Virtual Simulation of Dance by Integrating VR Technology and Motion Capture Technology

Ran Tao
College of Music and Dance, Huaqiao University, Xiamen 361021, China
E-mail: wudaotaoran@163.com

Keywords: virtual reality technology, three-dimensional bone extraction, motion capture, virtual simulation, graph convolutional neural network

Received: May 24, 2024

For the virtual simulation technology is difficult to meet the needs of dance movement interactivity, multi-person collaboration and real-time, a virtual dance simulation method combining virtual reality technology and motion capture technology is proposed. By using motion capture technology to capture dancers' movements and transform them into 3D movements of the virtual environment. Real-time feedback and interaction are then used to provide effective learning and performance tools for YH. The model was trained using Human3.6M dataset and obtained 81% accuracy on the self-built dataset. The validation accuracy on NTU RGB+D dataset and Kinetics dataset were 80% and 55% respectively. The accuracy of the proposed method of the study increased by 14.89%, 7.99% and 19.34% respectively compared to other methods. The accuracy of the study proposed dance virtual simulation method was 92% for all movement tests. The results show that the dance virtual simulation method designed by the study based on the fusion of virtual simulation technology and motion capture technology can achieve the interactivity of the limbs and improve the accuracy of the recognition of dance movements, which has a positive application value in the field of virtual dance experience. This study is expected to promote the further development of virtual dance experience and provide more opportunities and innovative space for dance enthusiasts and professional dancers. The study shows that this research is expected to promote the further development of virtual dance experience, providing more opportunities and innovation space for dance lovers and professional dancers.

Povzetek: Predlagana je izvirna metoda virtualne simulacije plesa, ki združuje tehnologijo virtualne resničnosti in zajem gibanja za izboljšanje interaktivnosti in prepoznavanja plesnih gibov. Prispevek omogoča realistično simulacijo in spodbuja inovacije v plesu ter izboljšuje učne in izvedbene izkušnje plesalcev.

1 Introduction

As an art form with a long history and profound cultural connotation, dance has been attracting many lovers and audiences with its unique artistic charm. However, traditional dance teaching and performance modes are often limited by venue, time and other factors, making it difficult to achieve efficient learning and creation [1-3]. At the same time, the development of dance art also needs continuous innovation and breakthrough to meet the needs and aesthetics of modern society. In this context, Virtual dance simulation, which integrates Virtual Reality (VR) and motion capture technology, has become a solution that attracts much attention. By introducing virtual reality technology into the field of dance, dancers are expected to explore freely in the virtual environment to achieve a higher degree of immersive learning and performance experience. At the same time, the application of motion capture technology also enables every minute movement of dancers to be accurately recorded, providing powerful data support for the improvement of skills [4-5]. Based on this, a virtual dance simulation method combining VR technology and motion capture technology is proposed. First, the dancers wear sensor gear to capture their body movements in real time and translate those movements into digital information. Through VR technology, users can experience simulated dance in a virtual environment, adjust various parameters of the dance, and get real-time feedback to improve their skills. The research aims to provide more ways to promote the development of dance innovation, enhance the virtual dance experience through interaction, and bring new possibilities to the field of dance. The innovation of the research is to integrate VR technology and motion capture technology to further improve the real-time performance of virtual simulation. This method is expected to improve the interactivity and realism of virtual dance, and provide a new way for dance innovation.

The study was divided into five parts. The first part introduces the research background, problems and solutions of dance virtual simulation. The second part

summarizes the research results of dance virtual simulation, and summarizes the difficulties and shortcomings of the methods. The third part introduces the dance virtual simulation method combining VR technology and motion capture technology. In the fourth part, a comparative experiment is designed to verify the accuracy and real-time performance of the proposed method by comparing it with Du's. The fifth part summarizes the research methods, analyzes the experimental results, and puts forward the shortcomings and prospects of the methods.

2 Related works

In recent years, virtual simulation experimentation techniques have developed rapidly and have had a significant impact on several fields. These technologies mainly use computer simulation and virtual reality to create near-reality experiences and environments. In order to realize the design of recommender system for shopping information search without requiring too much effort from users, Pfeiffer et al. proposed to utilize support vector machine combined with virtual reality technology to classify users' shopping search motives, so that shopping motives are identified in advance during the search process [6]. Wei et al. proposed a teaching platform for electrical engineering courses based on a virtual simulation platform, aiming to solve the limitations brought by the traditional way of teaching. Experiments show that the platform further improves students' comprehension and hands-on ability, which further improves the quality of teaching [7]. Li and other scholars proposed a virtual simulation experimental platform for chemistry experiments, aiming to further improve the teaching efficiency, while reducing the risks associated with experimental learning. Experiments show that this method can significantly improve students' innovative skills and the safety of experiments [8]. Hu and other researchers proposed a virtual simulation-based distributed wind power generation experimental system, aiming to further improve students' ability to deal with complex engineering problems. Experiments showed that the system can effectively improve students' ability to solve complex engineering [9].

On the other hand, with the development of VR technology, it has been widely used in many fields such as education and training, design and construction, military and prevention, etc. The combination of VR technology and motion capture technology can customise virtual environments according to the user's actions and reactions, providing a more realistic experience. To determine the feasibility of a commercially available virtual reality gaming system for upper extremity rehabilitation of community-based stroke survivors, Warland et al. conducted a study of postoperative rehabilitation training for stroke patients using a virtual reality system. Data from rehabilitation training movements of stroke patients were captured and analyzed, leading to the discovery of new ways to improve impairment and increase spontaneous use of functional activities [10]. Maciejewski and other researchers proposed a shooting tracking method with VR technology, aiming to improve the accuracy of military shooting training. Experiments show that the method has high selected attributes [11]. Yang and other researchers proposed a deep learning-based body pose recognition method for VR technology, aiming to capture human body movement data. Experiments showed that the method was able to effectively record various body postures [12]. Qiu et al. proposed a lighter and cheaper wireless inertial line motion capture system in order to improve the localization accuracy of human motion capture system. The human sensor network is used as the basis for foot displacement calculation using the zero-speed update algorithm, and the pose and foot trajectory fusion is performed according to the root unconstrained traversal, thus realizing the simultaneous reconstruction of human pose and displacement [13].

