Optimization of VR Human-computer Game Interaction Based on Improved PIFPAF Algorithm and Binocular Vision

Hong Zhu[1] and Bo Chen[2,*]
[1] School of Experimental Art, Hubei Institute of Fine Arts, Wuhan, China
[2] The School of Arts, Hubei University of Education, Wuhan, China
* Corresponding author: Bo Chen (chenbo1565@163.com)

Submitted: 12 Feb 2025 **Accepted:** 7 Apr 2025 **Published:** 26 Jul 2025

Abstract To make virtual reality human-computer games more accurate and provide users with an immersive gaming experience, the study combines the method of improved part intensity field and part association field (PIFPAF) with binocular vision to optimize the interaction of VR human-computer games. The experimental results indicated that the PIFPAF algorithms performed relatively well with number of errors and target keypoint correlation of 0.22 and 0.97, respectively. In terms of processing speed, the algorithm performed faster in both 640×480 and 320×240 resolutions, with 13 fps and 19 fps, respectively. Among the five predefined gestures, the "pointing" gesture was recognized correctly the largest number of times in 30 test sessions, with 29 successful identifications. In contrast, the "clenched fist" gesture had the fewest correct recognitions, totaling 26. The success of the suggested approach is confirmed by the experimental findings, which show that the optimized human-computer interaction system has high accuracy and processing speed. This study offers a fresh approach to the advancement of human-computer interaction technology and encourages technological integration innovation in the realm of virtual reality human-computer gaming.

Keywords: virtual reality; PIFPAF algorithm; binocular stereo vision; keypoint detection algorithm; dimensioning algorithm.

1. Introduction

As virtual reality (VR) technology advances quickly, it has progressively made its way into a variety of industries, including gaming, education, healthcare, design, and more, as a new interactive experience. The naturalness and intuitiveness of human-computer interaction (HCI) are also the key factors to enhance user experience [5, 7]. Traditional VR interaction methods, such as joysticks and keyboards, although satisfy users' needs to a certain extent, have certain limitations in simulating real-world interaction. It limits the user's immersion and interaction experience in the VR environment [17]. Although some deep learning-based stereo matching methods have made progress, they still face challenges such as high computational complexity, large hardware requirements, and poor adaptability to dynamic scenes in real-time applications. Optimizing the network structure, introducing lightweight models, utilizing parallel computing, and adaptive feature extraction techniques can improve their efficiency and real-time performance.

The part intensity field and part association field (PIFPAF) algorithm, originally proposed by Kreiss et al. [9], aims to solve the keypoint association problem in multi-person pose estimation. This algorithm can more accurately detect human body structure and associate keypoints by predicting local keypoints and spatial relationships between keypoints, especially suitable for complex interactive scenes. The development of PIFPAF is inspired by earlier work such as OpenPose, proposed by Cao et al. [3], which pioneered bottom-up keypoint detection in multi-person pose estimation.

In current HCI technology, traditional methods have many key problems in VR interaction scenarios, including chaotic keypoint matching during multi-person interaction, which leads to a decrease in tracking accuracy. In the case of occlusion, the loss of keypoint information is severe, which affects the accuracy of recognition. The high computational complexity affects real-time interaction performance. PIFPAF enhances local feature extraction by deep neural networks and optimizes the skeleton keypoint matching strategy, enabling it to infer the pose of the occluded part well even in occluded environments while maintaining computational efficiency. These features make it an ideal choice for optimizing VR HCI systems, enhancing immersive experiences and real-time performance, and promoting the development of intelligent interaction systems. Binocular vision can provide depth information to more accurately localize the user's body parts in complex environments [1, 4]. Therefore, to improve the naturalness and accuracy of VR human-computer game interaction (HCGI), this study investigates VR HCGI based on the improved PIFPAF algorithm with binocular vision.

The innovativeness of the research lies in the improvement of the existing PIFPAF algorithm in order to increase the accuracy and real-time performance of human pose estimation. It also combines binocular vision technology with the improved PIFPAF algorithm, thus proposing a new depth-aware interaction. The contribution of this research is to improve the accuracy of keypoint detection and the robustness of limb association through this algorithm. Its branch accurately locates keypoints using Gaussian heat maps and establishes human skeleton connections using vector fields, which is more resistant to occlusion compared to traditional methods. In addition, PIFPAF optimized the feature extraction network, adopted an efficient ResNet backbone network, and used adaptive scale inference to improve the ability to detect different human postures while reducing computational redundancy. These enhancements significantly improve real-time performance and enable efficient and accurate pose estimation in complex interactive scenarios, resulting in a smoother and more intuitive interactive experience.

The research will be carried out in four sections. The Section 2 is a review of the current research status of binocular vision and VR HCI. The Section 3 is the optimization study of human keypoint detection and HCI system. The Section 4 contains the experimental analysis of the research algorithms and system performance. In the Section 5 the results of experiments with the methodology proposed in this paper are discussed. The last Section 6 is a summary of the research.

2. Related works

With the advancement of computer graphics and HCI technology, VR technology has gradually matured, providing users with unprecedented immersive experiences. Optimization of HCI experience is also a hot research topic at present. Lyu aimed to explore the current state of HCI in the metaverse, and research results showed that key technologies such as 5G, blockchain, and HCI supported the development of the metaverse. In the future, humanized somatosensory connections in HCI could become a trend [14]. Ramadoss proposed an optimized non-invasive human-machine interaction model to improve the accuracy of human motion recognition in HCI. The research results showed that this method had significant effects on human motion and target recognition, reducing noise by 7.2% and improving accuracy to 97.2% [15]. Li proposed an interaction design model that combined artificial intelligence (AI) and voice information to enhance the HCI experience in VR environments. Research showed that this model promoted the application and development of VR technology in multiple fields such as gaming, fitness, and education by optimizing the HCI design [11].

Keypoint detection algorithms and stereo binocular vision can effectively detect and track human posture and motion. Read proposed a research method that comprehensively examined the use of binocular vision and stereoscopic vision in order to explore the advantages and disadvantages of binocular vision and its mechanisms. The research results indicated that although binocular vision reduced the overall field of view, it enhanced obstacle avoidance and contrast sensitivity [16]. Bonnen et al. proposed a research method that combined eye and body tracking to explore the role of binocular vision in complex terrain walking. The research results indicated that binocular vision was crucial for locating a foothold, and its absence could systematically affect gaze strategies, increasing perceptual uncertainty and making the gaze more inclined towards a nearby foothold [2]. Lin et al. proposed a recognition method based on improved ResNet and skeleton keypoints to improve the accuracy of single image human action recognition, and constructed a multi task network. The research results showed that this method could accurately recognize human movements under different human motion, background light, and occlusion conditions. Compared with the original network and main recognition algorithms, it had an advantage in accuracy and balances network parameters, solving the problems of large network and slow operation [12]. Zhang proposed a method that combined efficient network structure, training strategy, and post-processing techniques to address the challenge of human keypoint detection in a single image. The research results indicated that this method effectively improved the detection accuracy and outperforms the latest technology on the benchmark of keypoint detection [20]. To improve the accuracy and practicality of the fall detection system, Inturi proposed a new visual based fall detection scheme. The research results indicated that the system could effectively detect five types of falls and six types of daily activities, and performed well on the UP-FALL dataset [6].

VR and HCI technology have made significant progress in recent years. They generate highly realistic 3D virtual environments through computers, allowing users to interact with the virtual world in an immersive way. They are widely used in various fields such as gaming, education, healthcare, and design. In the VR field, major manufacturers have introduced several innovative devices. In 2023, Sony released the PlayStation VR2, which featured internal and external tracking, eye tracking, a high-definition display, and a controller with adaptive triggering and haptic feedback to enhance the gaming experience. In 2024, Apple released the Apple Vision Pro, a fully enclosed mixed reality headset that emphasizes video perspective functionality. Although it lacks the external controller of traditional VR headsets, it is described as a spatial computer. In terms of HCI, with the advancement of technologies such as computer graphics and AI, HCI is gradually shifting from traditional keyboard- and mouse-based interaction modes to more natural and intelligent interaction methods. For example, interaction methods based on gesture recognition, speech recognition, eye tracking, and other technologies are gradually emerging. The integration of eye tracking technology enables the system to optimize rendering based on the user's point of view, improving performance and immersion. In addition, the development of hand tracking and gesture recognition technology allows users to interact with virtual environments in a more natural way, reducing reliance on traditional controllers. Together, these advances are driving the evolution of VR and HCI technologies, providing users with a more intuitive and immersive experience.

