Transactions on Computational and Scientific Methods | Vo. 5, No. 1, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Deep Learning-Based Gesture Key Point Detection for Human-Computer Interaction Applications

Shiyu Duan
Carnegie Mellon University, Pittsburgh, USA
naomiduansy@gmail.com

Abstract: This paper studies a natural human-computer interaction system based on gesture key point detection, aiming to achieve accurate interaction between users and virtual devices by efficiently extracting hand key points. The system uses a deep learning model to process gesture images, extract the positions of key points such as finger joints and palms, and convert dynamic gestures into specific instructions through timing analysis. In the simulation experiment, users control virtual devices through gestures and complete remote-control tasks such as light switches, robotic arm operations, and drone path planning. The experimental results show that the system exhibits a high success rate and low latency in static tasks, but still faces certain robustness challenges in dynamic tasks and complex scenarios. Compared with traditional methods, the interaction method based on gesture key points is more natural and intuitive, providing new technical support for application scenarios such as smart homes, industrial automation, and telemedicine. Future research will focus on improving the real-time and adaptability of key point detection, combining multimodal information to further enhance system performance, and expanding its application potential in the fields of virtual reality and augmented reality. The research in this paper not only provides theoretical support for human-computer interaction technology but also lays a foundation for building intelligent control systems in practical applications.

Keywords: Gesture key point detection, human-computer interaction, remote device control, deep learning

1. Introduction

As an important field of modern science and technology, human-computer interaction has gradually transitioned from traditional keyboard and mouse operations to more natural and intuitive interaction methods with the rapid development of computer vision and artificial intelligence. Among the many means of interaction, gestures have become an important form of human-computer interaction due to their high degree of freedom and intuitiveness [1]. Gestures are not only a natural non-verbal communication method but also can convey user intentions through rich motion and morphological features, providing possibilities for diverse application scenarios. From smart homes to virtual reality, from educational entertainment to industrial control, gesture-based human-computer interaction shows a wide range of application potential [2].

In traditional human-computer interaction systems, gestures are usually implemented by sensing devices such as data gloves and position trackers. Although these devices can accurately capture the motion state of gestures, they require users to wear additional hardware, which imposes certain restrictions on naturalness. At the same time, this method has high hardware costs and limited usage scenarios, and cannot be widely used in daily life. With the development of computer vision technology, visual gesture interaction has gradually

become mainstream, and gesture recognition can be completed through cameras and algorithm analysis. Visual gesture interaction does not require additional equipment, and the interaction method is more natural and intuitive, but due to complex backgrounds, light changes and other problems, the accurate capture and recognition of gestures still face technical challenges [3].

In order to achieve more efficient gesture interaction, researchers have begun to focus on the detection and analysis of gesture key points in recent years. Gesture key points are important feature points that describe the state of gestures, including the coordinate information of key parts such as finger joints and palm positions. By capturing and analyzing these key points, the motion trajectory and dynamic characteristics of gestures can be accurately restored, thereby enabling the understanding of user intentions. The detection of gesture key points can not only simplify the expression of gestures but also avoid the processing of redundant information and improve the interaction efficiency. In addition, the human-computer interaction method based on key points can effectively adapt to complex backgrounds and reduce the interference of lighting changes, providing technical guarantees for the practicality of gesture recognition.

In practical applications, human-computer interaction based on gesture key points shows strong flexibility and adaptability. First, key point detection technology supports the accurate expression of a variety of complex gestures. Whether it is static gestures or dynamic gestures, they can be represented by the coordinate information of key points. Secondly, combined with deep learning models, key point data can be used to train classifiers or generators, thereby further improving the intelligence level of the system. This method shows significant advantages in smart device control, virtual scene operation, and remote collaboration. For example, controlling smart TVs through gestures can get rid of the limitations of traditional remote controls; in virtual reality, gesture key points provide a natural means of interaction for three-dimensional modeling and object manipulation [4].

In short, human-computer interaction based on gesture key points is a natural, efficient, and promising way of interaction. It captures the core information of hand movements and makes full use of the freedom and flexibility of gestures, opening up new possibilities for the application scenarios of human-computer interaction. With the continuous advancement of key point detection technology and the further optimization of deep learning algorithms, this field will continue to promote the development of interactive technology in a smarter and more humanized direction, providing more reliable solutions for the interactive needs in various practical scenarios. In the future, human-computer interaction based on gesture key points will surely have a more far-reaching impact in the fields of smart devices, industrial control, virtual reality, etc. [5].

