes less with increasing light intensities and, in some cases, the probability of gesture recognition decreases with increasing light intensities. Gesture recognition reaches a minimum of 89% in all tests, and its overall recognition accuracy is better than that of the eye tracking. The experimental results show that the gesture recognition system design has better adaptability to changes in lighting, and the eye tracking is more sensitive to changes in lighting. The overall accuracy of the system is better under good lighting conditions.

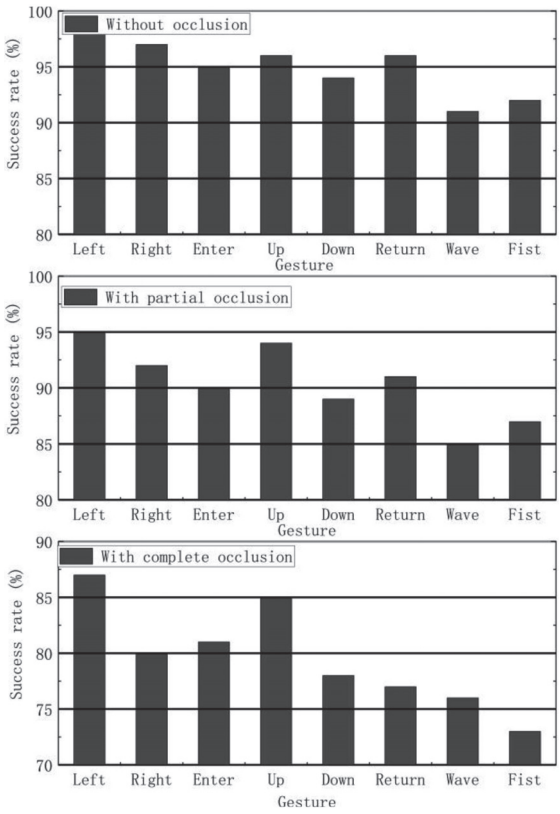

Figure 8 Test results for hand occlusion.

Table 2 Test results for robustness of eye tracking.

Interference Factor	Success Rate (%)	Average Offset Error (cm)	Average Response Time (s)
No Interference	98	0.5	0.1
Head Movement	95	1.2	0.18
High-Frequency Blinking	94	1.1	0.15
Wearing Frameless Glasses	96	0.7	0.12
Wearing Sunglasses	85	2	0.25

4.2 Robustness Testing

The purpose of this experiment is to evaluate the stability of intelligent human-computer interaction systems. Firstly, its performance in hand occlusion scenarios is tested. A laboratory space is set up and arranged as a regular living room. Three types of hand occlusion scenarios are designed: without occlusion, with partial occlusion, and with complete occlusion, to ensure the different positions and states of the hand in the cameras.After selecting the gestures, we randomly test each gesture with fast and slow speeds. Then we test the gestures under each occlusion condition and record the system's successful recognition rate. The results obtained are shown in Figure 8:

Figure 8 shows the success rate of gesture recognition for the intelligent human-computer interaction system in three scenarios: without occlusion, with partial occlusion, and with complete occlusion. From the results, it can be seen that the recognition rate of the system is highest without occlusion, with a maximum success rate of 98%. In the case of partial occlusion, the overall success rate decreases, with the highest success rate reaching 95% and the lowest success rate dropping to 85%. When completely occluded, the success rate decreases significantly, with the success rates of wave and fist gestures dropping to 76% and 73%, respectively. Experiments show that the intelligent human-computer interaction system is more responsive to simple gestures such as "left" and "right", but its performance is weakened when dealing with complex gestures and severe occlusion.

The robustness of the system in terms of eye tracking function is tested. Gaze interference factors are added in the laboratory space, including head movement, high-frequency blinking, and wearing of glasses. Under each interference factor, several target points are set, and the user is required to gaze at the target points under these interference conditions, and the system records the gaze positions. One hundred experiments are conducted under each interference condition, and the system's eye tracking success rate, average offset error (distance from the target point), and average response time are recorded. The results obtained are shown in Table 2:

Table 2 shows the tracking performance of the intelligent human-computer interaction system in the presence of different gaze interference factors. Without interference, the system has a success rate of 98%, an average offset error of 0.5cm, and an average response time of 0.1 seconds, indicating that the system has high accuracy and response speed under ideal

conditions. When interference such as head movement and high-frequency blinking is added, the success rates decrease to 95% and 94% respectively, the average offset error increases to 1.2cm and 1.1cm, and the response time also increases. The wearing of frameless glasses has a relatively small impact on the system, while the wearing of sunglasses reduces the success rate to 85% and increases the average offset error to 2cm, indicating that visible light occlusion has a significant impact on system performance. Overall, the system has good robustness, although its tracking precision and response time are poor when sunglasses are worn.