To summarize, many scholars have researched on HCI technology. Moreover, there are more studies on the acquisition of information about human motion and posture using binocular vision technology or human keypoint detection algorithm, and certain results have been achieved. However, most of the scholars only use a single algorithm model and do not improve the model's deficiencies. Most researchers focus on action recognition, trajectory prediction, and interactive feedback when researching VR interactive technology. However, these studies have certain limitations when faced with complex actions and multi-person interaction scenarios. First, current mainstream methods often rely on deep learning-based motion capture or pose estimation algorithms, such as convolutional neural networks, recurrent neural networks, and their variants, when dealing with complex actions. Although these methods can achieve good recognition results in simple single-person interaction scenarios, there are bottlenecks in the recognition and prediction of complex actions, mainly due to the difficulty of the model to accurately capture high-speed and nonlinear motion trajectories, especially when multiple joints are involved in coordinated motion. Existing methods have weak temporal modeling capabilities and are difficult to accurately predict subsequent actions. Furthermore, in multi-person interaction scenarios, traditional methods typically use data fusion based on visual or inertial sensors to analyze user interaction behavior. However, these methods are susceptible to data noise and environmental disturbances when faced with multiple occlusions, dynamic background changes, or synchronized interactions. This can result in interaction delays, increased error rates in action recognition, and ultimately reducing the immersion and real-time feedback effects of the game. On the other hand, traditional optimization algorithms typically rely on rule-based or reinforcement learning frameworks, such as Markov decision processes and reinforcement learning, when dealing with path planning and action generation problems in VR HCI. However, these methods have a high computational overhead in high-dimensional state spaces, making it difficult to respond to the complex action needs of users in real time. In addition, traditional reinforcement learning models are unable to efficiently model the dynamic interaction relationships between different users in multiuser collaborative interactions, making it difficult for the system to adapt to changing interaction patterns. It can be concluded that in response to the complexity and real-time requirements of VR HCGI scenarios, the existing research has the limitation of balancing computational efficiency, interaction accuracy, and real-time response capability, which has become a key challenge to further enhance the VR interaction experience.

For the above reasons, in this research the PIFPAF algorithm is improved by combining it with the enhanced binocular vision technology to locate the user in 3D, so as to optimize the VR HCGI system.

3. Optimization of VR interpersonal game interaction

3.1. Keypoint dimensional enhancement algorithm based on improved binocular vision technique

VR technology can use computer-generated 3D images and sounds to simulate the real sensory experience of humans [10, 22]. However, in VR human-computer games, users' hand movements and body movements have a high degree of complexity and diversity [8]. The keypoint detection algorithm can improve the accuracy of human motion detection in VR games by accurately locating human joints. These algorithms can capture player poses in real time, reducing errors caused by occlusion or complex movements, making interactions smoother. Combined with deep learning models, keypoint detection can optimize limb tracking, enabling the system to more accurately understand the player's intent. It can also improve the accuracy of physical feedback, improve action matching, avoid delays, enhance immersion, and provide strong technical support for motion interaction and prediction in VR games. Based on this, the study adopts the human body keypoint detection algorithm for keypoint positioning of human body images. When designing keypoint detection algorithms to accurately capture various complex user actions in VR environments, the following key factors need to be considered. Firstly, the algorithm should have high robustness to cope with challenges such as occlusion, lighting changes, and complex backgrounds. Secondly, it is necessary to ensure real-time

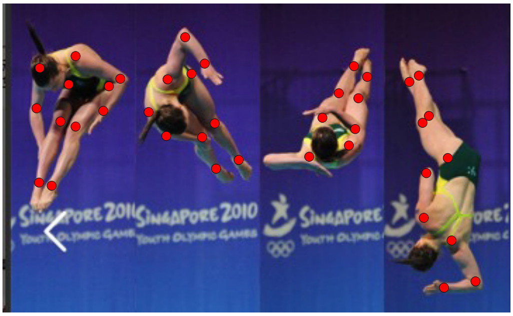


Fig. 1. Image of human keypoint positioning

performance to meet the low latency requirements of VR interaction. In addition, the algorithm should be able to adapt to users of different body types and action patterns, and have good generalization ability. Finally, it is necessary to optimize computational efficiency to achieve efficient operation with limited hardware resources.

Figure 1 illustrates the precise keypoint positioning. In this Figure, the keypoint detection algorithm for human keypoint positioning is mainly distributed in the joints of face and limb joints and torso. Facial keypoints are mainly used in VR applications for facial expression recognition, user authentication, and immersive interactive experiences. By accurately capturing facial movements, real-time facial expression mapping of virtual characters can be achieved, enhancing the authenticity of social interactions. In addition, facial keypoints can optimize voice synchronization and improve character performance. In security, they can be used for identity recognition, ensuring personalized settings and data security. For immersive control, the combination of eye tracking can provide a more natural way of visual interaction, improving the responsiveness and user experience of VR systems [13, 19]. In dynamic VR environments, facial keypoint detection faces challenges such as high real-time requirements, insufficient robustness in complex scenes, and limited computing resources. To address these obstacles, the following strategies can be adopted: firstly, optimize the algorithm structure and introduce lightweight CNN to reduce computational complexity and improve processing speed; Secondly, by combining multimodal information such as depth information and optical flow information, the robustness of facial keypoint detection is enhanced; Finally, parallel computing technology is utilized to further enhance the real-time performance of the algorithm. In addition, an adaptive feature extraction method attention mechanism is adopted to dynamically focus on key facial regions, improving detection accuracy. In order to effectively improve the accuracy and real-time performance of facial keypoint detection in dynamic VR environments with limited computing resources.

The data interaction function of the VR system is shown in Figure 2. In this Figure,

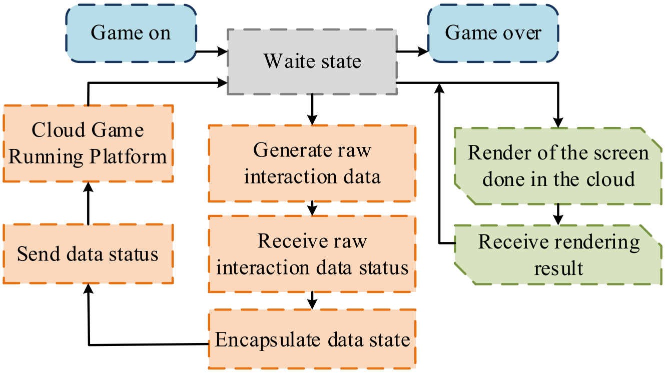

Game on
Waite state
Game over
Cloud Game Running Platform
Generate raw interaction data
Render of the screen done in the cloud
Send data status
Receive raw interaction data status
Receive rendering result
Encapsulate data state

Fig. 2. Interactive activity diagram of the interaction controller

the interactive controller has five states, starting from the initial waiting state. After the game starts, it enters the state of receiving raw interactive data and recording player operations. Then it encapsulates the data and sends the data status, processes and uploads the interactive data to the cloud. Next, it enters the state of receiving rendering results and receives rendered images from the cloud. After the game ends, the controller returns to the waiting state and prepares for the next interaction [5]. In VR games, interactive controllers must manage different states, including idle, active, interactive, and feedback, to ensure a smooth user experience. Real-time state switching determines response speed, such as the accuracy of gesture recognition, physical collision detection, and environmental feedback. Accurate state management can reduce latency, improve immersion, optimize the allocation of computing resources, and prevent lag. The combination of intelligent predictive algorithms and adaptive control strategies can enhance real-time interaction capabilities, making player interaction in virtual environments more natural and fluid. Moreover, its combination of intelligent prediction algorithms and adaptive control strategies can enhance real-time interaction capabilities, making players' operations in virtual environments more natural and smooth.

The real-time state switching of interactive controllers is extremely important for the accuracy of gesture recognition and physical collision detection. It reduces the delay between user actions and system feedback, improving the real-time response. In addition, dynamic computing resource allocation optimizes processing efficiency, prioritizing critical interaction tasks such as gesture recognition or collision detection. Real time state switching also enhances the naturalness of interaction, allowing the system to smoothly transition between different interaction modes based on user intent, improving the user experience. It also optimizes error handling, allowing the system to quickly adjust strategies to address recognition errors and reduce the negative impact on the experience. Ultimately, real-time state switching enables the controller to adapt

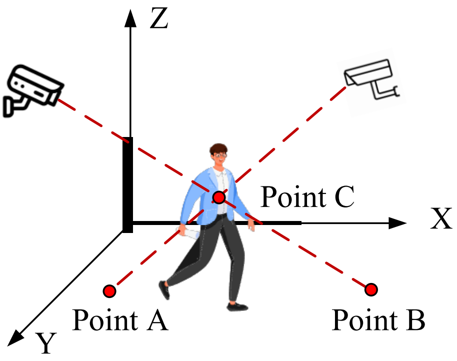


Fig. 3. Binocular positioning principle.

to complex scenarios, handle concurrent operations, ensure the accuracy and timeliness of interactive operations, and significantly improve the interaction quality of VR systems. However, in this study a universal 2D human keypoint definition methodwas used. This method lacks depth information and is difficult to accurately recover human posture, especially in occluded or complex motion scenes. Second, changes in perspective can cause keypoint positions to shift, which affects the stability of posture estimation. In addition, 2D methods are difficult to capture the 3D rotation information of human joints, which limits the accuracy of VR interaction. Therefore, research is needed to increase the dimensionality of human keypoint localization, combined with 3D keypoint detection or deep learning models, to improve the accuracy of human pose recognition in VR environments.