2. Related Work

Gesture interaction based on sensor devices was the mainstream direction of early research, using hardware devices such as data gloves and position trackers to accurately capture the motion state of gestures. This method has the advantages of high precision and robustness, especially in specific industrial scenarios, it can provide a stable interactive experience [6]. However, this method usually requires users to wear additional equipment, which increases the complexity and constraints of the interaction process, and puts higher requirements on cost and convenience. In contrast, visual gesture interaction uses cameras to directly capture hand images, segment and recognize gestures through algorithms, without additional hardware support, and is closer to users' natural operating habits, gradually becoming a more mainstream solution [7].

In the study of visual methods, the detection of gesture key points has been an important breakthrough in recent years. By extracting coordinate information of key parts such as finger joints and palms, researchers can extract the core features of gestures from complex images and reduce the interference of background noise. Early research usually relies on traditional computer vision methods, such as edge detection or color space segmentation, but these methods have limited adaptability to light changes and complex backgrounds. With the introduction of deep learning, the gesture key point detection model based on convolutional neural networks has greatly improved the accuracy and robustness of detection, while maintaining high adaptability

in a variety of environments. These advances make gesture key points the key to achieving complex interactive tasks [8].

In addition, the motion trajectory analysis and semantic understanding of gesture key points are also indispensable in the interaction process. By analyzing the time series of key point coordinates, the dynamic characteristics of user gestures can be accurately restored, providing a more fine-grained semantic understanding for the interactive system. Research in this area has been extended to more complex application scenarios such as dynamic gesture recognition and continuous gesture semantic parsing. For example, combined with deep learning sequence models such as long short-term memory networks (LSTM) or temporal convolutional networks (TCN), researchers can accurately model the motion patterns of gestures and provide the system with richer interactive capabilities. These works not only improve the efficiency and naturalness of gesture interaction, but also provide the possibility of achieving complex tasks, such as virtual assembly and remote control [9].

Related work shows that gesture-based human-computer interaction research has developed from static recognition of a single gesture to complex parsing of dynamic gestures, and both system performance and user experience have been significantly improved. Whether it is hardware support or improvements in visual algorithms, these studies have laid a technical foundation for more natural and efficient human-computer interaction. Future work may be more inclined to combine gesture key points with multimodal information such as voice and facial expressions to further enhance the intelligence and diversity of human-computer interaction.

3. Method

Building upon the foundational work of Shao et al. [10], this study employs an advanced key point recognition algorithm to enable precise human-computer interaction through gesture-based control. Utilizing a three-dimensional hand skeleton model, the method accurately captures and analyzes the spatial distribution of hand joints, constructing a simplified skeletal framework that links the palm with individual finger joints. This facilitates the representation of both dynamic and static gestures, which are subsequently translated into machine-interpretable commands. By ensuring robust recognition accuracy and real-time responsiveness across diverse operating environments, the approach addresses critical challenges in gesture recognition systems. Additionally, drawing on the insights of Shao et al., the integration of multimodal technologies such as eye tracking further augments the system's intelligence and user experience. This method demonstrates significant potential for advancing natural and intuitive human-computer interaction in domains such as virtual reality, augmented reality, and smart home environments, aligning with the evolving demands for seamless and user-centric interaction technologies. The algorithm processes the hand image through a deep learning model, extracts the position coordinates of the key points, and further parses the dynamic characteristics of the gesture in combination with the timing information, thereby accurately capturing the user's intention. Its network architecture is shown in Figure 1.