In summary, virtual simulation is used in fields such as medicine, aviation, and the military to provide lower-risk hands-on training environments that help learners gain experience without endangering the real world. However, it still suffers from low real-time performance and lack of experimental data. Therefore, this research proposes a dance virtual simulation that integrates VR technology and motion capture technology in conjunction with motion capture technology, which is expected to improve the problems such as low accuracy of VR technology through motion capture technology. In addition, the study further elaborates the comparison of the proposed method with the existing methods as shown in Table 1.

Table 1: Comparison of different methods

Authors	Year of publication	Methodologies	Key results and findings	Limitations
Zhao et al. [6]	2020	3DS MAX modelling, animation software and Unreal Engine 4 game engine	The platform has a relatively high focus on testing and has high testing accuracy and security	Depth image acquisition equipment is not easy to carry
Wei et al. [7]	2023	Teaching platform for electrical engineering course based on virtual simulation platform	Addresses the limitations of traditional teaching methods on students' understanding of the frontiers of science	Less interactive and real-time

Li et al. [8]	2022	Virtual simulation lab platform for chemistry experiments	Improved teaching efficiency	Poor interactivity and real-time performance
Hu et al. [9]	2021	Virtual simulation-based experimental system for distributed wind power generation	Virtual simulation-based experimental system for distributed wind power generation Improved students' ability to cope with complex engineering problems	Poor interactivity
Braun et al. [10]	2022	Motion capture method for firefighter training based on VR technology	Effectively improved firefighter training	Occlusion of human motion capture
Maciejewski et al. [11]	2020	VR technology shot tracking methods	Achieved higher selected attributes	Less real-time
Yang et al. [12]	2021	Deep neural network-based body pose recognition method for VR technology	Improved field tracking efficiency	Poor interactivity
Foreman et al. [13]	2019	A training model for Parkinson's patients based on VR technology	Effectively improves the training effect of patients	Poor interactivity
This paper	-	Virtual simulation of dance based on graph convolutional modelling and VR	Effectively improves the interactivity of limbs and the recognition accuracy of dance movements	-

3 Research on virtual dance simulation method based on VR technology

The dance virtual simulation method first requires capturing the dancer's movement data and then processing it into digitised information. Next, these data are used to create a virtual dance scene. Subsequently, the captured movements are transformed into a virtual performance. Finally, users can interact with the virtual dance through their devices and receive feedback in real time. This approach provides a more immersive dance experience.

3.1 3D bone extraction and human motion capture method

Traditional 3D skeleton extraction algorithms suffer from high computational cost, sensitivity to lighting, viewpoint changes and occlusion, and lack of accuracy and stability in complex scenes [14-15]. Unlike traditional methods, Openpose adopts a bottom-up approach, which first detects the locations of human joints and then connects these nodes to construct the human skeleton [16]. In addition, Openpose focuses more on the detection of human joint points, and therefore can better handle problems such as self-obscuration. Figure 1 shows the workflow diagram of VGG to receive images and divide the feature graph into two branches.

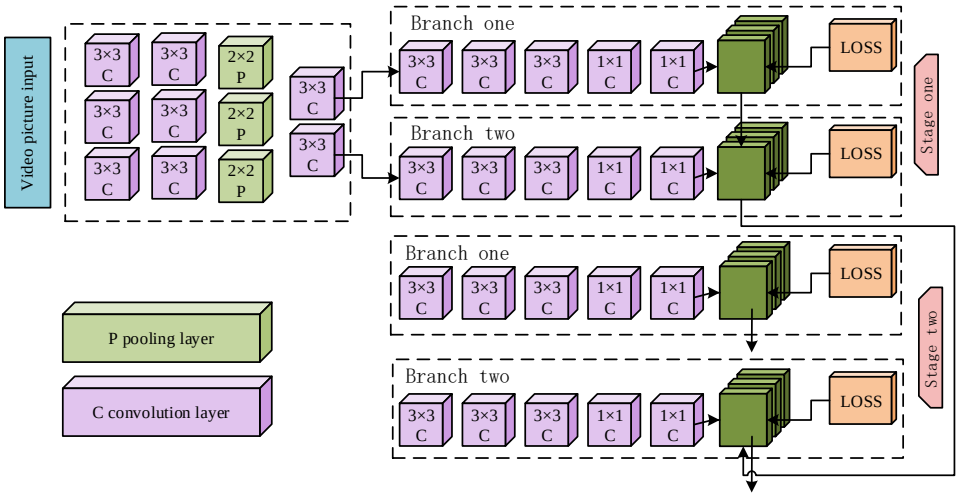


Figure 1: The VGG receives the picture and divides the feature map into two branches.

In Figure 1, the network model is divided into multiple stages, with the input to each stage being the output of the previous stage combined with the original feature map. The output of each stage consists of a set of positional confidence maps and a set of 2D vector fields. The network is optimised by computing a loss function that gradually learns to distinguish the left and right of the structure as the number of iterations increases. Eventually, the joint point coordinates are corresponded to each individual's body through the part affinity field. The calculation of the true value of the positional confidence map can be referred to Eq. (1).

$$S_{j,k}^{*}(P)=\exp\left(-\frac{\| p-x_{j,k} \|_2^2}{\sigma^2}\right) \tag{1}$$

In Eq. (1), $P\in R2$, denotes the current predicted picture, which is further expressed as the location of the j site of the k individual, and σ is a constant. The computation of the truth value of the vector field of L is shown in Eq. (2).

$$L_{c,k}^{*}(A)=\begin{cases} v & \text{If A falls on the k person's c link} \\ 0 & \text{other} \end{cases} \tag{2}$$

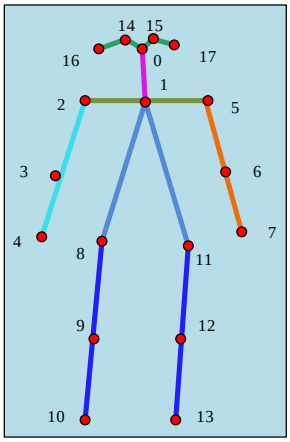
In Eq. (2), A denotes the pixel point, when A is on the c th link of the k th individual, then the unit vector between the two keypoints on this link is denoted by $L(A)$, otherwise it is taken as 0, which means that this pixel point is not on the human torso. The formula for vector divided by modulus length to find the unit vector is shown in Eq. (3).

$$v=(x_{j_2,k}-x_{j_1,k})/\| x_{j_2,k}-x_{j_1,k} \|_2 \tag{3}$$

In Eq. (3), v denotes the corresponding unit direction vector of the torso where the pixel point is located. This study utilised the 2D human pose estimation results to obtain 3D human skeleton information. The 2D human skeleton information was rapidly extracted by the Openpose tool and then matched to the pre-trained 3D model to achieve the conversion from 2D to 3D [17-18]. In the study, the Openpose open-source library was used to process the frame images of general video intercepts, and a version of GPU computing based on Linux system was chosen to complete the extraction of the 2D human skeleton. Figure 2 shows the numbering of the human joint points and the COCO-18 human skeleton.