The ability to perceive depth, as well as the position of objects in three dimensions, is crucial for HCI in VR. This is achieved through the use of binocular vision, which enables the calculation of disparity, thereby enhancing both immersion and spatial perception capabilities. In comparison with monocular vision, binocular vision has been demonstrated to facilitate more precise distance measurement. In contrast to technologies that rely on LiDAR or depth cameras, binocular vision offers several advantages, including cost efficiency, broader applicability, and enhanced performance under variable lighting conditions, thereby mitigating recognition failure. The principle of the method is shown in Figure 3. The method uses dual lenses to detect the point simultaneously. Binocular vision provides depth information for HCI in VR by simulating the stereoscopic imaging mechanism of the human eye, significantly enhancing immersion and spatial perception. It achieves three-dimensional spatial reconstruction through disparity calculation, optimizing users' spatial positioning and interactive experience in virtual environments. Binocular vision can capture user posture and gestures in real time, and achieve natural and smooth interaction with the help of keypoint detection technology, especially performing well in complex motion and occlusion scenes. In addition, binocular vision supports seamless integration of virtual and real environments, enhancing the interactive effects of augmented reality scenes. Its depth information can also optimize rendering performance by dynamically adjusting rendering resources, improving visual effects, and reducing computational resource waste. Binocular vision has strong adaptability and can work stably in different lighting and complex backgrounds, expanding the application scope of VR technology. However, the method is not adapted to more complex game scenes and is affected by light. The keypoint dimension enhancement algorithm for improving binocular vision technology can alleviate the problem of human occlusion in VR scenes by integrating deep learning with traditional visual geometry modeling methods. Its theoretical basis mainly comes from core technologies such as stereo matching, multi-view geometry, keypoint extraction, and dimension enhancement mapping. First, the algorithm relies on the disparity information of binocular vision by constructing the epipolar geometric relationship between the left and right cameras and combining it with a deep learning-based keypoint detection network to achieve accurate extraction of human joint points. Compared to monocular vision methods, binocular systems provide richer depth information, allowing the estimation of 3D positions based on unobstructed perspectives even when certain areas are obstructed. In addition, the keypoint dimensionality enhancement algorithm effectively completes missing keypoints caused by occlusion by high-dimensional mapping of low-dimensional 2D keypoint information, combined with spatiotemporal constraints and data-driven optimization strategies, such as Transformer based sequence modeling methods, and improves global consistency. The advantage of this method is that even if some joint points are occluded, the system can still infer their reasonable positions based on known joint topology relationships, thus reducing interaction errors caused by occlusion. Therefore, in this study the binocular stereo vision methodwill be improved. The the 2D pixel position by coordinate transformationwill be uplifted. Firstly, the calculation of converting the world coordinate system (CS) to the camera CS is shown in Equation (1).

$$\begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} = W \otimes \begin{bmatrix} X_E \\ Y_E \\ Z_E \end{bmatrix} + T\,, \tag{1}$$

where (X_E, Y_E, Z_E) is the world CS, (X_C, Y_C, Z_C) is the camera CS, W is the rotation matrix, and T is the translation vector. Then the camera CS is converted to the image CS. The specific calculation is shown in Equation (2).

$$Z_C \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \otimes \begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix}\,, \tag{2}$$

where (x, y, z) is the position of the same point in 2D image coordinates, a is the camera focal length, and Z_C is the depth coordinate. The conversion from image CS to pixel CS is performed. The specific calculation is shown in Equation (3).

$$\begin{bmatrix} m \\ n \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\mathrm{d}x} & 0 & m_0 \\ 0 & \frac{1}{\mathrm{d}y} & n_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{3}$$

where $\begin{bmatrix} m \\ n \\ 1 \end{bmatrix}$ is the transformed chi-square coordinates, $\frac{1}{\mathrm{d}x}$ and $\frac{1}{\mathrm{d}y}$ are the scaling factors, and m_0 and n_0 are the translations. In conclusion, it is possible to determine the transformation relationship between the global CS and the pixel CS. Equation (4) illustrates this particular computation.

$$Z_C \begin{bmatrix} m \\ n \\ 1 \end{bmatrix} = \lambda \otimes \begin{bmatrix} W & T \\ \vec{0} & 1 \end{bmatrix} \otimes \begin{bmatrix} X_E \\ Y_E \\ Z_E \\ 1 \end{bmatrix}, \tag{4}$$

where λ is the camera internal reference matrix. Camera calibration is critical for accurate 3D reconstruction, as it can eliminate lens distortion and provide internal and external parameters of the camera to improve reconstruction accuracy. The key steps include: image acquisition using a calibration board to obtain multi-angle images, feature point detection to extract corner or marker points, parameter estimation to compute internal parameters (focal length and principal points) and external parameters (position and rotation), aberration correction to correct for lens aberrations, and optimization and tuning to use nonlinear optimization to improve calibration accuracy. These steps ensure the accuracy of the camera model during the 3D reconstruction process and enhance the authenticity of spatial point cloud data. Among them, the internal reference calibration is calculated by the classical Zhang calibration method [21], which can find the distortion coefficient of the camera. The specific calculation is shown in Equation (5).

$$\mathrm{dist} = [\theta_1, \theta_2, \theta_3, \varphi_1, \varphi_2], \tag{5}$$

where θ is the radial distortion coefficient, and φ is the tangential aberration coefficient. Camera external parameter calibration can be carried out by changing the camera position and updating the position and attitude of the camera in the world CS. The specific flow of the keypoint dimensional enhancement algorithm is shown in Figure 4. In this Figure it can be seen that the keypoint dimensional enhancement algorithm consists of two parts: determining the internal and external parameters of the camera and increasing the dimension calculation of the data. Among them, the camera calibration stage

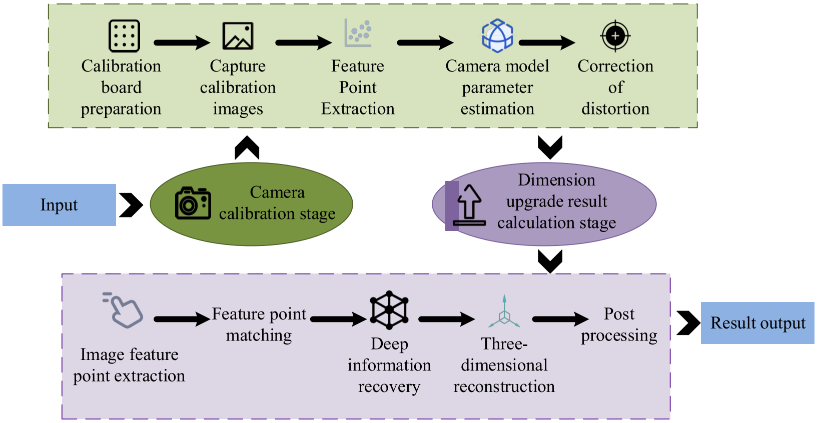

Calibration board preparation
Capture calibration images
Feature Point Extraction
Camera model parameter estimation
Correction of distortion
Input
Camera calibration stage
Dimension upgrade result calculation stage
Image feature point extraction
Feature point matching
Deep information recovery
Three-dimensional reconstruction
Post processing
Result output

Fig. 4. The overall flow chart of the dimensional enhancement algorithm.

is the preparatory stage of the algorithm, which needs to be performed only once. The stage of calculating the result of increasing dimension needs to extract the feature points in the 2D image to be processed and match them with the feature points of the known 3D structure. The geometric structure of the 3D scene is reconstructed based on the position of each feature point in the 3D space and the recovered depth information. In addition to this, the reconstructed 3D model is optimized and smoothed to improve the accuracy and visual effect. Finally, the upscaled results such as depth map are output. The dimensionality enhancement algorithm for feature point extraction and depth recovery can effectively improve 3D scene reconstruction. First, key feature points can be extracted by deep learning or traditional methods to improve matching accuracy. Then, binocular disparity estimation or deep neural network can be combined to recover depth information. Next, dimensionality boosting algorithms are used to optimize point cloud distribution, enhance geometric details of sparse regions, and improve reconstruction accuracy. Finally, by integrating multi-perspective information and correcting errors, the 3D model becomes more accurate and coherent, resulting in higher quality virtual environment reconstruction. The advantages of the keypoint dimensionality enhancement algorithm based on improved binocular vision technology in terms of speed and accuracy are mainly reflected in efficient stereo matching, optimized depth estimation, and keypoint reconstruction strategies. Compared with traditional methods, this algorithm improves the efficiency of stereo matching by introducing an adaptive disparity optimization strategy and a multi-scale feature fusion mechanism, making depth computation more stable and reliable, while reducing computational overhead and improving real-time performance. In addition, this method combines sparse point cloud completion technology in the keypoint reconstruction process, resulting in higher human keypoint reconstruction accuracy, especially in complex interactive scenes, which can provide more accurate motion capture. The theoretical basis for dealing with human occlusion in VR scenes lies in the disparity redundancy and depth compensation properties of binocular vision. Specifically, in the binocular imaging process, different camera angles can provide redundant information, allowing partially occluded keypoints to be inferred from unobstructed angles. This process alleviates the problem of keypoint loss caused by occlusion in monocular methods. Furthermore, the algorithm constructs spatial topological constraints based on graph neural networks, thereby enabling mutual constraints between detected keypoints and inferring the position information of partially occluded areas. This enables a more complete reconstruction of human body structure. Compared with traditional methods, this improved algorithm can handle human keypoint detection in complex scenes more stably. Even in occlusion situations, it can improve the prediction accuracy of keypoints through multi-view feature compensation and spatial relationship inference. At the same time, combined with optimized computation processes, it reduces computational complexity, making it faster and more accurate in interactive VR scenes.

