A Research of Report Analysis of Gestures in VR Games

Jinhao Zhang

Bachuan Quantum High school, Chongqing, China

zhangjinhao@bcschool.cn

Abstract. This article, through literature review and theoretical analysis, points out the advantages and limitations of gesture interaction in terms of operational naturalness and scene adaptability, and provides reasonable improvement ideas and the major trends in the future direction of VR. It focuses on analyzing five typical gesture recognition technologies, such as the data glove technology based on images, which uses the AR toolkit on fingertips and inverse kinematics technology to infer the movement trajectory of the user's hand. RGB image 3D hand pose recognition technology, which analyzes a single RGB image to infer 2D and 3D images and guess the player's gesture situation. Sparse multi-channel surface electromyography (sEMG) recognition uses a 4-stream model architecture, with a Butterworth filter for noise reduction during preprocessing, and μ-law transformation to amplify weak signals, along with channel attention to amplify features, maintaining high accuracy on the DB1 and DB9 datasets. The DTW-RCE neural network recognition technology for IMU sensors processes time-series data collected by inertial measurement units (IMUs). In the data processing flow, a 9-axis IMU sensor is used as the data acquisition end, synchronously obtaining three-axis acceleration and three-axis angular velocity data at a sampling frequency of 100Hz. The IMU and acoustic sensor multimodal recognition technology enhances robustness by integrating wrist acoustic signals, recording limb movements with three-axis accelerometers and gyroscopes. At the same time, it summarizes the existing challenges in real-time performance, environmental adaptability, and generalization ability of the technology, and proposes optimization solutions.

Keywords: Preprocessing; accelerometers; gyroscopes; flexible dry electrodes; dynamic resolution.

1. Introduction

The term "Virtual Reality (VR)" was first coined in 1989 by Jaron Lanier, the founder of VPL Research, an American company [1]. As a computer-based technology, VR has been widely applied in fields including aviation, healthcare, and military affairs, with the core objective of delivering realistic scenarios to users and enabling interactions between users and non-human intelligent agents [2-4]. Since the 1990s, VR has been integrated into gaming applications.

With the advancement and development of technology, VR games have become increasingly sophisticated, gaining growing popularity among users who are now able to access and experience them on a wider scale. Traditional Human-computer Interaction (HCI) methods such as keyboards, touchscreens, and physical button have achieved maturity and widespread adoption. However, they exhibit inherent limitations in VR gaming contexts, including the requirement for physical contact with devices and the associated operational complexity. Against this backdrop, gesture interaction, an emerging HCI paradigm, has been developed and now plays a pivotal role in VR gaming. This technology enables contactless interaction: computers can map users' real-world gestures and movements to virtual scenarios, thereby facilitating the execution of a series of in-game actions. Specifically, the use of RGB images to infer 3D gesture models [5] and Neural network gesture recognition of IMU [6] has broken the constraint of physical contact imposed by traditional HCI methods.

However, gesture recognition technology remains immature, and its application in VR gaming is plagued by critical drawbacks. These include difficulties in capturing fast-moving gestures, imprecise recognition, high susceptibility to environmental factors (e.g., lighting conditions),

challenges in device miniaturization, and excessively high hardware costs that are misaligned with consumer spending habits.

Taking gesture recognition as the research subject and VR gaming as the contextual backdrop, this study identifies the current application challenges of gesture recognition through literature review and theoretical analysis, with the aim of enhancing the HCI experience for VR gamer.

2. Method

2.1. Image-based data glove technology

Through low-cost image data gloves and lightweight design, gesture recognition and reconstruction in VR scenes are achieved through visual tracking and inverse kinematics technology [7]. Among them, the typical technologies of the palm of the hand are analyzed in detail, covering technical principles, core innovations, experimental performances and other contents. Fixed wear take a miniature camera, AR Toolkit on their fingertips worn also repair the damaged by CRC algorithm design, reduce the error recognition rate, improve tracking stability, at the same time, the formula for computing the fingertips $P = Q - dN$ computing fingertips center (Q) for design center, the normal vector N for design, d is half the side length of the dice and through multi-coordinate system transformation (camera coordinate system → real hand coordinate system → virtual world coordinate system), the transformation formula is $P' = M_v\, SM_K M_c P$, where S is the scale matrix, to achieve gesture mapping.

