www.cadanda.com Computer-Aided Design AND APPLICATIONS

Computer-Aided Hierarchical Interactive Gesture Modeling Under Virtual Reality Environment

Lihua Zhu[1] and Na Lu[2]

[1]School of Computer Science and Information
Engineering, Anyang Institute of Technology, Henan, Anyang 455000, China,
zhulihuagg@sohu.com

[2]School of Computer Science and Information
Engineering, Anyang Institute of Technology, Henan, Anyang 455000, China, luna8203@sina.com

Corresponding author: Lihua Zhu, zhulihuagg@sohu.com

Abstract. The first mock exam is based on the interaction gesture in virtual reality. There are many different types of interaction gestures. In this paper, a dynamic visual gesture interaction method suitable for desktop VR is designed with the help of computational assistant tools and gesture recognition method. Through the interactive experimental analysis of the visualization system, the designed dynamic gesture interaction method works well in the immersive visualization system, provides users with a high sense of immersion and helps users understand the data more effectively. In addition, aiming at the shortcomings of leap motion gesture tracking in desktop VR, we use computer vision technology and deep learning gesture recognition method to propose a strategy. By calculating the self-occlusion degree of hand images captured by deep camera, we recognize simple gestures with low self-occlusion degree by fast model-based method, the complex gesture image uses the trained convolution neural network for joint point recognition, which not only ensures the recognition accuracy, but also improves the problem of low frame number when the deep learning method is applied to VR. Finally, the superiority and reliability of computer-aided hierarchical interactive gesture modeling for virtual reality are verified by experiments.

Keywords: Computer aided system; interactive gesture modeling; virtual reality; Leap Motion
DOI: https://doi.org/10.14733/cadaps.2022.S5.43-53

1 INTRODUCTION

Video gesture is a research hotspot. Especially in the virtual reality environment, gesture interaction has very attractive characteristics. At present, the understanding of interactive gestures is still a problem that cannot be well solved [1]. The main reason is that it includes many different types and is a dynamic time-varying process, the first mock exam is still limited to using single model or classifier to recognize the gesture. There are usually two ways to capture gesture recognition information: one is to collect hand data with data gloves, and the other is to capture images or video frames with a camera to analyze hand data.

Human motion capture is widely used in VR, AR, movies, games, human-computer interaction and object processing. Especially in VR applications, in order to provide a stronger sense of immersion and enhance the sense of interaction with virtual objects, accurate motion capture is particularly important. Among them, whether human posture capture or gesture capture has a lot of application and development prospects. The representative products of motion capture system based on computer vision include Kinect for capturing body movements, leap motion for capturing gestures and Intel RealSense for recognizing expressions and gestures. This kind of equipment can track multiple targets in the detection area with high accuracy; At the same time, the detection target does not need to wear any sensor equipment, and there are few motion constraints. However, the human posture capture method using visual technology is easy to be disturbed by the external environment, and the user's movement space limit also reduces its practicability; The optical motion capture system detects marker points through multiple cameras placed in the scene space, among which motion analysis in the United States is a representative. However, due to the large amount of signals collected by multiple cameras and the complex spatial solution algorithm, the real-time performance of the system and the operation performance of the data processing unit are greatly reduced, and the price is relatively expensive; The principle of motion capture system based on inertial sensor is to attach inertial sensor devices such as accelerometer, gyroscope and magnetometer to important motion nodes of human body, and realize motion capture through algorithm according to the motion data of key positions. In the past decade, significant progress has been made, this research has made great progress. However, due to self-similarity, occlusion, wide range of joints and changing hand shapes, or the complex and diverse posture of the human body, as well as the lack of depth sensors, Singh professor point out that it is still challenging to estimate 3D human key points from a single image [2].