3.2. Optimization study of PIFPAF algorithm

The study adopts an improved binocular vision technique for human posture keypoint positioning in VR human-computer games. However, to realize user action recognition and animation simulation in game interaction systems, the study needs to further improve the applicability and detection effect of the algorithm in different game scenarios. PIFPAF is an advanced human pose estimation method, which is especially suitable for multi-person pose detection in low-resolution and crowded scenes [18]. Its network structure is shown in Figure 5.

The key to the PIFPAF algorithm in human pose estimation lies in the synergistic effect of the two branches, part intensity field (PIF) and part association field (PAF). The PIF branch is mainly used to detect the location information of human keypoints, improve the accuracy of keypoint detection by predicting the density distribution of each joint, and combine Gaussian distribution to enhance local features, so that the model can accurately locate keypoints even when dealing with complex backgrounds and occlusion situations. As a high-precision positioning mechanism for each keypoint, it not only regresses the continuous spatial coordinates of keypoints, but also effectively enhances the robustness of the model to occlusion, attitude distortion, and low resolution keypoints through the collaborative modeling of heatmaps and displacement vectors. Especially, in human-computer interaction scenarios such as VR and virtual reality, the PIF branch can provide more accurate responses to local human features with higher density pixel level supervision, thereby significantly improving the system's perception ability. Gaussian distribution is used in the keypoint regression process to model the position distribution of each predicted keypoint. By generating a two-dimensional Gaussian heatmap centered on the keypoint on the feature map, accurate weighting of local

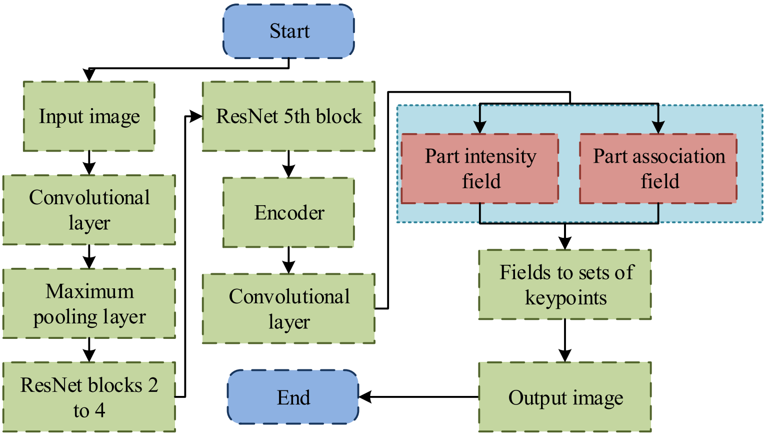


Fig. 5. Structural diagram of the PIPAF network

areas is achieved, thereby improving the accuracy of keypoint localization. In complex environments such as occlusion, lighting changes, or multi person interactions, Gaussian distribution can highlight the saliency of key areas, effectively suppress background interference, and enable the model to extract keypoint information more stably and accurately. This mechanism significantly enhances the robustness and detection performance of the PIFPAF model under high noise conditions. The PAF branch is responsible for learning the correlation information between different joints in the human body, using vector fields to represent the topological relationships between different joints, thereby maintaining structural consistency in multi-person interaction scenarios and effectively reducing the keypoint confusion problem. The combination of the two branches enables PIFPAF to achieve higher robustness in posture estimation. The PIF branch ensures accurate detection of keypoints, while the PAF branch ensures the rationality of the human body structure, especially in challenging scenarios such as occlusion, complex movements, and multi-person interaction. PAF can effectively utilize joint connection relationships for posture correction.

In addition, compared to traditional regression-based methods, PIFPAF's end-to-end optimization strategy allows the network to globally optimize posture estimation in terms of the entire structure, achieving a better balance between speed and accuracy. The network first receives raw image data and inputs it into the PIFPAF model, extracts features through convolutional layers, and downsamples at the max pooling layer. The encoder consists of multiple residual blocks to process features in depth. Then, the two branches separately generate keypoint field predictions. Finally, the decoder converts the feature map into a set of keypoints. Further research is conducted to optimize the

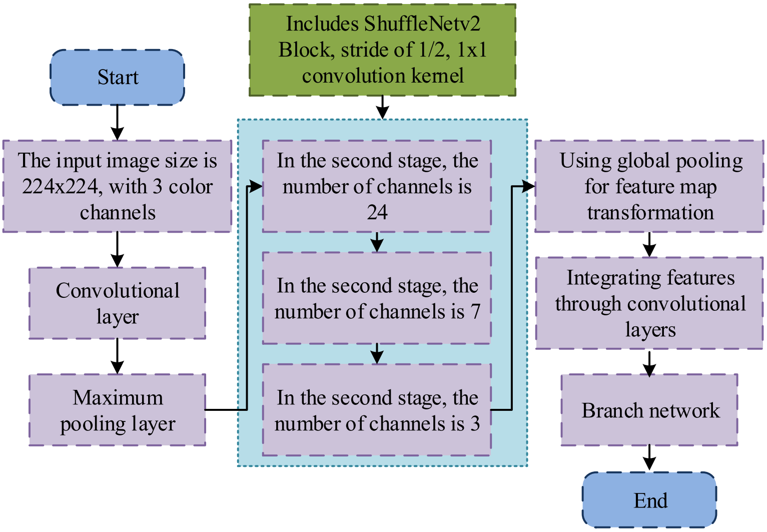

Start
The input image size is 224x224, with 3 color channels
Convolutional layer
Maximum pooling layer
Includes ShuffleNetv2 Block, stride of 1/2, 1x1 convolution kernel
In the second stage, the number of channels is 24
In the second stage, the number of channels is 7
In the second stage, the number of channels is 3
Using global pooling for feature map transformation
Integrating features through convolutional layers
Branch network
End

Fig. 6. Network structure in the improved PIFPAF feature extraction stage.

ResNet Block feature extraction network structure in the algorithm, in order to propose an improved PIFPAF algorithm. Its structure is shown in Figure 6.

ResNet has a large number of parameters, making training difficult. As shown in Figure 6, the convolutional layer is responsible for extracting local features of the image, which is crucial for identifying keypoints as it can capture key visual patterns in the image. Residual blocks alleviate the gradient vanishing problem in deep network training by introducing shortcut connections, allowing the network to train deeper layers more effectively and extract richer feature representations. These deep level features are particularly important for precise keypoint detection in complex environments, as they provide more contextual information and details. In addition, replacing ResNet Block with ShuffleNetv2 Block is to improve processing speed while maintaining accuracy. The design of ShuffleNetv2 Block is more lightweight and suitable for real-time applications, which is crucial for fast response and smooth interaction experience in VR environments. Feature extraction is performed on the original data after the replacement network, and the results are fed into the two-branch network for regression. The two-branch network's PIF branch is utilized to locate the important human body components and forecast each one's size, position, and confidence. Its output parameter set is calculated as shown in Equation (6).

$$P^{ij} = \{p_a^{ij}, p_x^{ij}, p_y^{ij}, p_o^{ij}, p_\tau^{ij}\}, \tag{6}$$

where i and j are the coordinates of the network output, p_a is the confidence map of the pixel, p_o^{ij} is the correction parameter for computing the loss function, p_τ^{ij} is the Gaussian smoothing parameter, and p_x^{ij} and p_y^{ij} are the components of the offset vectors in the x and y directions of the keypoints closest to the pixel, respectively. The computation of the Gaussian function is specifically shown in Equation (7).

$$G(x,y) = \frac{1}{2\pi\tau^2} e^{-\frac{x^2+y^2}{2\tau^2}} , \tag{7}$$

where τ is the 2D form of the Gaussian function. Its bandwidth is positively correlated with the influence range of the function. Based on the calculation of the parameters and the function, the prediction results of the keypoint location can be obtained. Its calculation is specified in Equation (8).