ID	0	1	2	3	4	5
Joint	Nose	Chest	Right shoulder	Right elbow	Right wrist	Left shoulder
ID	6	7	8	9	10	11
Joint	Left elbow	Left wrist	Right hipbone	Right knee	Right heel	Left hipbone
ID	12	13	14	15	16	17
Joint	Right heel	Left heel	Oculus dexter	Oculus sinister	Right ear	Left ear

(a) Human body node serial number corresponding table



(b) COCO-18 skeleton diagram

Figure 2: The number of the human node and the figure of the COCO-18 human skeleton.

In Figure 2, there are three selection states available for human keypoints, namely 15, 18 and 25. through experiments, it is found that selecting 18 keypoints works best. Openpose is used to obtain an ordered set of 2D joint point data from general video frames. By pairing the 3D/2D pose datasets, each 2D pose information is converted from the trained "virtual camera" to the original 3D pose set. Therefore, any 2D skeleton $p'_{2d}=(x',y')$ that needs to match 3D skeleton information can be described by Eq. (4).

$$p(p_{3D\_i}\mid p'_{2D})\propto \exp\left(-\frac{\|W_iP_{3D\_i}-p'_{2D}}{\sigma^2}\right) \tag{4}$$

In Eq. (4), P_{3D} denotes the 3D pose library and W_i denotes the corresponding virtual camera transform matrix. Since the algorithm is mainly used for action recognition, the skeleton needs to be extracted from consecutive frames containing the action video stream. This reflects the fact that over a period of time, the human body pose does not change significantly from frame to frame, since the human body moves continuously. Therefore, the human motion poses in any two consecutive frames over a period of time satisfy the relationship described in Eq. (5).

$$B=\sum_{i=1}^{16}\sqrt{(p_mx-p_nx)^2+(p_my-p_ny)^2+(p_mz-p_nz)^2}\ \mathrm{B}<\iota \tag{5}$$

In Eq. (5), B denotes the Euclidean distance between individual joints in different two 3D poses; ι is a scalar parameter, and m,n denotes the similarity in the video frame sequence. For the improved matching algorithm, the first frame is required to use global matching, where Euclidean distances are computed across the entire 3D pose library to find the minimum value. Instead of using global matching for subsequent frames, matching is performed in the λ sample data before and after, starting from the position matched in the previous frame. This is done because the sample poses of the 3D pose library conform to the principle of continuity and localisation of human motion, as detailed in Eq. (6).

$$B_{nin}=\min(B_i) \tag{6}$$

The i in Eq. (6) is denoted as the frame sequence number in the video stream. When processing graph data, node features and structural information need to be considered together. Traditional manual rules are difficult to capture complex patterns, so graph convolutional neural network, a method for deep learning, is introduced, in which the convolution operation is shown in Eq. (7).

$$h_j^{a+1}=\sigma(\sum_{j\in N_a}\frac{1}{c_{aj}}h_j^lw_{R_j}^l) \tag{7}$$

In Eq. (7), l denotes the layer in the neural network; a denotes the node in the graph structure; c_{aj} denotes the normalisation factor, which is the reciprocal of the node degree; N_a denotes the neighbouring nodes of the node a; R_j denotes the type of the node a, and w denotes the weight parameter. The matching algorithm usually obtains consecutive frames over a period of time, containing information about the human 3D skeleton at different points in time, with multiple joint points in each frame. In order to capture the overall skeleton motion characteristics, the skeleton spatio-temporal graph, which is an undirected graph, is constructed, see Figure 3.

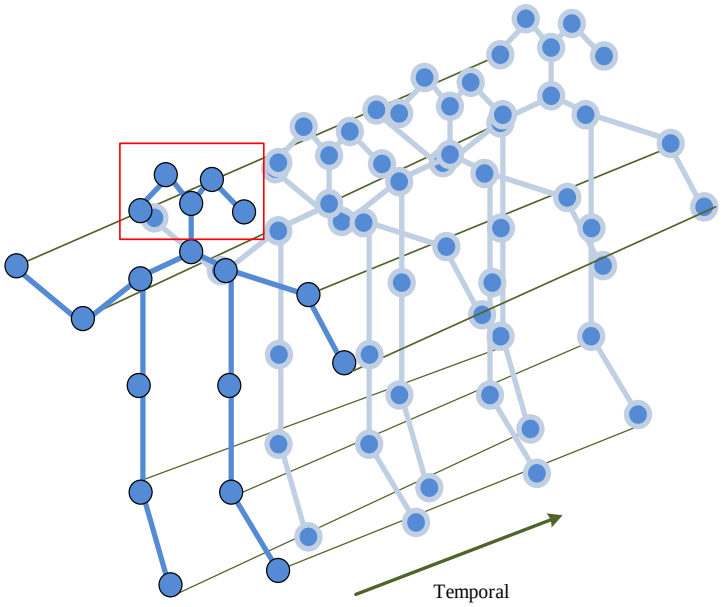
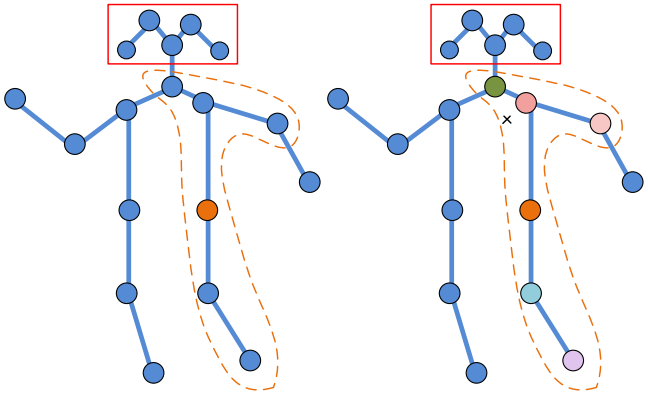


Figure 3: Space-time diagram of human skeleton.