2 RELATED STUDIES

Jeschke et al. [3] proposed to use the directional radial distribution feature and the ability to globally describe the hand posture to locate the fingertip, so as to realize gesture recognition. Although the shape features of each gesture are different, it is easy to make recognition errors for gestures with similar contour. The gesture recognition algorithm based on Hausdorff distance template matching proposed in document [4] calculates the main direction of gesture and the coordinates of pixels in the target area, which solves the problem of gesture rotation, but fails to solve the problems of light transformation, skin color interference and so on. Oudah et al. [5] first carries out scale invariant feature transformation and feature extraction on the image, and then carried out gesture recognition with multi classification support vector machine (SVM). This method has high gesture recognition efficiency, but the complexity of the algorithm is high, so that the recognition speed becomes slow. The hierarchical gesture recognition method combining finger detection and gradient direction histogram proposed in document, Sun et al. [6] selected a corresponding classifier from the pre trained classifier set according to the number of fingers, and uses the selected classifier to complete the recognition task. This method has high recognition accuracy, but it is too manual, Moreover, using skin color model to extract hand region does not solve the problem of skin color interference. Roberts et al. [7] proposed multimodal video gesture recognition based on nonparametric feature matching. This method uses multiple feature-matching conditional random fields and uses the appearance information near the moving hand in RGB image to capture the detailed movement of fingers. It has too much calculation, long time-consuming and low real-time performance. Yi et al. [8] segmented the gesture area based on skin color detection, used the gesture pixel coordinates to obtain the main direction of the gesture as the spatial gesture feature, and finally used the Hausdorff like distance template matching method to recognize the gesture. The algorithm has high robustness and recognition accuracy, but the amount of calculation of template matching is large, and there are errors when there is skin color interference. Therefore, the single frame detection method initialized per frame is easier to recover from the estimation error. Recently, depth learning provides a new direction for estimating the hand from depth images. Hybrid method is the latest trend of hand tracking, which combines generation and discrimination methods. It can independently overcome the limitations of their respective methods and integrate their advantages. Whether initialization or recovery from errors, the generation method is effectively supplemented by discrimination method.

Vision based gesture recognition needs to go through three stages: segmentation, representation and recognition. In the gesture segmentation stage, the palm is segmented from the environment [9]. In this link, the influence of background on gesture segmentation is mainly solved. Gesture representation represents the human hand model through certain features. The more commonly used contour features are fingertip features and palm features. Firstly, the palm is partitioned and the weights are set for different regions to determine the palm, so as to reduce the influence of the finger part on the palm position; Although this method can obtain better palm position, it has high requirements for regional division; In this paper, a fingertip recognition method based on contour curvature is adopted. In addition, the palm position is determined by calculating the direction of the maximum inscribed circle of the palm contour, which has limited applicability for specific gestures. Gesture recognition is based on features to distinguish the meaning of human gestures. The common methods are template matching method and classification algorithm [10].

3 COMPUTER AIDED HIERARCHICAL INTERACTIVE GESTURE MODELING UNDER VIRTUAL REALITY ENVIRONMENT

3.1 Framework Design of Computer-Aided Gesture Interaction Design System

Virtual reality technology uses computer technology and three-dimensional modeling technology, combined with various sensing devices, to provide users with a more real virtual world. At present, the most commonly used VR technology is head mount display (HMD). Because the virtual reality headgear completely covers the user's line of sight, the traditional interaction mode of mouse and keyboard has great limitations. In order to provide a good user experience of data visualization, combined with virtual reality technology and dynamic gesture recognition technology, this experiment proposes a three-dimensional visual hierarchical gesture interaction system for desktop VR. The basic framework of the design is shown in Figure 1.

As shown in Figure 1, the computer-aided hierarchical interactive gesture recognition system for virtual reality consists of two parts: data collection and gesture recognition. In the computer-aided hierarchical interactive gesture recognition system for virtual reality, the human-computer interaction strategy implements the reasoning and Simulation of interactive behavior through double-layer perception mechanism and active object technology, so as to realize the natural and harmonious human-computer interface of gesture interactive system. The main principles of human-computer interaction strategy include: 1) integrating interactive information in the computer-aided hierarchical interactive gesture recognition system environment for virtual reality, so that the static virtual scene and virtual object contain rich interactive semantic information; 2) On the basis of integrating interactive information, a double-layer perception mechanism is used to perceive the action information of the user's operation corresponding to the user's physical world;

The inner layer corresponds to the virtual world of the active object, by creating a small world situation centered on the active object, perceives the user's interaction intention and predicts the object's interaction behavior; 3) On the basis of perceiving the user's interaction intention, the object's interaction behavior simulation is executed, and the interaction behavior is fed back to the user by multi-channel integration. The essence of gesture interaction is the acquisition, processing and feedback of interactive information generated in the process of human system interaction.