$$F(x,y) = \sum_{ij} p_a^{ij} G(x,y|p_x^{ij}, p_y^{ij}, p_\tau^{ij}) , \tag{8}$$

where $F(x,y)$ is the keypoint position prediction function. PAF, on the other hand, is used to connect the detected body parts through the association information to form a complete human posture. Its output parameter set is calculated as shown in Equation (9).

$$A^{ij} = \{a_a^{ij}, a_{x1}^{ij}, a_{y1}^{ij}, a_{o1}^{ij}, a_{x2}^{ij}, a_{y2}^{ij}, a_{o2}^{ij}\} , \tag{9}$$

where a_{x1}^{ij}, a_{y1}^{ij}, a_{x2}^{ij}, and a_{y2}^{ij} are the components of the offset vector on the horizontal axis x and vertical axis y, respectively, and a_{o1}^{ij} and a_{o2}^{ij} are correction functions. The result of the output of the network structure includes three types of outputs, and the loss values of the three types of outputs are calculated and summed to obtain the total output of the network. The specific calculation is shown in Equation (10).

$$\text{LOSSES} = \text{BCELoss} + \text{SCALELoss} + \text{REGLoss} , \tag{10}$$

where BCELoss is the confidence correlation output, REGLoss is the offset vector correlation output, and SCALELoss is the target scale related output.

In this way, in this study an optimized HCGI system is constructed. The local game interaction system and the cloud game running platform make up the majority of the system. Among them, the operation process of the game interaction system is as follows. First, the images are captured by two GB cameras to obtain the raw images of the current frame. Then, the AI performs algorithm calculations such as keypoint detection, keypoint uplift, gesture recognition, etc. on the captured images to generate the composed raw data. Then the data generated by the AI module is encapsulated to form a JSON file and sent to the cloud via SOCKET communication. The operation process of the cloud game platform first requires preliminary data processing, including data reception, parsing, and operation. According to the processed data, the game is rendered, i.e. the processed data is used to generate JPG images. Then the rendered image data is sent back to the local machine via SOCKET communication, and the rendering effect is played in the local display module.

4. Performance analysis of optimized VR HCGI system

4.1. Experimental environment and data sources

In the experiment an improved binocular vision technology's keypoint dimensionality enhancement algorithm is used to evaluate the ability of the system to handle human occlusion and interactive performance in VR scenes. The AR game HCI experimental verification between HCI system and cloud game platform can be conducted. The software development environment for the experiment is Windows 10 operating system, PyCharm Community development tool, and PyTorch GPU deep learning environment. The hardware environment is RTX2060 6 G GPU and 16 G memory. In addition, the experimental environment also includes a high-performance GPU computing platform, and uses the Unity 3D engine and OptiTrack optical motion capture system to build a high-precision interactive VR test environment. The data acquisition of the experiment adopts a binocular stereo camera array to capture keypoint information under different occlusion conditions, and optimizes keypoint dimensionality and pose estimation through deep learning networks. The experimental setup includes several scenarios such as single person, multiple people, partial occlusion, and heavy occlusion to test the adaptability and robustness of the algorithm in different complex environments. Specific evaluation metrics include spatial accuracy indicators such as keypoint prediction accuracy, mean joint error, and posture estimation accuracy. The inference speed and computational complexity of the algorithm are measured simultaneously to evaluate its real-time performance. In addition, the experiment uses trajectory smoothness and latency indicators to verify the smoothness of the interaction, ensuring that the algorithm remains efficient and stable in complex interaction processes. The specific experimental scene is shown in Figure 7.

The layout and environmental characteristics of the experimental scenes have a significant impact on the performance of keypoint dimensionality enhancement algorithms in binocular stereo vision technology. As shown in Figure 7, the experiment is conducted in a specially designed VR interactive space with uniform and controllable lighting conditions to reduce the impact of lighting changes on depth estimation. The lighting equipment adopts a multi-angle light source arrangement at the top and side to ensure sufficient illumination in different directions while avoiding strong shadows or overexposure, thereby improving the image quality obtained by the binocular camera. The experimental space needs to ensure that participants have sufficient activity space to simulate real-world VR application scenarios. In the experimental environment, some obstacles

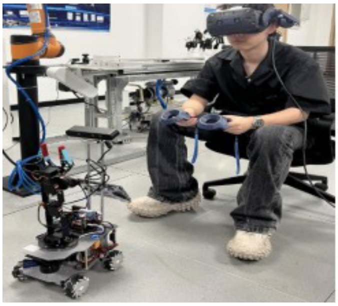


Fig. 7. Example of the experimental scene – a specially designed VR interactive space.

such as tables and chairs, simulated walls, or virtual interactive devices are appropriately placed to test the performance of the algorithm in complex occlusion situations. The presence of these obstacles can obscure keypoints of the human body, increasing the difficulty of inferring depth information. In addition, the scene may contain dynamically moving objects, such as other test subjects or virtual interactive elements, which may affect the stability of the stereo matching algorithms. By introducing different types of occlusion, such as partial occlusion and global occlusion, the adaptability of the algorithm in environments of varying complexity are evaluated. Environmental factors have a significant impact on the experimental results. First, lighting conditions can affect the quality of binocular matching. Too dark or high contrast environments can lead to errors in disparity calculation, thereby reducing the accuracy of keypoint dimensionality enhancement. Second, the arrangement of obstacles affects the occlusion pattern. If the occlusion is large or has strong reflective properties, it may introduce additional depth estimation noise. In addition, the background texture characteristics of experimental scenes can also affect the robustness of binocular matching. In complex backgrounds, false matches may increase. Therefore, it is necessary to optimize the background appropriately, such as using low-texture backgrounds to reduce interference. Finally, it is also necessary to consider the installation position and angle of the camera to ensure that the obtained binocular disparity information can fully cover important parts of the human body while avoiding depth distortion caused by viewing angle deviation.

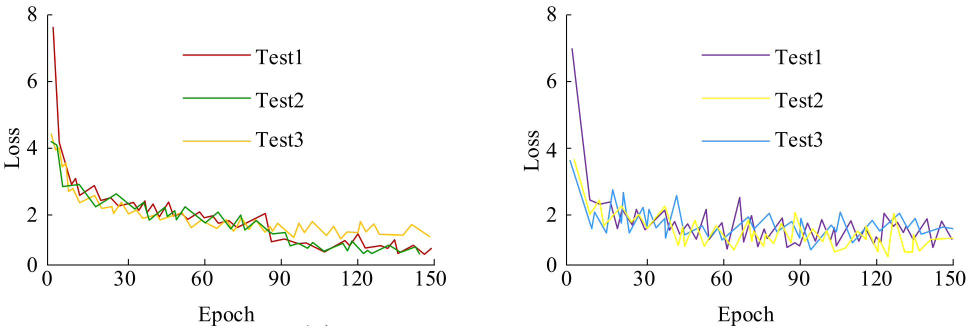
The experiment faces several challenges and limitations during implementation and testing. First, human occlusion is complex and highly unpredictable, especially in multi-person interactions, where the uncertainty of the occlusion region affects the accuracy of keypoint dimensionality enhancement. Second, binocular stereo vision relies on high-quality image matching, but depth estimation can be subject to errors under changing lighting, dynamic backgrounds, or reflections. In addition, improving the algorithm has a high computational complexity, and optimizing computational efficiency while ensuring real-time performance has become a key issue. During the experimental process, it is necessary to balance the accuracy of data annotation with the size of the data, ensuring that the occlusion data used for training is sufficiently rich to improve the generalization ability of the model. Hardware limitations are also a factor. Although high-performance GPUs are used for inference, the computational cost of the model still needs to be controlled to avoid delays that affect the VR interactive experience. Finally, due to the involvement of multiple device synchronizations in VR interaction, such as motion capture systems, VR headsets, and binocular cameras, time synchronization errors can affect the overall experimental results, requiring additional calibration steps to improve system consistency.

A total of 100 participants were recruited for the study, including 50 males and 50 females. The age distribution of the selected subjects includes children, adolescents, middle-aged, and elderly groups. Body types include lean, normal, and overweight. Moreover, all the participants are without any motor dysfunction.

4.2. Performance analysis of the keypoint dimensional enhancement algorithm with improved PIFPAF algorithm

To investigate the effect of different training strategies of the upscaled human keypoint detection algorithm on the loss function of the dataset, the study uses Basenet and Headsnet to train the dataset, respectively. Basene uses pre-trained model initialization and fine tuning on multi-scale feature maps. During the training process, random data augmentation is used to improve generalization ability, while the Adam optimizer is used to dynamically adjust the learning rate to avoid gradient oscillations. Headsnet uses a multi-task loss function combined with keypoint heatmaps and depth information monitoring to improve its ability to recover occluded areas. A total of 120 rounds of experiments were conducted. There were three groups of experiments. Test 1 and Test 3 were all trainded with Basene and Headsnet, respectively. Test 2 contained 50 rounds of each of the two types of training. The loss function variation curves obtained from the experiments are shown in Figure 8.