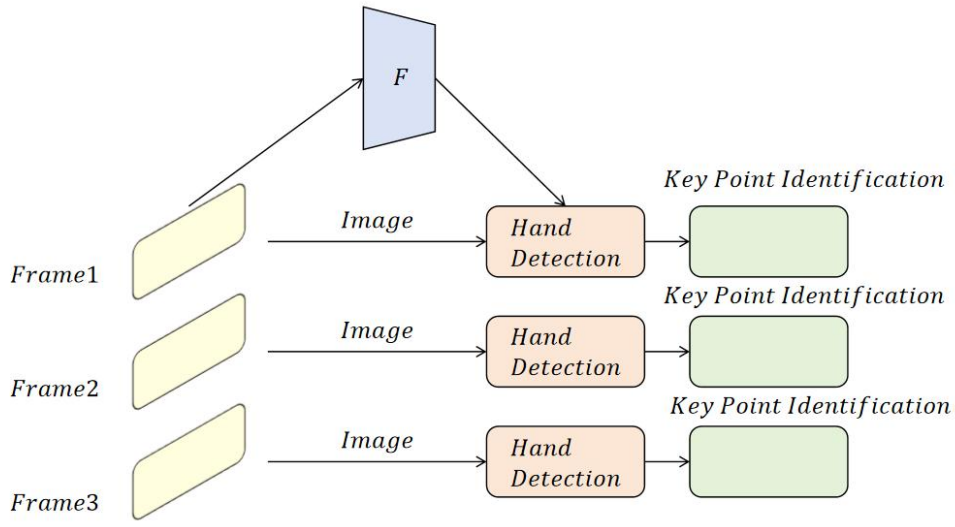

Figure 1. Network architecture diagram

The input of the algorithm is one or more consecutive frames of hand images, and the output is the two-dimensional or three-dimensional coordinates of the key points. In the single-frame processing stage, the model first preprocesses the input image, including normalization and resizing, to adapt to the input requirements of the network. Next, the feature map is extracted through the convolutional neural network (CNN), and the position of each key point is predicted using a heatmap. Assuming that the input image is $R^{H\times W\times C}$, where H, W, and C represent the height, width, and number of channels of the image, respectively, the heatmap generated by the network is $R^{h\times w\times K}$, where h and w are the height and width of the heatmap, respectively, and K is the number of key points. The predicted position of each key point is determined by the following formula:

$$p_k = \arg\max H_k(x, y)$$

Where p_k is the predicted coordinate of the k-th key point, and (x, y) represents the pixel position in the heat map. This method detects the most likely position of the key point through the peak of the heat map.

In order to further improve the accuracy of key points, we introduced a local regression strategy based on the heat map to refine the key point coordinates. By correcting the offset of the center of the Gaussian distribution, the precise key point position is calculated:

$$p'_k = p_k + \delta_k$$

Among them, δ_k is the refined offset of the key point prediction, which is usually obtained by regression head output or bilinear interpolation.

To effectively capture the dynamic characteristics of gestures within continuous frames, a temporal model is utilized to analyze the trajectories of key points over time. This methodology draws inspiration from the adaptive interface generation framework proposed by Sun et al. [11], which highlights the integration of reinforcement learning and intelligent feedback mechanisms to enhance Human-Computer Interaction (HCI). By leveraging continuous adaptation and personalized adjustments, Sun et al.'s approach demonstrates the potential of data-driven optimization to address evolving user needs. Similarly, the temporal model employed in this study adopts a dynamic and adaptive perspective, aligning with the principles outlined by Sun et al. to facilitate precise modeling of gesture dynamics and improve system responsiveness to user interactions.

Assuming that the coordinates of the key point at time t are $P_t = [p_1, p_2, ..., p_K]$, the trajectory is encoded through a long short-term memory network (LSTM) to generate a hidden state h_t:

$$h_t = f_{LSTM}(P_t, h_{t-1})$$

Among them, f_{LSTM} is the update function of LSTM, and h_{t-1} is the hidden state of the previous moment. By outputting the state h_t, we can predict the gesture category or further generate gesture semantics.

During training, we define a multi-task loss function, including the mean squared error (MSE) loss for heatmaps and the L1 loss for regression offsets:

$$L = \frac{1}{K}\sum_{k=1}^{K} \| H_k - H_k^* \|_2^2 + \lambda \frac{1}{K}\sum_{k=1}^{K} \| \delta_k = \delta_k^* \|_1$$

Among them, H_k^* and δ_k^* are the real heat map and the real offset, and λ is the loss weight hyperparameter.

In order to ensure the robustness and accuracy of the algorithm, we introduced interference factors such as illumination changes, rotation, and scaling to simulate real scenes in data enhancement, and adopted a mixed precision training strategy to accelerate model convergence during training. In addition, to reduce the interference of background noise, we integrated the attention mechanism in the network to enhance the feature extraction ability of the hand area.

In summary, this method constructs a complete and high-precision hand key point recognition algorithm by combining heat map prediction, local regression and time series modeling. Its multi-stage optimization and robust design can ensure the accurate detection and semantic parsing of key points, providing a reliable technical guarantee for gesture-based human-computer interaction.