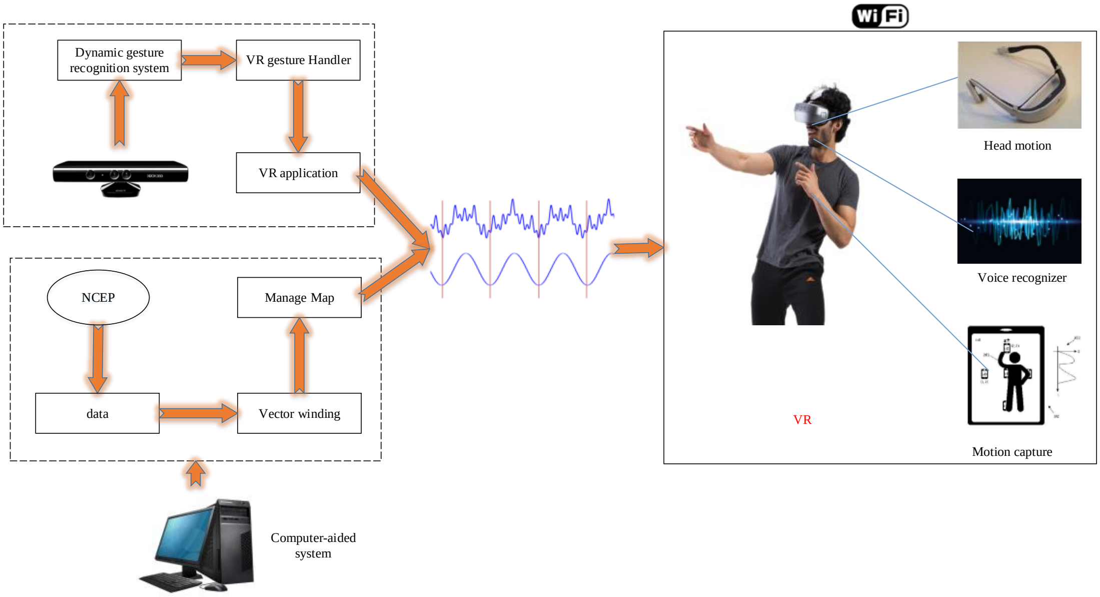

Figure 1: Framework design of computer aided gesture interaction design system.

This paper gives a broader connotation to the concept of interactive information, that is, it holds that all knowledge and phenomena meaningful to the completion of interactive tasks in the process of human-computer interaction are regarded as interactive information, which can be the interactive behavior input by users, the interactive behavior output by the system, the interactive intention of users, the characteristics of interactive objects Interactive semantic information and rules and knowledge related to interactive tasks. We regard human-computer interaction as a process of continuous generation, transmission and transformation of interactive information between human and computer. Through the two-way flow of interactive information, we constantly change the relationship between "human machine environment" and their state, and finally promote the completion of interactive tasks. Therefore, the goal of human-computer interaction in the virtual assembly environment to be studied in this paper is to enable the smooth and barrier free transmission and translation of interactive information between human and computer, reduce the cognitive gap between human and computer as much as possible, and minimize the distance between the user's mental model of task and the task execution mode presented by computer during human-computer interaction. Therefore, the process design of computer aided gesture interaction design system is shown in Figure 2. As shown in the figure, the process is mainly divided into three parts: data acquisition, feature extraction and gesture recognition.