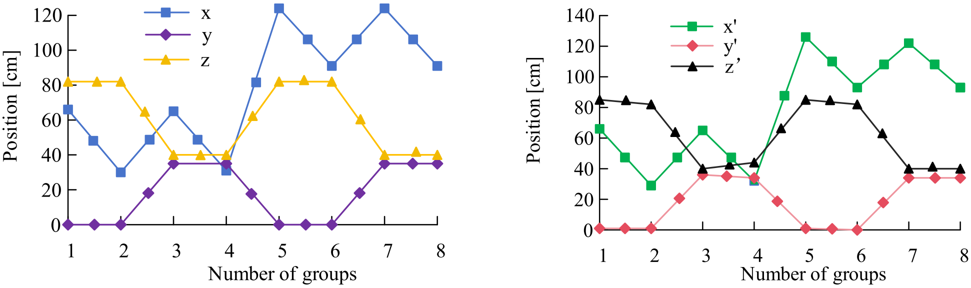
In Figure 8a, the loss functions of all three groups on the training set decrease with the number of training rounds, and decrease rapidly at the beginning and then stabilize. Among them, the starting value of the loss function of Test 1 is much higher than that of the other two groups, which is 8. The starting values of Test 2 and Test 3 are 4.2 and 4.3, respectively. The loss functions of Test 2 and Test 3 have a close trend in the early stage. However, in the later stage when Test 2 and Test 1 are stabilized, the loss function changes are closer to each other and both of them are roughly stabilized at about 1.5. Test 3 stabilizes with a slightly higher loss value, fluctuating within a range around 2. In Figure 8b, the loss function value of the three experimental training sets on the validation set decreases with the increase of training rounds. It decreases drastically in a short period of time, after which the change decreases. However, the volatility is relatively large, and all of them fluctuate in the range of 0.5 to 2.5. This may be



a Training set loss curve.

b Verification set loss curve.

Fig. 8. Loss function of the experiment on the dataset.



a World coordinates of real points.

b World coordinates of predicted points.

Fig. 9. Test results of the dimensional enhancement algorithm on the target position.

due to the diversity of data in the validation set or the difference in the generalization ability of the model on different data. It can be observed that Test 2 has more rounds in the validation set in which the loss function value achieves the minimum value.The experimental results show that the experimental group Test 2, which combines two network training strategies, has the best training effect. To further verify the feasibility of this keypoint dimensional enhancement algorithm, the experiment is to localize the target object by this algorithm and compare the error between the predicted and actual position. The world coordinates of the actual point are (x, y, z) and the world coordinates of the predicted point are (x', y', z'). A total of 8 experiments are conducted and the specific results are shown in Figure 9.

In Figure 9a, the position change range of the target object on the x-axis is large and fluctuates in the range of $[30, 120]$ cm. The position change curves of y-axis and z-axis are almost symmetrical to each other, and their change ranges are $[0, 35]$ cm and

Tab. 1. Results of performance comparison of different algorithms.

Algorithm name	Improved PIFPAF algorithm	OpenPose	ResNet50 + YoloV3
NE [%]	0.22	0.25	0.19
OKS	0.97	0.96	0.98
640×480 [fps]	13	4.2	0.75
320×240 [fps]	19	15	0.80
Extendibility	High	Middle	Middle
CPU utilization ratio [%]	15.4	23.1	30.5

$[40, 82]$ cm, respectively. In Figure 9b, the variation ranges of the target object in x-axis, y-axis and z-axis are $[29, 126]$, $[1, 36]$, $[40, 85]$ cm. Comparing the coordinate change curves of the target in Figure 9a and 9b, it can be observed that the coordinates of the predicted object position and the actual position using the keypoint dimensional enhancement algorithm are very similar to each other, and the total average absolute error is 2.11 cm. The experimental findings demonstrate that the keypoint dimensional enhancement method is capable of precisely capturing the target's shifting location in space. It has good robustness, and can adapt to different environments and changes in conditions. To investigate the performance of the improved PIFPAF algorithm, OpenPose, and ResNet50 + YoloV3 algorithms are compared. Two metrics, number of errors (NE) and object keypoint similarity (OKS) are calculated. Moreover, the speed of the algorithms is compared for different resolution images. OpenPose is a multi-stage CNN-based pose estimation algorithm that uses a bottom-up approach to detect human keypoints and analyzes limb structure through keypoint correlation. It is suitable for multi-person pose estimation. ResNet50 + YoloV3 combines deep residual networks with object detection algorithms, using ResNet50 to extract human features and YoloV3 for efficient object detection and localization, ensuring the accuracy and real-time performance of keypoint detection. The performance comparison between the two in VR HCI scenarios can help analyze the accuracy and speed advantages of keypoint detection. Table 1 displays the individual experimental outcomes.

In Table 1, the ResNet50 + YoloV3 algorithm performs best on the NE and OKS metrics with 0.19% and 0.98, respectively, with the lowest error rate and the highest keypoint similarity. OpenPose has the worst performance on both metrics, which may be related to the bottom-up approach adopted by OpenPose. The improved PIFPAF algorithm, on the other hand, performs in the middle, with NE and OKS of 0.22 and 0.97, respectively. However, in terms of processing speed, the improved PIFPAF algorithm performs faster in both 640×480 and 320×240 resolutions, with 13 fps and 19 fps, respectively. OpenPose's processing speed at 640×480 resolution is somewhere in between at 4.2 fps. However, the processing speed at 320×240 resolution is 15 fps, which is not much different from the improved PIFPAF algorithm. The processing speed of ResNet50 + YoloV3 is significantly lower than the other two algorithms, which may be

Tab. 2. Statistics of gesture recognition efficiency in 30 tests for each gesture.

Gesture	Gesture 1 *open palm*	Gesture 2 *clench fist*	Gesture 3 *thumbs up*	Gesture 4 *victory symbol*	Gesture 5 *pointing*
Correct identification in 1st run	28	26	29	27	28
Correct identification in 2nd run	1	3	0	1	1
Correct identification in 3rd run	0	1	1	1	1
Correct identification in 4th run	1	0	0	1	0
Correct identification in 5th run	0	0	0	0	0

due to the fact that the algorithm has sacrificed some of its speed in order to obtain higher accuracy. Due to the improvement of the algorithm structure and the use of parallel processing techniques, the testing findings demonstrate that the revised PIFPAF algorithm greatly boosts processing speed while maintaining higher accuracy.

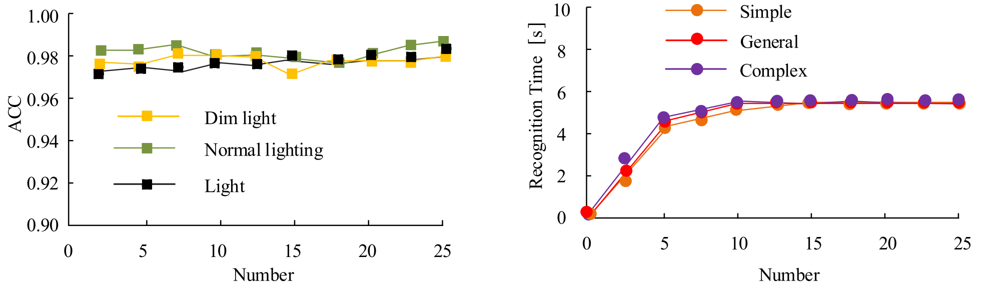
4.3. Performance analysis of optimized VR HCGI system

To further verify the recognition accuracy of the VR human-computer gaming system using the improved PIFPAF algorithm and binocular vision optimization for the actual user actions, in the experiment five static gestures. Moreover, 30 sets of tests were conducted to recognize the specified gestures in the interaction actions.

In order to standardize the evaluation of gesture recognition accuracy in interactive systems, five commonly used static gestures were defined and assigned identification numbers. Gesture 1 is *open palm*, Gesture 2 is *clench fist*, Gesture 3 is *thumbs up*, Gesture 4 is *victory symbol*, and Gesture 5 is *pointing*. These gestures were selected due to their clear and recognizable visual features, and were repeatedly tested throughout the entire recognition experiment to maintain consistent gesture identifiers.

The total number of times each gesture was correctly recognized in the repeated test is shown in Table 2. It can be seen that almost all of the five gestures were recognized by the interactive system in one recognition run. Among them, Gesture 3 is recognized in one recognition the largest number of times: 29, and Gesture 2 is recognized in one recognition the smallest number of times: 26. Moreover, almost most of the gestures are fully recognized in the first three runs. The results of the experiment demonstrate that the interaction system can meet the needs of the player by accurately identifying the user's gesture movements while they are playing.