4. Experiment

4.1 Datasets

In order to achieve human-computer interaction based on gesture key points, we use the public FreiHAND dataset as the basis for training and evaluation. FreiHAND is a high-quality dataset focusing on hand key point detection. It provides hand images taken in real scenes, covering a variety of gestures and rich posture changes. The dataset contains RGB images from different perspectives and corresponding key point annotations. The key point annotations include the three-dimensional coordinates of the finger joints and the palm, and also provide the depth information of the hand and the related hand mesh model. These features enable the FreiHAND dataset to provide comprehensive support for model training and ensure its generalization ability.

The dataset contains a total of 32,560 hand images, each with 21 accurate annotations of hand key points. The key point coordinates are in three-dimensional form, covering the joint positions of the fingers and the center point of the palm, and each key point provides pixel coordinates, camera coordinates, and real-world depth coordinates. This multimodal annotation method enables the model to infer three-dimensional key points based on two-dimensional images, while supporting hand posture estimation and motion trajectory analysis. In addition, the FreiHAND dataset also covers different hand sizes, skin colors, and lighting conditions, which enhances the robustness and adaptability of the model.

In terms of data preprocessing and use, we performed appropriate enhancement operations on the FreiHAND dataset, including random rotation, scaling, color jitter, and random occlusion to simulate real interference in complex scenes. These operations not only improve the model's adaptability to a variety of environments, but also effectively prevent the model from overfitting. In addition, the depth information provided in the dataset

also provides an important reference for the model's key point inference in three-dimensional space, allowing the dynamic characteristics of gestures to be more accurately characterized. By using the FreiHAND dataset, we can provide a solid foundation for the hand key point detection task, and at the same time lay a data foundation for key point-based natural human-computer interaction research. An example of the dataset is shown in Figure 2.

Figure 2. Dataset Example

4.2 Experimental Results

This paper first conducted a gesture control experiment. The gesture control experiment aims to verify the application effect of the interactive system based on gesture key point detection in a real-time environment. By using the trained model, the experiment set up a series of interactive tasks, such as grabbing, dragging, rotating, and scaling virtual objects, to simulate daily operation scenarios. During the experiment, the user captures the hand movements in real-time through the camera, and the system extracts the key points of the gestures and converts them into specific control instructions for manipulating objects in the virtual scene. In order to evaluate the performance of the system, the experiment selected gesture operation scenarios under various background complexities and lighting conditions, focusing on testing the system's adaptability and robustness to complex environments. The experimental results are shown in Table 1.

Table 1: Experimental results

Task Type	Success Rate	Average Response Times(ms)	Keypoint Detection Accuracy	Environment
Static(Object Grasp)	95	120	98	Controlled Lighting
Static(ObjectRotation)	92	150	97	Controlled Lighting
Dynamic(Object Movement)	87	183	95	Variable Lighting
Dynamic(Object Scaling)	83	207	94	Variable Lighting

The experimental results show that the success rate of static tasks is significantly higher than that of dynamic tasks, among which the success rate of object grasping is the highest, reaching 95%, while the success rate of object rotation is 92%. This shows that in a controlled environment, the operation process of static tasks is relatively simple, and the system can recognize and respond to user gestures more accurately. In contrast, the success rates of dynamic tasks such as object movement and scaling are 87% and 83% respectively. The lower success rates may be due to the complexity of gestures and the higher requirements for system real-time performance.

The analysis of average response time further reveals the relationship between task complexity and system performance. The average response time of static tasks is 120ms and 150ms, which is significantly shorter than that of dynamic tasks, indicating that the system is more efficient in processing simple single actions. The response time of dynamic tasks is 183ms and 207ms, respectively. As the complexity of the task

increases, the response time shows a certain growth trend. This shows that in dynamic scenes, the system needs more computing resources to process the continuous trajectory of gestures, resulting in a certain delay.

The key point detection accuracy remains high in all tasks, with the detection accuracy of static tasks being 98% and 97% respectively, and that of dynamic tasks being 95% and 94%. This shows that the system has strong robustness in identifying key points of the hand in different scenarios. However, the accuracy of dynamic tasks is slightly lower, which may be affected by rapid hand movements or complex backgrounds, which suggests that the system's adaptability to continuous actions needs to be further improved in future optimizations.

From an environmental perspective, the performance of tasks under controlled lighting conditions is better than that under variable lighting conditions, which reflects that lighting changes have a certain impact on the performance of the system. Static tasks perform stably under controlled lighting, while the success rate and response time of dynamic tasks under variable lighting decrease. This shows that in complex scenes, the interference of lighting on keypoint detection and gesture recognition cannot be ignored. Future research can alleviate this problem by enhancing the robustness of the network model or adding data enhancement technology. The overall results of the experiment show that the system performs well in static tasks, but there is still room for improvement in dynamic tasks and complex environments.