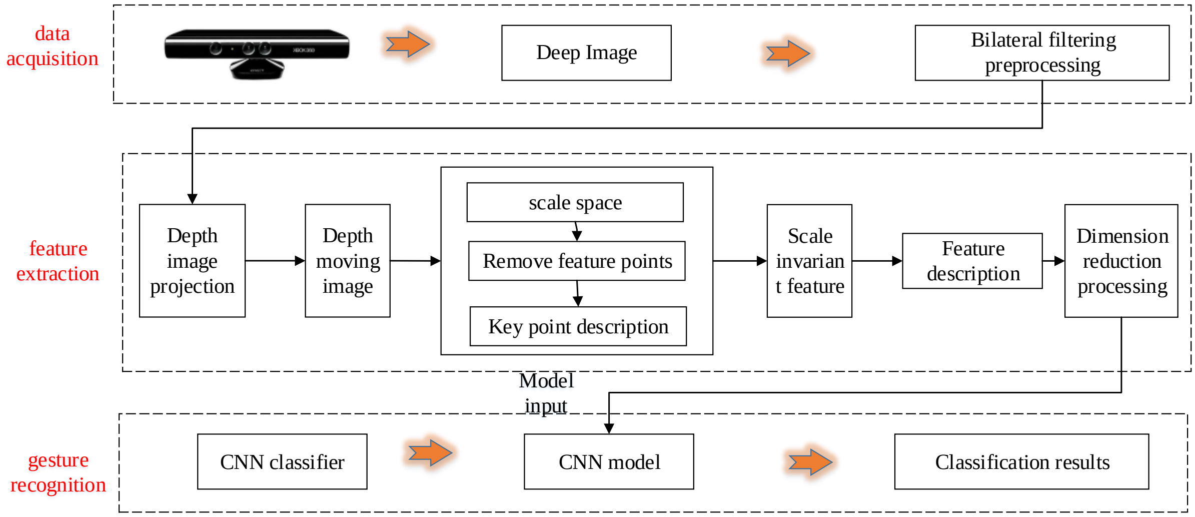

Figure 2: Process design of computer aided gesture interaction design system.

3.2 Gesture Interaction Modeling of Single Depth Image for Virtual Reality

In VR experience, image rendering takes up a lot of computing resources. In a large number of current methods, there is still a tradeoff between real-time and accurate performance. The technology with high precision usually works at low frame rate, so it is not suitable for interactive systems in spatial immersion scenes. In order to track gestures in real time with high accuracy, a gesture recognition method based on fast model is proposed. The basic framework of the algorithm is shown in Figure 3. The method based on fast model can be divided into the following steps: obtaining depth image frame, k-nearest neighbor method, extracting hand, threshold segmentation image and fast point cloud fitting.

The k-nearest neighbor method: it uses the points and takes this as the ball center to extract a sphere with a radius of 10cm.

Threshold segmentation: Using the outer motion perception mechanism in the double-layer perception model and the interactive situation information of human hand, the workspace situation and motion characteristics of human hand are analyzed, and the types of user interactive gestures are recognized. If it is recognized as a swimming gesture, it can be mapped into the viewpoint roaming of the virtual scene through the human hand action; If it is recognized as a grasping gesture, the selection operation of the virtual object can be performed; If it is recognized as an assembly gesture, it is transformed into the interactive intention of the user to perform a certain operation task. Interactive information acquisition completes the mapping from the real world to the virtual environment at two levels, including the mapping of user's motion and action from the real world to the virtual world and the mapping of user's interactive intention from the real world to the virtual world. Interactive intention reasoning is an intermediate link between the acquisition of interactive information and the execution of interactive behavior. Its role is particularly important. It directly determines whether the interactive system can output the interactive behavior expected by users. The main function of interactive intention reasoning is to predict and trigger potential interactive behavior.