The experiment further investigates the recognition of the optimized interactive system under different light or environment complexity conditions. Figure 10 displays the specific outcomes. The graphs in this Figure show the trend of the recognition accuracy of the system with the number of tests under different lighting environments. Figure 10b shows the variation of the recognition time with the number of tests under different environmental complexities. The recognition accuracy under the three lighting conditions



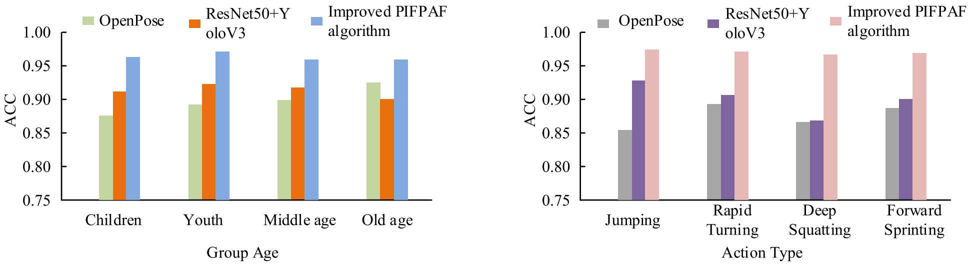
a The recognition accuracy of the system in different lighting environments.

b The recognition time of the system in environments of different levels of complexity.

Fig. 10. System identification performance analysis under different environmental conditions.

shown in Figure 10a is relatively close, with small fluctuations around 0.98. This indicates that the system can maintain stable recognition performance under different lighting conditions. This may be due to the system's strong lighting robustness in the feature extraction and recognition algorithms, which can effectively adapt to different lighting environments. As shown in Figure 10b, there is a slight difference in the recognition time among the three environments in the initial stage. When the number of recognition is 5, the recognition times for simple, normal, and complex environments are approximately 4.1 s, 4.2 s, and 4.3 s, respectively. As the number of tests increases, the detection time of the three environments gradually stabilizes around 5.2 s. It can be concluded that the complexity of the environment has a relatively small effect on the recognition time of the system, and the system can achieve consistent and stable processing efficiency in environments of different complexity after adapting to the environment.

Further experiments are conducted to compare the accuracy of posture recognition among different user groups and dynamic scenarios. Among them, the experiment selects four movement postures for recognition: jumping, fast turning, deep squatting, and forward sprinting. The results are shown in Figure 11. The graphs in Figures 11a and 11b show the comparison of pose recognition accuracy for different age groups and motion poses of each algorithm. In Figure 11a, the improved PIFPAF algorithm performs best in all age groups, with an accuracy rate higher than 0.95. OpenPose performs poorly in all age groups, especially in children and middle-aged populations, with accuracy rates below 0.90. ResNet50 + YOLOv3 performs well in young and middle-aged populations. This may be due to the large deviation between the body types of children and the elderly and the standard dataset, which affects the accuracy of the model's keypoint detection. As shown in Figure 11b, the improved PIFPAF algorithm performs best in all motion types, with an accuracy rate around 0.97. The recognition accuracy range of OpenPose is $[0.85, 0.89]$. The performance of ResNet50 + YOLOv3 is in between, with a maximum recognition accuracy of about 0.93 for children. The unstable recognition accuracy of



a The system's posture recognition accuracy for users of different ages.

b The recognition accuracy of different algorithms for different dynamic scenes.

Fig. 11. Comparison of recognition effects of various algorithms on different groups and motion postures.

Tab. 3. Statistical analysis of keypoint detection algorithm performance.

Index	OpenPose	ResNet50 + YoloV3	Improved PIFPAF algorithm	Standard deviation	p
Key-point detection accuracy [%]	85.2	89.5	94.3	3.2	$p < 0.05$
Interaction response time [ms]	120.4	98.7	85.6	8.5	$p < 0.01$
Block recovery accuracy [%]	72.8	81.2	92.5	4.1	$p < 0.01$
False detection rate [%]	14.6	10.2	5.8	2.8	$p < 0.05$

OpenPose and ResNet50 + YOLOv3 may be attributed to the fact that activities such as jumping and rapid turning result in brief losses of body keypoints. In contrast, activities like deep squatting and sprinting forward cause significant displacement, making it challenging for single frame-based posture recognition algorithms to track reliably. The improved PIFPAF optimizes keypoint matching through binocular vision, maintaining high detection accuracy even under various intense movements.

To verify the feasibility of the proposed method, the experiment further conducts a significance test on the optimized VR HCGI system. The specific results obtained are shown in Table 3. According to this Table, the optimized VR human-machine game interaction system performs better in several key indicators. Among them, the accuracy of keypoint detection reached 94.3%, which is significantly improved compared to OpenPose's 85.2% and ResNet50 + YoloV3's 89.5%, with $p < 0.05$. This improvement is mainly due to the keypoint dimension enhancement algorithm of binocular vision technology, which effectively enhances the ability to capture spatial information and improves the accuracy of keypoint recovery under occlusion. In terms of interaction response time, the average processing time of the improved algorithm is only 85.6 ms, which is significantly optimized compared to the other two algorithms, with a $p < 0.01$. This optimization is due to the lightweight design of the algorithm structure, which

Tab. 4. Results of key point detection and 3D attitude estimation accuracy.

	Index	Experimental group 1	Experimental group 2	Control group 1	Control group 2
Key point detection accuracy	Average error of pixel standard	1.2	1.5	2	2.5
	Match success rate [%]	95.2	93.7	89.5	88.7
3D pose estimation accuracy	Average joint error [cm]	1.5	1.8	2.2	2.8
	Pose estimation accuracy [%]	94.3	92.5	88.7	85.6

makes the inference process more efficient and reduces the computational overhead. In terms of occlusion restoration accuracy, the improved algorithm reaches 92.5%, $p < 0.01$, The false detection rate of the proposed method is reduced to 5.8% ($p < 0.05$), further verifying the robustness of the improved algorithm.To further verify the effectiveness of epipolar geometry in improving the accuracy of keypoint detection in VR systems based on binocular vision, as well as the influence of coordinate transformation on the accuracy of 3D pose estimation, two experimental groups were set up: high calibration accuracy + epipolar geometry and high calibration accuracy + epipolar geometry, and two control groups: low calibration accuracy + epipolar geometry and low calibration accuracy + epipolar geometry for comparative experiments.

The results obtained are shown in Table 4. In terms of keypoint detection accuracy, the average pixel error of experimental group 1 is 1.2, and the matching success rate is 95.2%, both of which are better than the average error of 1.5 and the success rate of 93.7% in experimental group 2. The performance of the control group was poor, with an average error of 2 and 2.5 for control group 1 and control group 2, respectively, and a success rate of 89.5% and 88.7%, respectively. In terms of 3D pose estimation accuracy, the average joint error of experimental group 1 is 1.5 cm, and the pose estimation accuracy is 94.3%, which is also better than experimental group 2. The control group had larger errors of 2.2 cm and 2.8 cm, respectively, and lower accuracy rates of 88.7% and 85.6%. In summary, the experimental group outperformed the control group in keypoint detection and 3D pose estimation, indicating that high-precision camera calibration and coordinate transformation methods can significantly improve the accuracy of 3D pose estimation, reduce average joint error, and improve the accuracy of pose estimation. Experimental group 1 performed the best, indicating that the algorithm using epipolar geometry constraints significantly outperformed the algorithm without epipolar geometry in terms of keypoint detection accuracy and 3D pose estimation accuracy.

5. Discussion

The experimental results showed that the optimized VR HCGI system outperformed traditional methods in terms of keypoint detection accuracy, interaction response speed, and occlusion recovery ability, thus improving the real-time interaction experience in virtual environments. This optimization rendered VR devices more adaptable to complex action recognition, multi-user interaction, and occlusion environments, and it could be widely applied in immersive gaming, remote collaboration, rehabilitation training, and other fields. For example, in sports VR games, the system must accurately recognize large movements such as running and jumping to provide real feedback. In rehabilitation training, optimizing action recognition for different ages and physical conditions ensures safety and effectiveness. It provided a new solution for the development of VR interaction technology and laid the foundation for optimizing future intelligent HCI systems. This advantage was mainly due to the application of binocular vision technology combined with the keypoint dimensionality enhancement algorithm, which could more accurately restore occluded keypoints and improve the stability of detection. In terms of keypoint detection accuracy, experimental results indicated that the improved algorithm achieved 94.3%, which was significantly improved compared to OpenPose (85.2%) and ResNet50 + YoloV3 (89.5%). This advantage was mainly due to the deep information fusion of binocular vision, which allowed the system to exploit multi-view features and reduce the error of monocular methods in occluded scenes. In addition, the keypoint dimensionality enhancement algorithm enhanced local features and optimized globally, making keypoint localization more accurate. In terms of interaction response time, the optimized algorithm had an average processing time of 85.6 ms, which was nearly 30% less than OpenPose's 120.4 ms. This improvement was mainly due to the improved network structure, which used a lightweight CNN for feature extraction and reduced computational complexity by optimizing feature matching strategies, making inference faster. Compared to traditional deep learning methods, this algorithm was more suitable for real-time interactive applications and improved the user experience. In terms of occlusion restoration accuracy, the improved algorithm achieved 92.5%, a 20% improvement over OpenPose's 72.8%. This improvement was due to the introduction of binocular depth estimation, which allowed the system to make reasonable inferences based on the spatial information of other keypoints even when some keypoints were occluded. The accuracy was higher compared to methods based on monocular RGB images. In addition, by combining the Transformer structure for global feature modeling, the system could infer missing parts from the full pose distribution, further improving robustness. Compared with other studies, some existing research used long short-term memory networks or gated recurrent units for temporal modeling. However, their computational complexity was large and difficult to meet real-time interaction requirements. The improved model used in this study achieved a better balance between computational complexity and accuracy, and was suitable for efficient HCI systems in VR scenes.