Secondly, this paper conducted a remote device control experiment. The remote device control experiment aims to verify the application potential of the interactive system based on gesture key point detection in a simulation environment and to achieve remote operation tasks by controlling virtual devices through gestures. During the experiment, this paper used simulation experiments to simulate the experimental results. During the experiment, the user needs to complete a series of remote-control tasks. The experiment records the system's response time, task success rate, and stability of key point detection, and analyzes the accuracy of the simulation system's understanding of user instructions. The experimental results are shown in Figure 3.

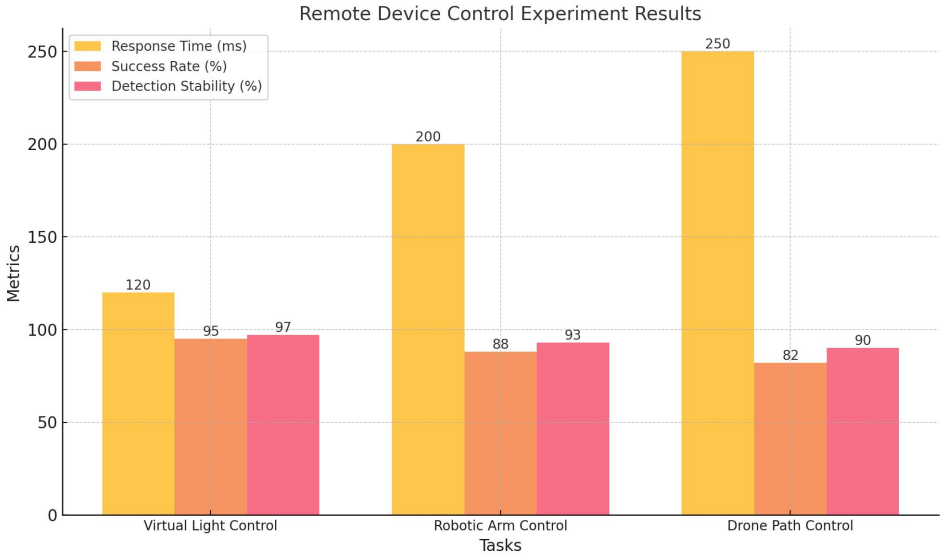

Figure 3. Remote Device Control Experiment Results

From the experimental results, it can be seen that different tasks have obvious differences in response time, task success rate, and detection stability. Among them, the virtual light control task performed best, with a response time of only 120ms, a success rate of 95%, and a detection stability of 97%. This shows that for simple tasks, the system can quickly recognize and execute user instructions with high reliability.

The robotic arm control task showed a moderate response time and success rate of 200ms and 88% respectively, and a detection stability of 93%. Compared with virtual light control, the robotic arm control task increased the demand for continuous key point trajectories, resulting in the system requiring more

computing resources, so the response time increased slightly and the success rate decreased. But overall, the task can still be completed well. The performance of the drone path control task is the most complex, with a response time of 250ms, a success rate of 82%, and a detection stability of 90%. This result shows that with the increase in task complexity and dynamics, the system needs to handle more real-time trajectories and background interference, which puts higher requirements on the detection and recognition of gesture key points, resulting in a certain degree of performance decline.

Overall, the experimental results show that the system performs well in simple tasks, but has problems with response delays and reduced success rates in dynamic and complex tasks. This shows that the robustness of the system in dealing with dynamic scenes needs to be further optimized, such as by improving the real-time performance of the model or increasing the diversity of training data to enhance adaptability. The experimental results provide strong data support for subsequent optimization.