G in Figure 3 contains T frames with N joints per frame, and the joints are connected by skeleton edges and trajectory edges. Unlike 2D or 3D convolutional neural networks, graph convolution has some differences in implementation. In spatio-temporal convolutional graph, the spatial graph within a single frame is represented by the adjacency matrix Λ and the self-connections are represented by the unit matrix I. The graph convolution of the spatial map in a single frame can be represented by Eq. (8).

$$f_{out}\Lambda^{-\frac{1}{2}}(A+I)\Lambda^{-\frac{1}{2}}.f_{in}W \tag{8}$$

In Eq. (8), f_{in} is an input feature graph, represented by the tensor (C,V,T), and the convolution process performs a standard two-dimensional convolution followed by multiplication by the normalised adjacency matrix. Throughout the graph convolution, the receptive field size and root node distance are set to 1, similar to CNN image convolution. The study increases the size of the receptive field, i.e., the value of D is set to 2 to increase the number of points sets adjacent to the root node, and the change in the receptive field is shown in Figure 4.



(a) Subsets with adjacent distance of 2

(b) Comprehensive distance and center of gravity distance division strategy

Figure 4: Subsets with adjacency distance as 2 and division strategies of comprehensive distance and centre of gravity distance.

In Figure 4, to adjust the size of the convolution kernel, the study focuses on the value of K in the spatial domain dimension. The neighbouring point set division strategy is redesigned, combining a new approach of distance and motion spatial configuration. Firstly, based on the distance from a node to the root node, the surrounding nodes are divided into three parts, the root node itself, the set of points with distance 1, and the set of points with distance 2. These sets are then divided into nodes close to the centre of the overall skeleton and nodes

far from the centre based on the spatial configuration, resulting in the division of the sensory field into five subsets.

3.2 Research on virtual simulation of human motion capture dance

Based on 3D skeleton extraction and skeleton movement classification techniques, a VR human-computer interaction system using body movements as operation commands is designed and applied to virtual reality projects. The user and processing terminal is responsible for ensuring the normal operation of the programme, processing the operation commands and directing the display of the head-mounted VR device. A Cardboard eyeglass case is used as the head-mounted terminal [19]. The Data Acquisition and Processing Terminal shoots videos of limb movements, collects data, analyses and extracts the human skeleton for movement classification. The module understands body semantics, converts semantics into interactive instructions, and sends them to users and processing terminals. The whole interaction process is shown in Figure 5.

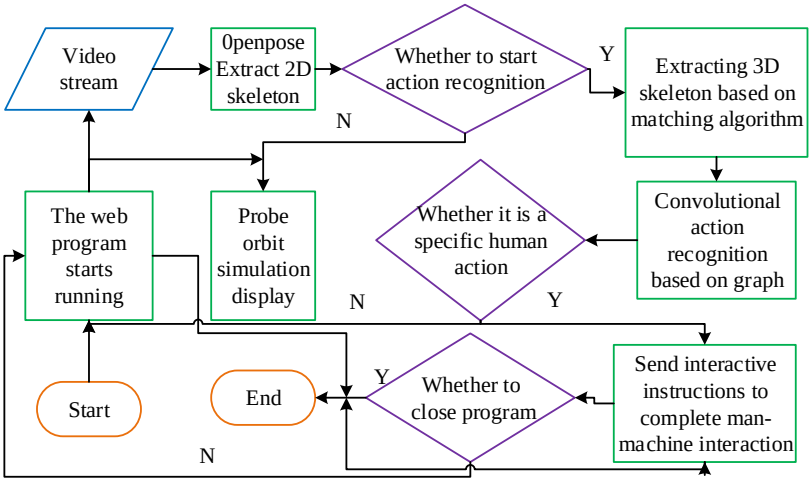


Figure 5: Flow chart of VR interactive system.

In Figure 5, the whole VR programme is first launched and developed using the Cardboard developer SDK built with the Unity virtual engine, which is eventually packaged and installed with Unity on the mobile phone. Then, the mobile phone is placed into the glasses case, the camera is activated, and the 2D skeleton of the human body is extracted in real time through Openpose, and video recording is initiated when the amplitude of movement reaches a specific threshold [20]. The 2D skeleton is converted to 3D skeleton by matching algorithm and inputted into action recognition model for classification. The classification results are mapped into interaction commands and fed back to the VR simulation programme. By enriching the action semantics, the operation commands can be extended to realise newer human-computer interaction modes, as shown in Figure 6.

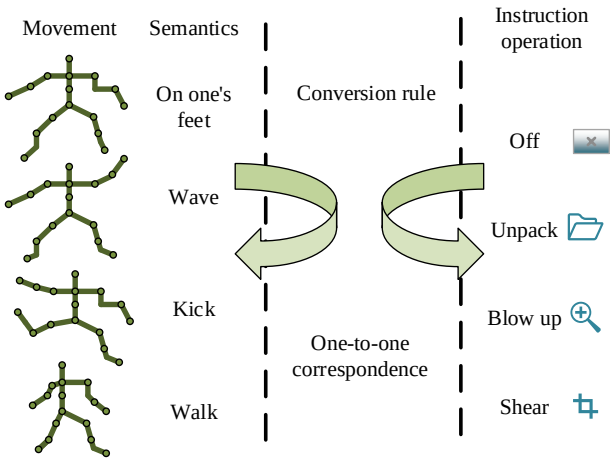


Figure 6: Relationship between body semantics and operational instructions.

In Figure 6, the study chooses to use the existing categorised actions, filtered to correspond one-to-one with the operation commands. This information is sent to the VR display module through the action recognition module, and after discrimination the delegate function is called to realise functions such as conversion of viewpoints, zooming in and zooming out, which are implemented on the Unity platform. For the purpose of the study, a large number of dance action models need to be built for training, and Eq. (9) shows the library of input samples.

$$X = \{g_{1,1}, g_{1,2}, g_{1,3}, \ldots, g_{c,n},\} \tag{9}$$

In Eq. (9), $g_{c,n}$ denotes the action fragment at the n th gesture, while the fragment contains the action gesture at m. The expression for the calculation of the action fragment at the m th gesture $g_{c,n}$ is shown in Eq. (10).

$$g_{c,n} = \{P_1, P_2, P_3, \ldots, P_m,\} \tag{10}$$

After a complete dance movement segment is projected to the output space, a "trajectory" will be formed in the output space, and a set of index numbers containing timing information will be obtained. The specific calculation formula is shown in Eq. (11).