2.2. RGB-based 3D-images inferring

It is composed of HandSegNet (precise segmentation of hand areas), PoseNet (encoder and decoder architecture output 2D key point score map), and PosePrior network (inferences 3D poses from 2D key points through standardized coordinates and perspective estimation, solving scale and rotation ambiguity) [5]. Its dataset was generated through the Mixamo 3D model and Blender rendering. Experiments show that in the coordinate system of 2D keypoint detection PCK value and PX, the area under curvature (AUC) of jointly reaches 0.49, and the AUC value of 3D pose estimation ours can reach 0.948. Superior to some deep data methods. The higher the AUC data, the better the model performs in complex scenarios.

2.3. Sparse multi-channel surface electromyographic signals

It adopts a 4-stream deep learning architecture: the first stream captures the instantaneous changes of muscle electrical signals through the TCN module, the second stream integrates LSTM and TCN to model long-range dependencies, the third stream combines CNN and TCN to extract spatio-temporal features, and the fourth stream uses skip connections to solve deep data loss [8]. In the preprocessing stage, a 4-order Butterworth filter is used for noise reduction. The weak signal is amplified by the μ-law transformation while maintaining the scale of the large amplitude signal. Finally, it is normalized to the range of [0,1]. The classification method uses the channel attention module to screen the key features. An average accuracy rate of 94.31% was achieved on the Ninapro DB1 dataset, and an accuracy rate of 93.37% was reached on DB9.

2.4. Based on IMU DTW-RCE neural network

An efficient algorithm is proposed for the time series data collected by IMU sensors to address the recognition degradation caused by the differences in gesture execution speed and duration [6]. The core innovation lies in the integration of DTW and RCE neural networks, supporting real-time learning and high-precision HGR. This method uses a 9-axis IMU sensor with a sampling rate of 100 Hz to collect three-axis acceleration and three-axis angular velocity. The data is preprocessed by an Arduino microcontroller and finally implemented on an FPGA at 150 MHz to ensure a millisecond-level response. The system avoids traditional CPU/GPU latency and is suitable for

embedded VR devices. Innovatively integrating DTW and RCE to achieve online incremental training. The experiment demonstrated 10 gestures. Multiple subjects executed them at different speeds, with each gesture repeated 20 times. The sequence length was 200 to 300 points, and the average accuracy rate was 98.6%. Single training lasts for 0.423 ms, and recognition lasts for 0.426 ms. Compared with traditional RCE, DTW-RCE has significantly improved robustness under speed variation.

2.5. Multimoding of IMU and acoustic sensors

This method addresses the bottleneck of a single IMU during arm movement and enhances robustness by fusing acoustic signals from the wrist [9]. The prototype uses 10 microphone units, but after optimization, only 1 is needed, proving that the concept is applicable to daily gesture recognition. The IMU collects three-axis acceleration and gyroscopes to record limb movements. Ten microphones are attached to the skin through the air cavity to capture the vibration of skin displacement caused by muscle contraction. The preamplifier processes the signal. The experiment avoided placing the microphone in the "wrinkle zone", and ultimately optimized it to a single microphone on the middle and front wrist. The system allows for natural forearm movement, unlike sEMG, without the need for specific muscle positioning. Innovative integrating acoustic signals to supplement the deficiencies of IMU under motion interference and enhance the accuracy of dynamic scenes. The experiment demonstrated 10 subjects, 13 daily gestures, each repeated 10 times. The IMU samples at 100 Hz and the acoustics at 44.1 kHz. The accuracy rate of the generalization model was 75%. Individual model is greater than 80%; The fusion is about 5% higher than that single IMU.

3. Discussion

The above mentioned typical technologies play a significant role in the development of VR gestures, but there are still some deficiencies. This study listed various challenges for these technologies and general solutions to the problems.