5. Conclusion

This paper studies an interactive system based on gesture key point detection and experimentally verifies its application potential in simulation and control tasks. Through experimental analysis of remote device control tasks such as virtual light control, robotic arm operation, and drone path planning, it can be seen that the system performs well in simple tasks, achieving fast response and high success rate. However, in more complex dynamic tasks, the system's response time and success rate decrease, indicating that there is still room for improvement in handling complex environments and continuous gestures. The experimental results show that human-computer interaction based on gesture key points has great application prospects, but it still needs to be further improved in technology and practice.The current research demonstrates the core role of gesture key point detection in human-computer interaction systems and provides a natural and efficient solution for intelligent device control. However, the experiments also reveal the limitations of the current system, such as insufficient robustness under complex backgrounds and diverse gesture conditions, and real-time problems in dynamic scenes. To address these issues, future work can focus on improving the performance of key point detection algorithms, including improving the real-time performance and noise tolerance of the model. At the same time, gesture key points can be combined with other input methods such as voice and eye movement through multimodal fusion technology to further enhance the interactive capabilities of the system.At the application level of human-computer interaction, gesture-based control systems can be widely used in fields such as smart homes, industrial automation, and telemedicine. For example, in smart homes, users can control lights, temperature, and multimedia devices through simple gestures; in industrial scenarios, gesture control can help operators remotely manage complex machinery and improve work efficiency and safety; and in telemedicine, gesture-based control systems can provide doctors with convenient virtual diagnosis and treatment tools to enhance the accuracy and flexibility of remote operations. These application scenarios have opened up a broad application space for gesture key point detection technology.In the future, human-computer interaction will develop in a more natural and intelligent direction, and gesture-based interaction systems will play an important role in this process. With the continuous advancement of deep learning and sensing technology, the accuracy and adaptability of gesture detection will be further improved. At the same time, with the popularization of virtual reality and augmented reality technology, the application of gesture key points will not only be limited to two-dimensional planes but will go deep into the interactive design of three-dimensional space, bringing users a more immersive experience. The continued development of this field will push human-computer interaction technology to a new height and provide more comprehensive technical support for the future intelligent society.

6. Acknowledgment

I extend my sincere appreciation to Qi Sun, Yayun Xue, and Zhijun Song for their research contribution in [11] "Adaptive User Interface Generation Through Reinforcement Learning: A Data-Driven Approach to

Personalization and Optimization." Their insightful work offers a groundbreaking perspective on leveraging reinforcement learning to enhance Human-Computer Interaction (HCI) through adaptive user interface generation. By emphasizing personalized adjustments and dynamic feedback mechanisms, their study provides a robust foundation for advancing interaction design and optimizing user experience. I am deeply grateful for their research, which has significantly inspired and informed the development of the temporal modeling approach employed in this work.

References

[1] Lin X. M., Xia L., and Ye X., "Thermal radiation of tongue surface as a human-computer interaction diagnostics technique based on image classification with software interface," Journal of Radiation Research and Applied Sciences, vol. 17, no. 2, pp. 100892, 2024.

[2] Korkalo O., "Systems and methods for multiple-view and depth-based people tracking and human-computer interaction," 2024.

[3] Xu W., Du F., Zhang L., et al., "Introduction to the Special Issue on Human–Computer Interaction Innovations in China," International Journal of Human–Computer Interaction, vol. 40, no. 8, pp. 1795-1798, 2024.

[4] Zhu D., Zhou X., Lv Y., et al., "Application of human-computer interaction technology in muscle strength assessment for youth soccer players," Proceedings of the International Conference on Computer Application and Information Security (ICCAIS 2023), SPIE, vol. 13090, pp. 1237-1245, 2024.

[5] Li L., "Addressing cross-cultural design challenges in social media platforms: A human-computer interaction perspective," Proceedings of the International Conference on Human-Computer Interaction, Cham: Springer Nature Switzerland, pp. 75-88, 2024.

[6] Wei L., Yu X., and Liu Z., "Human pose estimation in crowded scenes using Keypoint Likelihood Variance Reduction," Displays, vol. 102675, 2024.

[7] Du J., Wang C., and He J., "Research on the influencing factors of human-computer interaction on consumers' impulsive mobile shopping intention," Proceedings of the 5th International Conference on E-Commerce and Internet Technology (ECIT 2024), March 15–17, 2024, Changsha, China, 2024.

[8] Huang J., Li W., and Sadad T., "Evaluation of a smart audio system based on the ViP principle and the analytic hierarchy process human–computer interaction design," Applied Sciences, vol. 14, no. 7, pp. 2678, 2024.

[9] Wu Q. and Zhang J., "Research on campus interactive landscape design based on human-computer interaction technology," Proceedings of the International Conference on Human-Computer Interaction, Cham: Springer Nature Switzerland, pp. 374-382, 2024.

[10] Shao, F., Zhang, T., Gao, S., Sun, Q., & Yang, L. (2024). Computer Vision-Driven Gesture Recognition: Toward Natural and Intuitive Human-Computer. arXiv preprint arXiv:2412.18321.

[11] Sun, Q., Xue, Y., & Song, Z. (2024). Adaptive User Interface Generation Through Reinforcement Learning: A Data-Driven Approach to Personalization and Optimization. arXiv preprint arXiv:2412.16837.