5. CONCLUSION

The method proposed in this study, which involves integrating computer vision and deep learning technologies, significantly improves the performance of eye tracking and gesture recognition in the intelligent human-computer interaction system.. The precision of eye tracking under varying lighting conditions and user postures is optimized through infrared ray sources and edge detection technology. For gesture recognition, the application of multi-scale CNN and self-attention mechanism enhances the system's ability to extract hand features, especially in scenarios with occlusion and complex backgrounds. In addition, this study effectively fuses eye data and gesture data, improving the naturalness and accuracy of interaction through shared convolutional layers and batch normalization processing. The experimental results validate the effectiveness of the proposed method and demonstrate its potential in applications such as smart home control. In future research, more complex user interaction scenarios can be further explored, and algorithms can be optimized to adapt to more diverse environments and user behaviors, while emphasizing the real-time performance and scalability of the system to promote the widespread application of intelligent human-computer interaction technology.

FUNDING

This work was supported by the Education Department of Jilin Province 2024 Annual Vocational Education and Adult Education Teaching Reform Research Project. Project name: Application and Research of Virtual Simulation Mixed Teaching Mode Based on VR/AR Technology in Higher Vocational Colleges. Project number: 2024ZCY427

REFERENCES

1. Niehorster, D.C., Nyström, M., Hessels, R.S. et al. The fundamentals of eye tracking part 4: Tools for conducting an eye tracking study. Behav Res 57, 46 (2025).
2. Adhanom I B, MacNeilage P, Folmer E. Eye tracking in virtual reality: a broad review of applications and challenges. Virtual Reality, 2023, 27(2): 1481–1505.
3. Plopski A, Hirzle T, Norouzi N, et al. The eye in extended reality: A survey on gaze interaction and eye tracking in head-worn extended reality. ACM Computing Surveys (CSUR), 2022, 55(3): 1–39.
4. Kaduk T, Goeke C, Finger H, Konig P. Webcam eye tracking close to laboratory standards: Comparing a new webcam-based system and the EyeLink 1000. Behavior Research Methods, 2024, 56(5): 5002–5022.
5. Dunn M J, Alexander R G, Amiebenomo O M, Arblaster G, Atan D, Erichsen J T, et al. Minimal reporting guideline for research involving eye tracking (2023 edition). Behavior Research Methods, 2024, 56(5): 4351–4357.
6. Boerman S C, Muller C M. Understanding which cues people use to identify influencer marketing on Instagram: an eye tracking study and experiment. International Journal of Advertising, 2022, 41(1): 6–29.
7. Qi J, Ma L, Cui Z, Yu Y. Computer vision-based hand gesture recognition for human-robot interaction: a review. Complex & Intelligent Systems, 2024, 10(1): 1581–1606.
8. Li Y, Zhang D, Chen J, Wan J, Zhang F, Hu Y, et al. Towards domain-independent and real-time gesture recognition using mmwave signal. IEEE Transactions on Mobile Computing, 2022, 22(12): 7355–7369.
9. Lee M, Bae J. Real-time gesture recognition in the view of repeating characteristics of sign languages. IEEE Transactions on Industrial Informatics, 2022, 18(12): 8818–8828.
10. Gao R, Li W, Xie Y, Yi E, Wang L, Wu D, et al. Towards robust gesture recognition by characterizing the sensing quality of WiFi signals. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2022, 6(1): 1–26.
11. Chua S N D, Chin K Y R, Lim S F, Jain P. Hand gesture control for human–computer interaction with Deep Learning. Journal of Electrical Engineering & Technology, 2022, 17(3): 1961–1970.
12. Jain R, Jain M, Jain R, Madan S. Human Computer Interaction – Hand Gesture Recognition. Advanced Journal of Graduate Research, 2022, 11(1): 1–9.
13. Chakravarthi S S, Rao B, Challa N P, Ranjana R, Rai A. Gesture Recognition for Enhancing Human Computer Interaction. Journal of Scientific & Industrial Research, 2023, 82(04): 438–443.
14. Li P, Zhao L. A novel art gesture recognition model based on two channel region-based convolution neural network for explainable human-computer interaction understanding. Computer Science and Information Systems, 2022, 19(3): 1371–1388.
15. Ma G R, He J X, Chen C H, Niu Y F, Zhang L, Zhou T Y. Trigger motion and interface optimization of an eye-controlled human-computer interaction system based on voluntary eye blinks. Human–Computer Interaction, 2024, 39(5–6): 472–502.
16. Sevcenko N, Appel T, Ninaus M, Moeller K, Gerjets P. Theory-based approach for assessing cognitive load during time-critical resource-managing human–computer interactions: An eye-tracking study. Journal on Multimodal User Interfaces, 2023, 17(1): 1–19.
17. Wang L, Wang C, Zhang Y, Gao L. An integrated neural network model for eye-tracking during human-computer interaction. Mathematical Biosciences and Engineering, 2023, 20(8): 13974–13988.
18. Wang J, Xu S, Dai Y, Gao S. An eye tracking and brain–computer interface-based human–environment interactive system for amyotrophic lateral sclerosis patients. IEEE Sensors Journal, 2022, 23(20): 24095–24106.
19. Bharath M R R,Kumar L D, Jayasuriya C, Sundaram T M. Controlling mouse and virtual keyboard using eye-tracking by computer vision. Journal of Algebraic Statistics, 2022, 13(3): 3354–3368.