3.1. Challenges and improvement

3.1.1 The issue of precise collection

When playing VR games, people often experience unstable sEMG signals due to muscle fatigue and inconsistent muscle contraction rates, as well as electrode position shifts. Even with the TCN and LSTM fusion framework, a channel attention mechanism is still needed for screening. However, our extreme movements are misjudged due to sudden signal changes. It is possible to adopt "flexible dry electrodes" to replace traditional gel electrodes, reducing skin irritation and displacement effects, while using the "dynamic baseline calibration" algorithm to counteract signal drift caused by fatigue. For rapid movements, an "instantaneous feature caching mechanism" is added to the multi-stream TCN model to retain the key electromyographic features at the peak moment of the movement and reduce the misjudgment of sudden change signals. The ambiguity of 3D pose restoration: Users' gestures need to be precisely projected into the virtual scene in 3D space. However, a single RGB image cannot present depth information, so there is often ambiguity when estimating from 2D to 3D images, making it difficult to distinguish the difference between the two. Embed a 6-axis IMU (accelerometer + gyroscope) in the hand device, combine spatial posture with visual 2D key points, and optimize 3D posture in real time through the "Iterative Closest Point (ICP) algorithm". Similarly, in the IMU+DTW+RCE method, the IMU sensor is vulnerable to electromagnetic interference, drift or noise, resulting in a decline in accuracy in complex environments. Moreover, the recognition performance is highly dependent on the quality of the template. If the initial template is improperly selected or the dataset is insufficient, the accuracy is prone to drop. In the IMU+ acoustic fusion method, a single IMU fails during arm movement, and the acoustic signal is easily affected by external noise or skin contact variations, resulting in an

increase in the false recognition rate of dynamic scenes. For IMU-related methods, uncertainty calibration (such as UAC neural networks) can be introduced to reduce the computational complexity of DTW, enhance noise robustness and fine motion capture.

3.1.2 Real-time performance and low latency issues

The computational cost of complex models is extremely high. For instance, although the multi-stream deep learning architecture achieves a recognition accuracy rate of 94.31% (on the DB1 data), the model contains multiple convolutional layers and attention modules, which poses pressure for real-time inference on embedded VR devices (such as head-mounted displays). The frame rate of the three-stage process (HandSegNet,PoseNet, PosePrior) on a regular GPU is only 23fps, which is difficult to meet the real-time tracking requirements of high-speed actions. It is possible to reduce the number of multi-stream convolution kernels while retaining the core features. A dynamic resolution strategy is adopted for the visual model. When performing rapid actions, the image resolution is reduced to 320×240, with priority given to ensuring a frame rate greater than 60 frames. When performing fine movements, restore to 640×480 to balance accuracy and speed. Hardware optimization can also be carried out for the convolution and attention modules of the gesture recognition model. The data preprocessing and calibration are time-consuming, and the sEMG signal needs to undergo μ-law transformation, Minmax normalization and other preprocessing. The IBDG prototype needs to be manually calibrated by the user. These steps will increase the startup delay. If a user logs out of the game halfway, recalibration is required. The solution can be automated calibration. After the sEMG device is turned on, the baseline calibration can be automatically completed within 10 seconds through "random gesture guidance". Visual tracking replaces manual alignment calibration boards through scene feature point matching. Breakpoint resumption mechanism: If the device is shifted, the calibration parameters can be quickly rebuilt based on the "valid data of the last 3 seconds" without interrupting the game. Similarly, in the IMU+DTW+RCE method, the DTW algorithm involves constructing and calculating the cost matrix, which requires a time complexity of $O(n^2)$. Especially when dealing with long sequence gestures, the computational burden is heavy, making it difficult for resource-constrained devices to run in real time. In the IMU+ acoustic fusion method, the feature extraction reaches over 7,000, and the mRMR screening process is complex. It may encounter vanishing gradients or overfitting, and the computational load is high, making it unsuitable for low power.