$$O_{c,n}(t) = (o_t\}, t \in T \tag{11}$$

In Eq. (11), $O_{c,n}$ is used to identify the index sequence of dance moves for each category. According to the histogram statistics rule, a histogram of a sequence of movements containing n gestures can be represented by Eq. (12).

$$H(o_u)L_{c,n} = \frac{f_u}{n} \tag{12}$$

In Eq. (12), f denotes the frequency of occurrence of the u th output node in the dance action, and n denotes the number of gestures included in the dance. The new input movements are computed by matching the Euclidean distance with the known movement templates to discriminate the category of unknown movements. In a dance self-learning system, this recognition process can be done offline or online. Action variability is calculated by the normalised inner product of the eigenvectors of the two gestures, as described in Eq. (13).

$$d(p_i, p_j) = \sqrt{\sum_{k=1}^{N} w_k \left(\frac{f_{i,k} - f_{j,k}}{f_k(\max) - f_k(\min)}\right)} \tag{13}$$

In Eq. (13), $f_{i,k}$ and $f_{j,k}$ are the values of the k th feature vector for the postures p_i and p_j respectively; $f_k(\max)$ represents the maximum value of the k th feature; $f_k(\min)$ represents the minimum value of the k th feature, and w_k is the weight of the k th feature. In the study, the joint angle Euclidean distance is used to evaluate the accuracy of the simulated movement, and the specific calculation formula is shown in Eq. (14).

$$E_{avg} = \frac{\sum_{i'=1}^{n} D(R_{i'}, C_{i'})}{n} \tag{14}$$

In Eq. (14), E_{avg} denotes the average error, and n denotes the total number of frames describing this action. $R_{i'}$ denotes the actual value of the joint; $C_{i'}$ denotes the simulated value of the dummy, and $D(R_{i'}, C_{i'})$ denotes the Euclidean distance between the actual value and the simulated value of the data in the $i' = 1$ th frame. The common frame is the basis of the algorithm and is used to ensure that the two action segments before and after the action choreography have similar poses for a smooth transition. It is understood that for any two poses of data, the centre of gravity distance between them can be measured using Eq. (15).

$$D_R(f_i, f_i) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{15}$$

In Eq. (15), (x, y, z) denotes the coordinates of the centre of gravity of the human body, respectively. In summary, the whole process of skeleton-based action recognition and graph convolution-based human action recognition method mainly starts from data acquisition, goes through motion capture, data processing, feature extraction, action recognition, virtual reality environment setting, and finally realises the user's interaction through the VR device, and improves the skills and experience according to the real-time feedback. The details are shown in Figure 7.

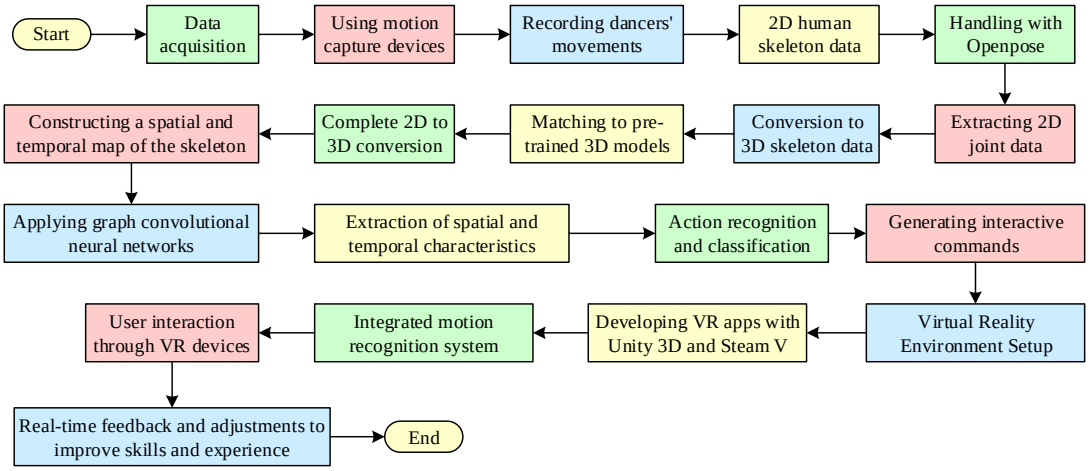


Figure 7: Implementation process of dance motion capture technology based on virtual technology.

In Figure 7, the study uses the Human3.6M dataset to pre-train the 3D model and conducts simulation experiments of the proposed method with the self-constructed dataset and the NTU RGB+D dataset and the Kinetics dataset.Human3.6M is a large-scale multiview human motion capture dataset, which provides detailed information about the human body's joint positions, and is often used for the research of 3D human motion analysis NTU RGB+D is a multimodal dataset that provides depth information and colour images and is commonly used for 3D action recognition and understanding research. This dataset records a wide range of human actions and provides detailed annotation information. Kinetics is a large-scale video dataset commonly used for action recognition and understanding research. It contains multiple categories of action videos, each labelled with an action category. The above dataset is able to accurately verify the effectiveness of the proposed method of the study, so the subsequent simulation experiments are conducted around the above dataset. In addition, based on the obtained data, the study uses Openpose to extract 2D skeleton information from video frames and uses a matching algorithm to convert the data into 3D skeleton data. Meanwhile, the data are fused and preprocessed by data overlay, feature fusion, and spatial alignment.

4 Virtual technology-based dance motion capture technology validation simulation experiments

The study was conducted using an Intel Chihuahua E5-2600 CPU with a software configuration that included CUDA 8.0, CUDNN 6.0, Python 3.5, and Tensorflow 1.4. The object of the study was the NTU RGB+D dataset containing action samples, each consisting of an RGB video, a depth map sequence, 3D skeletal data, and an infrared video. The 3D skeletal data consisted of each frame of the 25 3D positions of major body joints. The dataset covers a total of 60 different action categories, which are divided into three main categories: everyday actions, interactive actions and complex actions.

4.1 Simulation experiment of dance movement recognition based on virtual technology

The Human3.6M dataset, which contains about 3.6 million labelled human movement data samples and their corresponding RGB images, is used as the experimental training data in the study. The whole dataset can be divided into 11 broad categories, each consisting of 15 subcategories. These major classes mainly represent 11 different professional modelling experimenters, while the subclasses mainly cover different human movements. Figure 8 shows the results obtained by three different models in the Human3.6M dataset using the mean error as an evaluation metric.