In the process of VR interaction, synchronizing facial keypoint data with voice input to improve character performance and realism faces many challenges. Firstly, the synchronization of data collection is a crucial issue. The capture of facial keypoints relies on visual sensors, while voice input relies on audio devices, and there are differences in sampling rate and processing speed between the two, resulting in difficulties in aligning data on the timeline. Secondly, the real-time requirements are extremely high. The VR environment requires a low latency interactive experience, and the synchronization processing of facial expressions and speech needs to be completed in a very short time, which puts extremely high demands on the efficiency of algorithms and hardware performance. In addition, robustness in complex scenarios is also a challenge. In environments with multiple interactions or noisy backgrounds, the accuracy of facial keypoint detection and speech recognition can be affected, which in turn affects the synchronization effect. Finally, the difference in personalized expression is also a problem. There are significant differences in facial expressions and voice tones among different users. How to preserve these personalized features during synchronization while achieving natural and smooth interaction is a direction that needs further research in the future.

6. Conclusion

To capture and analyze the user's gesture in real time to ensure real-time performance in VR environment so as to provide a smoother and intuitive interaction experience, in this study the PIFPAF algorithm was improved. It was also combined with binocular vision technology to optimize the VR HCGI operation. The experimental results indicated that the loss functions of all three tested groups decreased with the increase of training rounds and then stabilized. Among them, Test 2 and Test 1 were closer to each other in terms of the variation of the loss function as they stabilized on the training set, both roughly stabilizing around 1.5. Test 3 had slightly higher loss values as it stabilized, fluctuating in the range around 2. The loss function values of the three sets of experimental training on the validation set fluctuated in the range of 0.5 to 2.5 in the later stages, and Test 2 had the highest number of rounds in which the loss function value achieved the minimum in the validation set. The overall results of the performance verification of the keypoint dimensional enhancement algorithm revealed that before and after using this algorithm, the predicted object positions were very similar to the coordinates of the actual positions, and the total average absolute error was 2.11 cm. The experimental results indicated that the experimental group combining the two network training strategies had the best training effect, and the keypoint dimensional enhancement algorithm could accurately capture the moving position of the target in space with good feasibility. The study demonstrated that the proposed method exhibited favorable applicability and accuracy in gaming scenarios.

However, the research is still limited by experimental conditions in specific environments, and the robustness under complex lighting and extreme occlusion conditions still needs to be improved. Therefore, future research will optimize the generalization ability of the model and combine it with deep learning to improve interaction accuracy. This can improve the real-time and accuracy of VR interaction systems and provide new ideas for the development of intelligent HCI technology.

Authors' declarations

Conflict of interest

The authors have no conflict of interest to report.

Data availability

The information on data are included in the manuscript.

References

[1] M. Akram, S. Siddique, and M. G. Alharbi. Clustering algorithm with strength of connectedness for m-polar fuzzy network models. *Mathematical Biosciences and Engineering* 19(1):420–455, 2022. doi:10.3934/mbe.2022021.

[2] K. Bonnen, J. S. Matthis, A. Gibaldi, M. S. Banks, D. M. Levi, et al. Binocular vision and the control of foot placement during walking in natural terrain. *Scientific reports* 11(1):20881, 2021. doi:10.1038/s41598-021-99846-0.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43(1):172–186, 2021. doi:10.1109/TPAMI.2019.2929257.

[4] B. M. S. Hasan and A. M. Abdulazeez. A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining* 2(1):20–30, 2021. doi:10.30880/jscdm.2021.02.01.003.

[5] K. Head-Marsden, J. Flick, C. J. Ciccarino, and P. Narang. Quantum information and algorithms for correlated quantum matter. *Chemical Reviews* 121(5):3061–3120, 2021. doi:10.1021/acs.chemrev.0c00620.

[6] A. R. Inturi, V. M. Manikandan, and V. Garrapally. A novel vision-based fall detection scheme using keypoints of human skeleton with long short-term memory network. *Arabian Journal for Science and Engineering* 48(2):1143–1155, 2023. doi:10.1007/s13369-022-06684-x.

[7] J. Katona. A review of human–computer interaction and virtual reality research fields in cognitive infocommunications. *Applied Sciences* 11(6):2646, 2021. doi:10.3390/app11062646.

[8] T. Kosch, R. Welsch, L. Chuang, and A. Schmidt. The placebo effect of artificial intelligence in human–computer interaction. *ACM Transactions on Computer-Human Interaction* 29(6):1–32, 2023. doi:10.1145/3529225.

[9] S. Kreiss, L. Bertoni, and A. Alahi. PifPaf: Composite fields for human pose estimation. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11969–11978. IEEE Computer Society, 2019. doi:10.1109/CVPR.2019.01225.

[10] Z. Q. Lan, J. X. Wang, and L. Q. Wang. Multi-view line matching based on multi-view stereo vision and leiden graph clustering. *Journal of Geo-Information Science* 26(7):1629–1645, 2024. doi:10.12082/dqxxkx.2024.240080.

[11] K. Li and X. Li. AI driven human–computer interaction design framework of virtual environment based on comprehensive semantic data analysis with feature extraction. *International Journal of Speech Technology* 25(4):863–877, 2022. doi:10.1007/s10772-021-09954-5.

[12] Y. Lin, W. Chi, W. Sun, S. Liu, and D. Fan. Human action recognition algorithm based on improved resnet and skeletal keypoints in single image. *Mathematical Problems in Engineering* 2020(1):6954174, 2020. doi:10.1155/2020/6954174.

[13] N. S. Logan, H. Radhakrishnan, F. E. Cruickshank, P. M. Allen, P. K. Bandela, et al. Imi accommodation and binocular vision in myopia development and progression. *Investigative Ophthalmology & Visual Science* 62(5):4, 2021. doi:10.1167/iovs.62.5.4.

[14] Z. Lyu. State-of-the-art human-computer-interaction in metaverse. *International Journal of Human–Computer Interaction* 40(21):6690–6708, 2024. doi:10.1080/10447318.2023.2248833.

[15] J. Ramadoss, J. Venkatesh, S. Joshi, P. K. Shukla, S. S. Jamal, et al. Computer vision for human-computer interaction using noninvasive technology. *Scientific Programming* 2021(1):3902030, 2021. doi:10.1155/2021/3902030.

[16] J. C. A. Read. Binocular vision and stereopsis across the animal kingdom. *Annual Review of Vision Science* 7(1):389–415, 2021. doi:10.1146/annurev-vision-093019-113212.

[17] G. R. E. Said. Metaverse-based learning opportunities and challenges: a phenomenological metaverse human–computer interaction study. *Electronics* 12(6):1379, 2023. doi:10.3390/electronics12061379.

[18] S. Seinfeld, T. Feuchtner, A. Maselli, and J. Müller. User representations in human-computer interaction. *Human–Computer Interaction* 36(5-6):400–438, 2021. doi:10.1080/07370024.2020.1724790.

[19] W. Xu, B. Chen, Y. Hu, and J. Li. A novel wide-band directional music algorithm using the strength proportion. *Sensors* 23(9):4562, 2023. doi:10.3390/s23094562.

[20] J. Zhang, Z. Chen, and D. Tao. Towards high performance human keypoint detection. *International Journal of Computer Vision* 129(9):2639–2662, 2021. doi:10.1007/s11263-021-01482-8.

[21] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In: *Seventh IEEE International Conference on Computer Vision*, vol. 1, pp. 666–673, 1999. doi:10.1109/ICCV.1999.791289.

[22] Z. Zimiao, X. Kai, W. Yanan, Z. Shihai, and Q. Yang. A simple and precise calibration method for binocular vision. *Measurement Science and Technology* 33(6):065016, 2022. doi:10.1088/1361-6501/ac4ce5.