3.1.3 Insufficient environmental adaptability and robustness

The accuracy of 2D key point detection of the environment-sensitive visual model for visual tracking decreases under low light conditions. For instance, AR Tool Kit Plus relies on binary tags for recognition. Under strong light, the reflection of the tags can lead to misrecognition. In VR games, players' movements are large and may exceed the effective tracking range of the device: The sEMG electrode has a limited coverage area and cannot capture all movements. The camera's field of view is fixed. If the player's hand extends beyond the edge of the screen, the tracking will fail. Similarly, in the IMU+DTW+RCE method, relying solely on IMU data leads to performance degradation when there is significant arm movement or external vibration interference, and the accuracy rate in multi-gesture interaction scenarios is insufficient, failing to meet the diverse requirements of VR games. In the IMU+ acoustic fusion method, the microphone needs to be closely attached to the wrist skin to capture vibrations, which may cause skin irritation, displacement or prolonged discomfort. External noise can interfere with the microphone and misidentify gestures. In the camera module, a dual mode of "auto exposure + infrared fill light" is added. In low-light scenes, infrared is automatically turned on (without affecting the human eye), and the binary marking is changed to "recent QR code + fluorescent border". In strong light, fluorescence recognition and positioning are used. Multi-device collaboration: Place one IMU on the wrist and one on the upper arm. Combined with the positioning of the head-mounted display, calculate the full-body movements to make up for the coverage limitations of a single device. When the hand approaches the tracking edge, predict the inertia and use the "LSTM Physics Engine" to

predict the movement trajectory for the next 0.5 seconds. Even if it briefly exceeds the range, there will be no lag. Combining sEMG, visual or in-ear sensors to supplement the IMU acoustic limitations enhances noise robustness and fine motion capture.

3.1.4 Hardware dependency and cost issues

In the IBDG and IMU+ acoustic fusion methods, specific hardware (such as micro cameras, microphone arrays or FPgas) is required, which increases the size and cost of the equipment (e.g., FPgas are relatively expensive), and is not conducive to the portability and consumer popularization of VR games. The sEMG electrodes and acoustic sensors need to be attached to the skin. Wearing them for a long time may cause discomfort or allergies, affecting the gaming experience. It is possible to develop integrated sensor chips (such as integrating IMU and microphones into smart bracelets) to reduce costs and achieve wireless operation. The wearable parts are custom-made with 3D printed flexible materials to adapt to different hand shapes of users and reduce discomfort. Lower the entry threshold through open-source hardware platforms (such as Arduino-compatible modules) and promote community optimization.

3.1.5 Insufficient generalization ability and multi-user adaptability

The generalization model accuracy of multimodal methods (such as.IMU+ acoustic fusion) is only 75%, and specific training by users is required. RGB images and the DTW-RCE method are sensitive to different skin colors, lighting or user habits, resulting in reduced cross-user performance. The dataset relies on synthetic data (such as the Mixamo model) and has poor generalization in real VR scenarios (such as multi-hand interaction). It is possible to utilize transfer learning and domain adaptation techniques to migrate from large-scale synthetic datasets to real user data. Introduce self-supervised learning to reduce annotation requirements, or develop adaptive mechanisms (such as real-time model fine-tuning in games) to support multi-user switching and seamless training, thereby enhancing the generalization rate.

3.2. Future prospects

However, it still faces multiple challenges and needs to focus on key directions for continuous breakthroughs in the future. technology continues to develop, core bottleneck has not been completely resolved, mainly embody in three aspects: signal and data level: state of sEMG signal is susceptible to muscle, electrode position offset and individual physiological differences, lead to unstable characteristics, extreme action easy cause misjudgment, there is ambiguity in 3D poses. RGB images without depth information have difficulty distinguishing different 3D poses of similar 2D projections. Hardware and experience: vision limited data glove, IMU, acoustic sensor contact equipment such as easy to cause the user is unwell, and part of the equipment depends on complex hardware (FPGA), increase the cost and difficulty of deployment. Model and real-time level: high precision model (such as flow more deep learning architecture) complexity is high, the pressure in the embedded real-time VR devices reasoning, some technical frame rate is low, difficult to meet the demand of high speed motion tracking; At the same time, data preprocessing and calibration are time-consuming and can easily undermine the immersive experience of the game.

It is speculated that the future needs to focus on two big innovation direction; Multimodal sensor fusion by integrating multi-source data such as visual [10], physiological, and motion data, feature-level or decision-level fusion is used to compensate for the deficiencies of a single modal, thereby enhancing the robustness of recognition in complex scenarios. Lightweight and generalization optimization; Develop lightweight deep learning models to reduce computational load; At the same time, optimize the hardware design to improve the user's wearing experience. In addition, the further study still needs to promote deep learning model improvement, personalized interaction design, professional in the field of application, promote VR gesture recognition technology in the game scene to achieve more accurate and stable human-computer interaction.