20. Gunawardena N, Ginige J A, Javadi B. Eye-tracking technologies in mobile devices using edge computing: A systematic review. ACM Computing Surveys, 2022, 55(8): 1–33.
21. Sekehravani E A, Babulak E, Masoodi M. Implementing canny edge detection algorithm for noisy image. Bulletin of Electrical Engineering and Informatics, 2020, 9(4): 1404–1410.
22. Wu F, Zhu C, Xu J, Bhatt M W, Sharma A. Research on image text recognition based on canny edge detection algorithm and k-means algorithm. International Journal of System Assurance Engineering and Management, 2022, 13(Suppl 1): 72–80.
23. Tian Wenqi, Li Zhen, Duan Xintao, Zhang Runze. An Adaptive Backlight Image Processing Algorithm Based on Threshold Segmentation. Computer & Digital Engineering, 2020, 48(10): 2465–2470.
24. Dai Yingxin, Wu Ruixian, Wang Zhiguo. Research on Adaptive Threshold Segmentation Algorithm Based on PET/CT. China Medical Devices, 2022, 37(3): 79–83.
25. Cong S, Zhou Y. A review of convolutional neural network architectures and their optimizations. Artificial Intelligence Review, 2023, 56(3): 1905–1969.
26. Habib G, Qureshi S. Optimization and acceleration of convolutional neural networks: A survey. Journal of King Saud University-Computer and Information Sciences, 2022, 34(7): 4244–4268.
27. Rajasekaran G, Raja Sekar J. Abnormal Crowd Behavior Detection Using Optimized Pyramidal Lucas-Kanade Technique. Intelligent Automation & Soft Computing, 2023, 35(2): 2399–2412.
28. Hambali R, Legono D, Jayadi R. The application of pyramid Lucas-Kanade optical flow method for tracking rain motion using high-resolution radar images. Jurnal Teknologi, 2020, 83(1): 105–115.
29. Ren Jiamin, Gong Ningsheng, Han Zhenyang. Multi-Target Tracking Algorithm Based on YOLOv3 and Kalman Filter. Computer Applications and Software, 2020, 37(5): 169–176.
30. Yin Kuang, Wang Hongbin, Hu Fan, Zhang Tie, Fang Jian, La Yuan. Research on Target Tracking Technology Based on Switched Kalman Filter. Machine Tools & Hydraulics, 2021, 49(12): 23–28.
31. Zhang Shujun, Peng Zhong, Li Hui. SAU-Net: Medical Image Segmentation Method Based on U-Net and Self-Attention. Acta Electronica Sinica, 2022, 50(10): 2433–2442.
32. Zhou Yutao, Wu Huayi, Cheng Hongquan, Zheng Jie, Li Xuexi. Pedestrian Trajectory Prediction Model Based on Self-Attention Mechanism and Group Behavior Characteristics. Geomatics and Information Science of Wuhan University, 2020, 45(12): 1989–1996.
33. Saxena D, Cao J. Generative adversarial networks (GANs) challenges, solutions, and future directions. ACM Computing Surveys (CSUR), 2021, 54(3): 1–42.
34. Jabbar A, Li X, Omar B. A survey on generative adversarial networks: Variants, applications, and training. ACM Computing Surveys (CSUR), 2021, 54(8): 1–49.
35. Wang, Yetong, Li, Guozhang, Xing, Kongduo, Alfred, Rayner. Automatic Classification and Recognition of Spatiotemporal High-resolution Image Data Based on Deep Neural Networks. Engineering Intelligent Systems, 2024, 32(4): 309–317.
36. Gu, Yan, Gu, Wanli, Wang, Qi, Kim, Duk-Hwan. Visual Design of Computer Human-Computer Interaction Interface Based on Wireless Network. Engineering Intelligent Systems, 2024, 32(3): 267–275.
37. Wei, Jie. Agile Supply Chain Management Collaboration Based on Artificial Intelligence Traceability System. Engineering Intelligent Systems, 2024,32(5): 401–410.

Ke Wang was born in Jilin, Jilin, P.R. China, in 1982. She received the Master degree from Northeast Electric Power University, P.R. China. Currently, she is working at the School of Software Engineering, Jilin Technology College of Electronic Information. Her research interests include computer applications and network security.
E-mail: 13843225515@163.com

Weihua Zhu was born in Jiaohe, Jilin, P.R. China, in May 1976. He received the Master degree from Harbin Institute of Technology, P.R. China. Currently, he works at the School of Software Engineering, Jilin Technology College of Electronic Information. His research interests include Intelligent network and Internet of Things technology applications.
E-mail: zhuweihua19760525@163.com