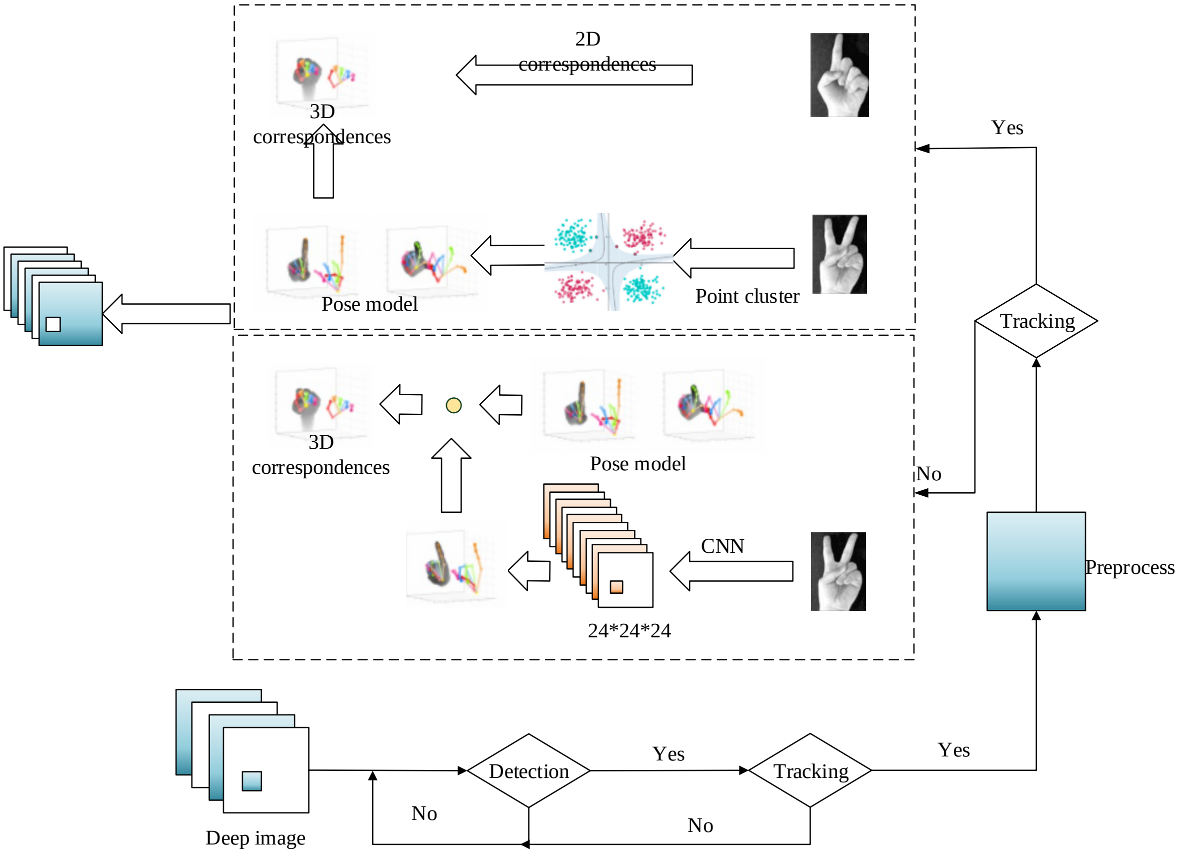

Figure 3: Gesture interaction modeling structure of single depth image for virtual reality.

Fast point cloud fitting: Let F be the common point of sensor input data. A 3D hand model $\theta=\{\theta_1,\cdots,\theta_{26}\}$ is given. Solve the optimization problem:

$$\min\{E_{3D}+E_{2D}+E_{wrist}\} \tag{1}$$

Where E_{3D} is geometric positioning, E_{2D} is alignment of 2D contour of hand model with 2D contour extracted from sensor data, E_{wrist} means minimize wrist energy. Therefore,

$$\begin{cases} E_{3D}=\omega_1\sum_{x\in P}\left\|x-\prod_n(x,\theta)\right\|_2 \\ E_{3D}=\omega_2\sum_{x\in P}\left\|p-\prod_n(p,\theta)\right\|_2^2 \\ E_{wrist}=\omega_2\left\|\prod_{2D}(k_0(\theta))-\prod_l(k_0(\theta))\right\|_2^2 \end{cases} \tag{2}$$

Minimizing the fitting energy alone can easily lead to distorted hand posture. Inspired by model integration and multi view voting, we adopt a new single CNN architecture to directly regress three-dimensional joints, estimate joints through end-to-end optimization and reasoning in single depth images, and train separate fully connected layers (FC) on multiple feature regions and merge them. Using multiple CNN requires a lot of memory and time, which is not practical for applications, especially virtual reality applications need to occupy a lot of computer resources. Inspired by CNN's multi branch set method, we regard the fusion of multiple branches as a single CNN as a generalized set type. The strategy is to fuse different scaling inputs or different images with multiple input branches. Therefore, a gesture recognition algorithm model based on fast CNN is designed in Figure 4.