4. Conclusion

This study conducts a comprehensive analysis of gesture recognition technology in the context of VR games. By sorting out the technical development background, comparing typical HCI modes, dissecting core gesture recognition on VR technical paths, and analyzing existing challenges, it clarifies the core value of gesture recognition in breaking the physical limitations of traditional interaction and enhancing the sense of immersion in VR games. The research finds that the current five typical technologies—based on image data gloves, RGB image 3D poses, sparse multi-channel surface electromyography signals, and IMU multimodal fusion—have achieved breakthroughs in recognition accuracy (e.g., the IMU+DTW-RCE method has an average accuracy of 98.6%), hardware lightweighting e.g.the acoustic fusion solution optimized with a single microphone, and scenario adaptability, providing diverse technical support for handle-free VR interaction. However, the practical application of the technology still faces four core bottlenecks. Firstly, in terms of signal acquisition, sEMG signals are easily interfered by muscle status, and 3D poses derived from RGB images have depth ambiguity. Secondly, in terms of real-time performance, high-precision models such as multi-stream deep learning have high computational costs, with an average frame rate of only 23fps on ordinary GPUs, which is insufficient to meet the tracking needs of high-speed movements. Thirdly, in terms of environmental robustness, visual tracking is significantly affected by lighting conditions, and the performance of IMUs degrades under large-scale movements. Fourthly, in terms of hardware and generalization, dedicated equipment increases costs, and the generalization accuracy of multimodal models is only 75%, leading to poor adaptability across users and real scenarios. The solutions proposed in this study, "dynamic resolution strategy," "multi-device collaborative tracking," and "transfer learning to optimize generalization," can effectively provide directions for technical optimization. In the future, gesture recognition technology for VR games should focus on two core directions: "multimodal fusion" and "lightweight generalization." On one hand, through feature-level fusion of multi-source data (including visual, physiological, and motion data), the defects of a single modal can be compensated. On the other hand, through the design of lightweight models such as model compression, edge computing and the integration of flexible hardware for example smart bracelet-style sensors, the balance between accuracy and cost can be achieved. With technological iteration, gesture recognition is expected to further break through existing limitations, promoting the upgrade of VR games from "handle control" to "natural gesture interaction." At the same time, it will provide technical references for HCI innovation in consumer electronics, medical care, and other fields, helping to build a more efficient and immersive intelligent interaction ecosystem.

References

[1] Lanier J. Who owns the future? Simon and Schuster; 2014.

[2] Guthridge R, Clinton-Lisell V. Evaluating the efficacy of virtual reality (VR) training devices for pilot training. Journal of Aviation Technology and Engineering. 2023;12(2):1.

[3] Bracq M-S, Michinov E, Jannin P. Virtual reality simulation in nontechnical skills training for healthcare professionals: A systematic review. Simulation in Healthcare. 2019;14(3):188–94.

[4] Lele A. Virtual reality and its military utility. Journal of Ambient Intelligence and Humanized Computing. 2013;4(1):17–26.

[5] Zimmermann C, Brox T. Learning to estimate 3D hand pose from single RGB images. In: IEEE International Conference on Computer Vision. 2017.

[6] Kim M, Cho J, Lee S, Jung Y. IMU sensor-based hand gesture recognition for human–machine interfaces. In: Italian National Conference on Sensors. 2019. p. 19.

[7] Pamplona VF, Fernandes LAF, Prauchner JL, Nedel L, Oliveira MM. The image-based data glove. 2008.

[8] Shin J, Miah ASM, Konnai S, Takahashi I, Hirooka K. Hand gesture recognition using sEMG signals with a multi-stream time-varying feature enhancement approach. Scientific Reports. 2024;14.

[9] Siddiqui N, Chan RHM. Multimodal hand gesture recognition using single IMU and acoustic measurements at wrist. PLoS ONE. 2020;15.

[10] Chung S, Lim J, Noh KJ, Kim G, Jeong H. Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning. Sensors. 2019;19(7):1716.