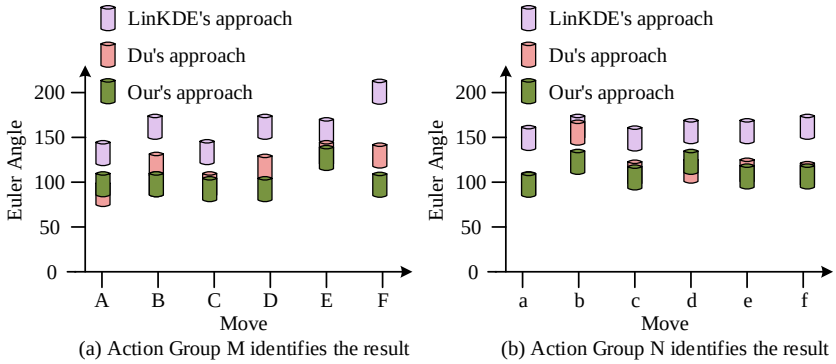


Figure 8: Comparison of identification errors of different methods on Human3.6M dataset.

According to Figure 8(a), it can be seen that the error of the proposed method is minimised when compared to other proposed methods such as Du's, except for A. The errors of the proposed method are 118, 95, 96, 137, and 112 on B, C, D, E, and F, respectively. according to Figure 8(b), the error of the proposed method is minimised when compared to other proposed methods such as Du's, except for d. The errors of the proposed method are 99, 120, 115, 117, and 112 on a, b, c, e, and f, respectively. the study is based on the method of the proposed method. 's et al. The method has the smallest error compared to all other methods proposed in the study, which are 99, 120, 115, 117, and 112 on a, b, c, e, and f, respectively. The study evaluates the effect of different sensory field delineation strategies proposed in the study on the recognition results by using the original dataset as a test set and the accuracy as an evaluating metric, and Figure 9 demonstrates the corresponding test results.

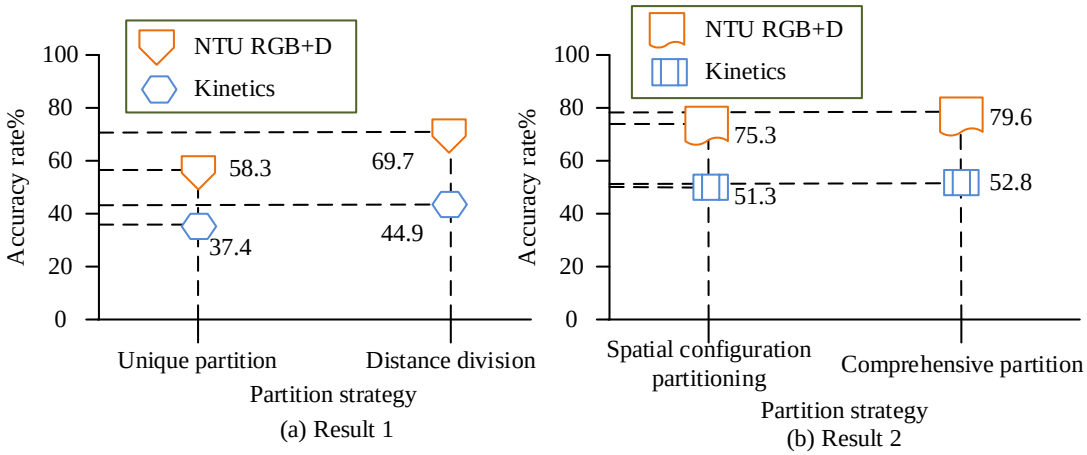


Figure 9: Results display of different partitioning strategies.

The experimental results shown in Figure 8 indicate that the graph convolution model performs differently with different subset division methods. The poor model performance when a single division method is used is mainly due to the fact that only a simple averaging of features is performed before the graph convolution process. The performance of the model progressively improves as the number of subset divisions increases, which highlights the importance of the size of the receptive field and the size of the convolution kernel. The improved model achieves a small increase in performance by enlarging the receptive field and introducing a new subset division strategy. To evaluate the effectiveness of this model, the researchers compared it to several other deep learning models for action recognition, including 3D convolutional neural networks, dual-stream convolutional neural networks and dual-stream recurrent neural networks. These experiments covered the new dataset, the NTU RGB+D dataset and the Kinetics dataset, and mainly evaluated the F1 values and accuracy of the model in the three datasets, and the specific results can be seen in Table 2.

Table 2: Comparative experimental results display.

Data set	Make by oneself		Kinetics		NTU RGB+D	
	F1 score (%)	Accuracy (%)	F1 score (%)	Accuracy (%)	F1 score (%)	Accuracy (%)
Textual convolution model	82	81	54	55	81	80
3D product model	50	57	55	56	73	75
Two-stream recursive convolution model	59	62	54	57	78	81
Two-flow convolution model	46	45	55	57	76	79

According to Table 2, the proposed method has the highest accuracy rate of 81 % in the dataset 'Make by oneself'. In the other two datasets, the accuracy rates are 55 % and 80 %. Compared with the other methods, the accuracy of the proposed method increased by 14.89 %, 7.99 %, and 19.34 %, respectively. In terms of F1 score, the proposed method is more advantageous in the 'Make by oneself' dataset and 'NTU RGB+D' dataset, while in the 'Kinetics' dataset, the 3D convolutional neural network F1 score is more advantageous in the 'Kinetics' dataset and the 'Kinetics' dataset. In the 'Kinetics' dataset, the 3D convolutional neural network F1 score is pseudo 55%, while the F1 score of the proposed method is 54%. This difference may be due to the fact that Kinetics is a large-scale video dataset, and 3D convolutional neural networks are better at capturing spatio-temporal features in videos. Overall, graph convolution as a messaging method acting on graph data, human keypoints can be used to build graph data based on human limb connections, while features that contain more information about human structure can be extracted by graph convolution. This result indicates that the graph convolution model performs well in real-time and can effectively combine temporal and spatial features. The model better captures the trajectory of the human skeleton at different moments by constructing a spatio-temporal graph based on the human skeleton sequence.