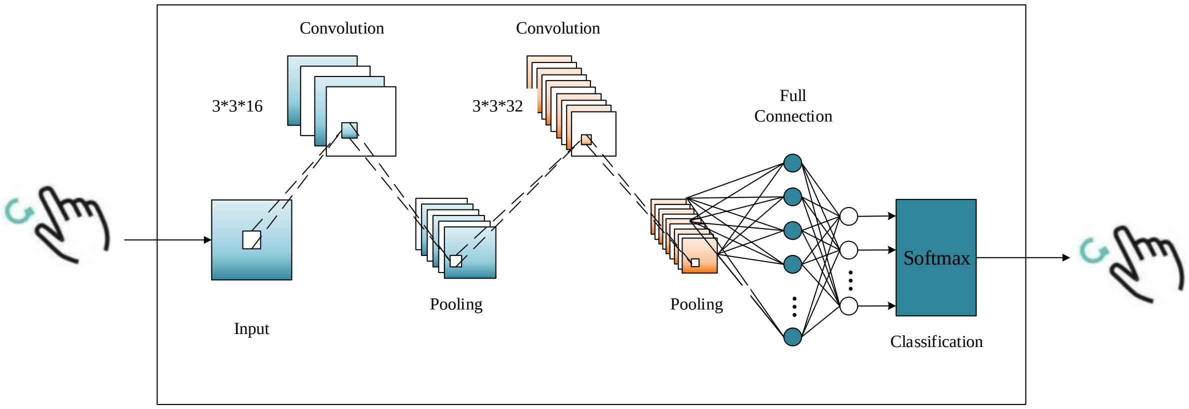

Figure 4: Structure of the CNN network.

As shown in Figure 4, the acquired gesture image is preliminarily bilateral filtered, and the processed image is preliminarily extracted through a 3*3*16 convolution network, and then through pooling operation. After repeating this twice, the image and feature information are sent to the full connection layer. Finally, the classification prediction is carried out by Softmax function.

4 ANALYSIS OF RESULTS

4.1 Simulation Environment and Data Sources

In order to verify the effectiveness of gesture recognition algorithm based on composite feature and dynamic threshold circle method, relevant experiments are carried out. The experiment is based on PC platform and Kinect sensor. PC has 2.9 GHz CPU and 8 GB ram; Kinect's sample image size is 320 × 240, frame rate 30 FPS, which can meet the hardware requirements of real-time recognition. Use Openni and OpenCv open-source libraries to build an experimental platform in visual studio.

The training process is summarized as: 1) The training samples are hand-made images containing gestures according to the experimental requirements, and various labels are marked for each gesture. 2) The image and the corresponding label are sent to the gesture detection part for ACF feature extraction, and the gesture part is detected and segmented. 3) CNN is used to decompose the segmented image into high-frequency and low-frequency blocks, extract hog features from high-frequency blocks, extract LBP features from low-frequency blocks, and cascade fuse the high-frequency and low-frequency block features to obtain the features finally used for gesture recognition. 4) The fused features and corresponding tags are input into CNN as training sets for training, and one-to-one multi class classification method is used. 5) In the test phase, the input image goes through the above process and finally outputs the label corresponding to the gesture.

4.2 Results for Accuracy Verification

Area ratio: first, in order to test the feasibility of the technical contribution of this chapter as an experiment, we test the optimized network structure by trying different hand ratios in order to obtain the best performance. According to the gestures set in the experiment, each experimenter changed the shooting angle and tried to change different light irradiation conditions at the same distance from the camera. Each gesture was done 10 times, a total of 50 experiments. We randomly selected 4000 frames for the experiment. The speed of different hand area ratios and the accuracy results of different hand area ratios are shown in Figure 5

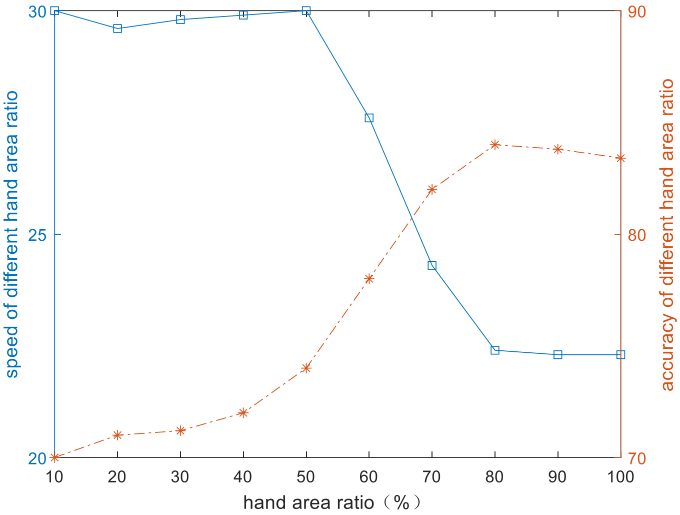

Figure 5: Accuracy verification of the model.