4.2 Motion capture based virtual simulation experiment for dance

In order to evaluate the reliability and feasibility of the interactive system, two key indexes of accuracy and real-time are considered in the experiment. In the laboratory environment, the action data of the test participants were collected, and the response accuracy rate of the VR interactive system based on action recognition under different action commands was recorded, so as to verify the effectiveness of the interaction process. At the same time, the response speed of the system is evaluated by measuring the response time difference between the proposed interaction method and the traditional viewpoint gaze interaction method. The study performed 100 interaction tests for each common action, and recorded the number of successes and the average interaction time, as shown in Figure 10.

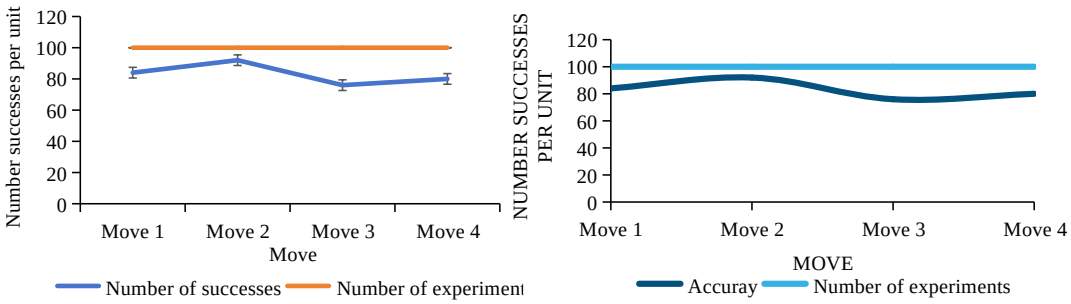


Figure 10: Experimental results of action recognition accuracy.

According to Figure 10, the accuracy rates of all movement tests exceeded 70%, with the highest being 92% and the lowest being 76%, confirming the feasibility of body movements in VR interaction. However, there are significant differences in the accuracy rates of different actions. Actions with a large amplitude and left-right direction extension were recognised better, while those with a small amplitude and front-back direction extension were recognised poorly. When switching between actions, too much speed and frequency can lead to misoperation or reduce the accuracy of the operation. Considering the importance of real-time to interactive experience and efficiency, the average VR interaction time based on body movements is compared with the traditional viewpoint gaze interaction time, and the relevant results are shown in Figure 11.

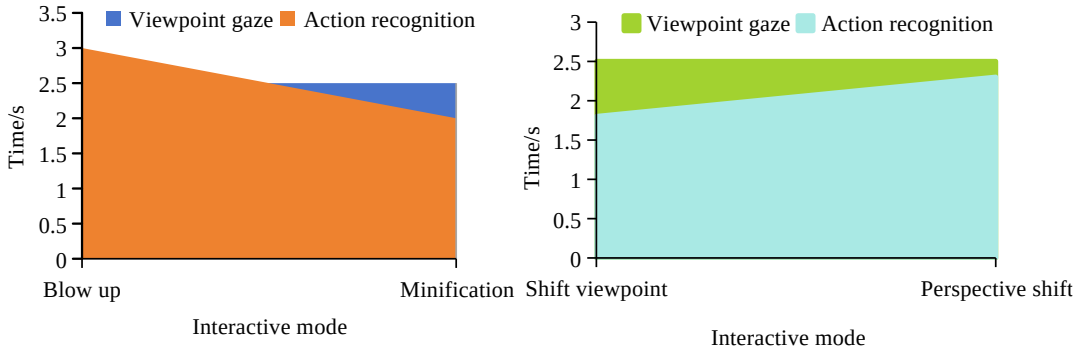
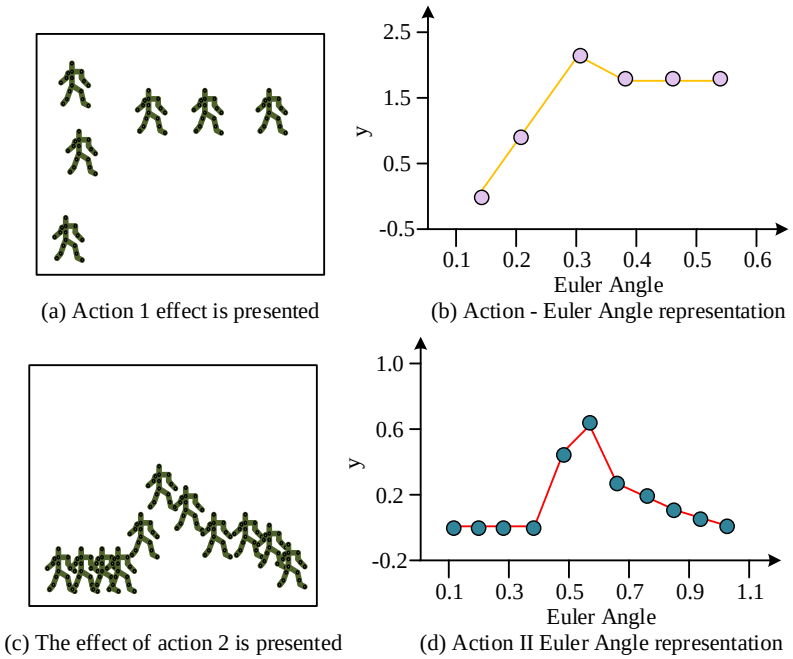


Figure 11: Action recognition interaction time comparison table (unit: second).

According to the data in Figure 11, the average results of 100 experiments show that the action recognition-based VR interaction time is generally shorter than the point-of-view gaze system. This suggests that action recognition VR interaction has advantages. The point-of-view gaze system requires an arrangement of fixed buttons in the VR space, which is not user-friendly and affects the immersion experience. In contrast, body-motion recognised interaction avoids these drawbacks, making the interaction smoother and more efficient. Therefore, body-motion-based VR interaction systems provide a more reliable alternative to VR glasses. The system developed through PC/Windows/VC5.9 software environment was used to simulate the jumping technical movements of divers and the serving technical movements of male volleyball players. The system uses Euler angles in the y-direction to represent the two movements in order to compare the similarity between the simulation modelling effects and the Euler angles to verify the effectiveness of the system. The display of the technical movements and the Euler angles are represented in Figure 12.



(a) Action 1 effect is presented

(b) Action - Euler Angle representation

(c) The effect of action 2 is presented

(d) Action II Euler Angle representation

Figure 12: Technical action effect presentation and Euler Angle representation.