We choose 60% area ratio as the equilibrium point. Under this threshold, both accuracy and speed have good performance.

4.3 Results for Superiority Verification

We select the success rate as the performance indicators. Under the condition of no skin color interference and background, each experimenter did 8 times for each gesture, a total of 3400 experiments. In addition, five scenes were selected as the background. The experimenter's upper body was displayed in the picture. In addition to the skin color of the hand, there were also the skin color of the face and arm. Each gesture was done 5 times, a total of 1000 experiments, the predictions are shown in Figure 6. As shown in Figure 6, the model proposed in this paper has certain advantages over Ren algorithm, but there is still a little gap compared with model algorithm.

In order to vertically compare the advantages of interactive gesture model in VR environment, the users are trained to understand the function and characteristics of gesture interaction in the system; Then users can complete a series of predefined interactive tasks in the way set by the system. The statistics of the usage of 10 users show that the interactive gesture understanding method in this paper has a good gesture recognition accuracy. As shown in Figure 7, the recognition speed of this method can meet the needs of real-time gesture interaction.

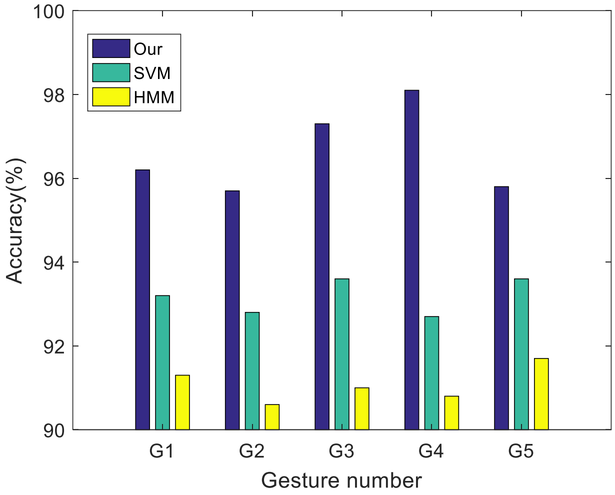

Figure 6: Superiority verification of the model.

Compared with the traditional method based on a single classifier, such as HMM Compared with Bayesian network or neural network, this method determines the basic types of gestures through rough classification, and then uses a variety of models for modeling and recognition, which improves the accuracy and efficiency of the analysis process.

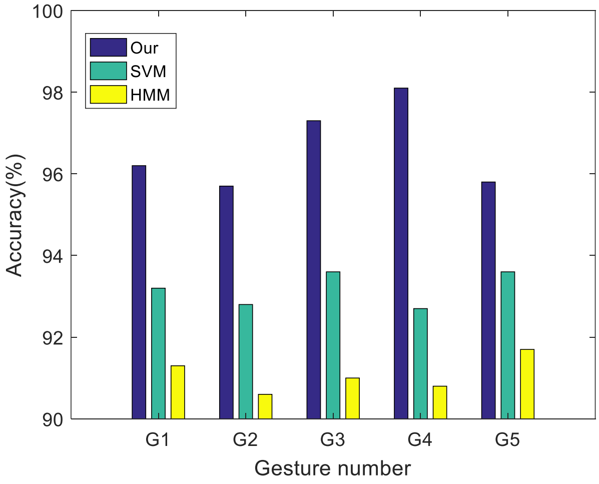

Figure 7: Superiority verification of the model.

We compare the method in this paper with SVM and HMM. Although the HMM method can automatically segment and recognize continuous video input, it is necessary to calculate the posterior probability of 13 sub HMMs in the multi HMM ring model for each frame of gesture image, which is too heavy; The SVM method needs to calculate the characteristics of gestures in each frame image: the change of hand area and position, and the relative position relationship between hand and other parts of the body, and calculate the posterior probability of dynamic Bayesian network. In fact, not all interactive gestures need semantic recognition, the recognition of gestures that only need tracking but do not need recognition will reduce the system efficiency. Therefore, the gesture interaction model proposed in this paper has strong advantages.

5 CONCLUSION

This paper studies human key point recognition and interaction for virtual reality. Firstly, leap motion gesture capture is used for interactive operation, which improves the operation experience of vector field climate data visualization system in virtual reality environment. In addition, taking desktop VR as the application background, a joint strategy is proposed to improve the real-time and accuracy of gesture recognition. Finally, through computer vision technology, the defects of the traditional inverse motion human motion capture system are improved. For different VR environments, various tracking methods are studied to try more suitable solutions. Therefore, a computer-aided immersive dynamic gesture interaction design system for desktop VR is designed. Aiming at the problem of large recognition error when jitter and self-occlusion are serious, we introduce the bare hand interaction mode through the front depth camera. It can help users focusing without moving their head and following the movement. By combining the model-based method and CNN network, the speed accuracy tradeoff strategy of gesture interaction is introduced in the desktop VR environment. The algorithm achieves high real-time and high precision.

Lihua Zhu, https://orcid.org/0000-0001-8407-9921
Na Lu, https://orcid.org/0000-0002-8816-600X

REFERENCES

[1] Li, X.; Wang, X.; Wan, P.-J.; Han, Z.; Leung, V.-C.: Hierarchical edge caching in device-to-device aided mobile networks: Modeling, optimization, and design, IEEE Journal on Selected Areas in Communications, 36(8), 2018, 1768-1785. https://doi: 10.1109/JSAC.2018.2844658

[2] Singh, G.; Mantri, A.; Sharma, O.; Kaur, R.: Virtual reality learning environment for enhancing electronics engineering laboratory experience, Computer Applications in Engineering Education, 29(1), 2021, 229-243. https://doi.org/10.1002/cae.22333

[3] Jeschke, A.-M.; de Groot, L.-E.; van der Woude, L.-H.-V.; Lansink, I. -O.; van Kouwenhove, L.; Hijmans, J. -M.: Gaze direction affects walking speed when using a self-paced treadmill with a virtual reality environment, Human movement science, 67, 2019, 102498. https://doi.org/ 10.1016/j.humov.2019.102498

[4] Zhao, H.; Swanson, A.-R.; Weitlauf, A.-S.; Warren, Z.-E.; Sarkar, N.: Hand-in-hand: A communication-enhancement collaborative virtual reality system for promoting social interaction in children with autism spectrum disorders, IEEE transactions on human-machine systems, 48(2), 2018, 136-148. https://doi.org/10.1109/THMS.2018.2791562

[5] Oudah, M.; Al-Naji, A.; Chahl, J.: Hand gesture recognition based on computer vision: a review of techniques, journal of Imaging, 6(8), 2020, 73. https://doi.org/10.1109/ICRA40945.2020.9197301

[6] Sun, Y.; Weng, Y.; Luo, B.; Li, G.; Tao, B.; Jiang, D.; Chen, D.: Gesture recognition algorithm based on multi-scale feature fusion in RGB-D images, IET Image Processing, 14(15), 2020, 3662-3668. https://doi.org/10.1049/iet-ipr.2020.0148

[7] Roberts, G.; Holmes, N.; Alexander, N.; Boto, E.; Leggett, J.; Hill, R. -M.; Brookes, M. -J.: Towards OPM-MEG in a virtual reality environment, NeuroImage, 199, 2019, 408-417. https://doi.org/10.1016/j.neuroimage.2019.06.010

[8] Yi, C.; Lu, D.; Xie, Q.; Liu, S.; Li, H.; Wei, M.; Wang, J.: Hierarchical tunnel modeling from 3D raw LiDAR point cloud, Computer-Aided Design, 114, 2019, 143-154. https://doi.org/10.1016/j.cad.2019.05.033

[9] Li, X.; Zhou, Z.; Liu, W.; Ji, M.: Wireless sEMG-based identification in a virtual reality environment, Microelectronics Reliability, 98(JUL.) 2019, 78-85. https://doi.org/10.1016/j.microrel.2019.04.007

[10] Chakraborty, B.-K.; Sarma, D.; Bhuyan, M.-K.; MacDorman, K.-F: Review of constraints on vision-based gesture recognition for human–computer interaction, Iet Computer Vision, 12(1), 2018, 3-15. https://doi.org/10.1049/iet-cvi.2017.0052