Figure 12 shows that the simulation of the two technical movements by the described system is highly consistent with the representation of the Euler angles in the y-direction. This demonstrates the accuracy of the system in terms of simulation modelling and shows that its simulation results are applicable to real sports training. The system also facilitates the adjustment and modification of the movements and allows the simulation results to be used for subsequent comparative analysis with the training videos. Finally, in order to further confirm the validity and application value of the methodology proposed in the study, the study was conducted through questionnaires, interviews, usage tracking, and skill enhancement assessment, and we collected feedback from 20 dancers and 5 coaches. The specific results are shown in Table 3.

Table 3: User experience feedback.

Classification	Recommendation score	Accuracy score	Interactivity score	Ease of use score	Recommendation
Dancers	95	96	84	90	Add more dance moves and tutorials to enrich the teaching content
Coaches	97	92	80	84	Provides personalised settings to suit different levels of dancers

The average results of the feedback from 20 dancers and 5 instructors on the use of the study's proposed system are shown in Table 3, with each rating item scored out of 100. It can be seen that the VR dance simulation system proposed by the study has significant potential to improve dance learning efficiency and engagement. It is expected to become an important tool for dance education and training through further technical optimisation and content expansion.

5 Discussion

The virtual dance simulation method proposed in the study achieves 81% accuracy on the self-built dataset, and 80% and 55% validation accuracy on the NTU RGB+D dataset and Kinetics dataset, respectively. Compared to existing methods in the literature, such as those of Zhao et al [6] and Wei et al [7], the proposed method of the study achieves an improvement in accuracy of 14.89% and 7.99%, respectively. This significant improvement may be due to the fact that the diversity and scale of the dataset play a better effect on model training. The multi-view and large-scale nature of the Human3.6M dataset provides rich learning samples for the model, which helps to improve the generalisation ability of the model. And compared with the traditional 3D convolutional neural network, the graph convolutional neural network better captures the spatio-temporal features of the human skeleton data, especially when dealing with non-regular mesh data. In addition, the combination of motion capture technology and VR technology provides more detailed and accurate motion data, which provides strong support for the accuracy of motion recognition. It is worth mentioning that the study innovatively combines graph convolutional modelling with VR technology for virtual simulation of dance movements. This approach not only improves the accuracy of movement recognition, but also provides a more natural and intuitive interaction in a virtual reality environment. In addition, through real-time feedback and interactivity, studying the proposed system provides an innovative learning tool for dance learners and performers.

Although the proposed system performed well in the experiments, there are still some errors and limitations. In the experiments, it was found that the accuracy of recognising small amplitude and vertically extended movements was low. This may be due to the accuracy limitation of the motion capture system in capturing subtle movements. In addition, although the system was able to provide real-time feedback in most cases, there is still room for improvement in the system response time for the recognition of certain complex movements. In terms of user adaptability, there exists the possibility that some users may need time to adapt to the interaction in the VR environment, especially for dancers who are not accustomed to using high-tech devices. In summary, the proposed method of the study provides a novel and effective approach in the field of virtual dance simulation. By combining VR technology and motion capture technology, it not only improves the accuracy of motion recognition, but also enhances the user's immersion and interaction experience. In the future, the algorithm will be further optimised to improve the recognition of small movements; and the system response time will be improved to ensure a smoother interaction experience. More affordable equipment and software solutions will be developed to lower the threshold of use. In addition, the research will explore more application scenarios, such as dance education and rehabilitation training, in order to give full play to the potential of virtual simulation technology in the field of dance.

6 Conclusion

VR technology and motion capture technology are areas that have made huge breakthroughs in recent years, and they have already revolutionised a number of fields, including entertainment, healthcare, education and training. One such area that has received a great deal of attention is virtual simulation of dance. This research fuses virtual reality technology with motion capture technology by first

using a motion capture system to accurately record the dancer's movement data, which is then fed into a VR system. Within the VR environment, this movement data was used to create an interactive, three-dimensional dance virtual character that allowed the user to experience the dance from a first-person perspective. The experimental results show that when operating on four common body movements and their corresponding commands, the accuracy of all movements tested exceeded 70%, with a high of 92% and a low of 76%, confirming the feasibility of body movements in VR interaction. In the experiments on the display and Euler angle representation of technical movements, the system's simulation of the two technical movements and the representation of the y-direction Euler angle are highly consistent. This proves the accuracy of the system in simulation modelling and shows that its simulation results are applicable to real sports training. The method improves the efficiency and engagement of dance learning through highly realistic visual and motor feedback, while providing researchers with a new way to analyse dance movements and improve teaching methods. However, the method still suffers from high cost, user adaptability problems, and limitations of dance naturalness, and more affordable equipment and software solutions can be further developed in the future.

References

[1] Li Q Y, Li Z H, & Han J (2021). A hybrid learning pedagogy for surmounting the challenges of the COVID-19 pandemic in the performing arts education. *Education & Information Technologies*, pp. 7635-7655. https://doi.org/10.1007/s10639-021-10612-1.

[2] Hsia L H, Hwang G J, & Lin C J (2022). A WSQ-based flipped learning approach to improving students' dance performance through reflection and effort promotion. *Interact Learn Envir*, pp. 229-244. *Interactive Learning Environments*,*1-16.* https://doi.org/10.1080/10494820.2019.1651744.

[3] Gregor S, Vaughan-Graham J, Wallace A, Walsh H, & Patterson K K (2021). Structuring community-based adapted dance programs for person's post-stroke: A qualitative study. *Disabil Rehabil*, pp. 2621-2631. https://doi.org/10.1080/09638288.2019.1708978.

[4] Gao P J, Zhao D, & Chen X A N (2020). Multi-dimensional data modelling of video image action recognition and motion capture in deep learning framework. *IET Image Process*, pp. 1257-1264. https://doi.org/10.1049/iet-ipr.2019.0588.

[5] Teer B (2021). Performance analysis of sports training based on random forest algorithm and infrared motion capture. *Journal of Intelligent & Fuzzy Systems*, pp. 6853-6863. http://dx.doi.org/10.3233/JIFS-189517.

[6] Pfeiffer J, Pfeiffer T, Meissner M, & Weiss E (2020). Eye-tracking-based classification of information search behavior using machine learning: Evidence from experiments in physical shops and virtual reality shopping environments. *Information Systems Research*, pp. 675-691. http://dx.doi.org/10.1287/isre.2019.0907.

[7] Wei M J, Zhang H X, & Fang T Y (2020). Enhancing the course teaching of power system analysis with virtual simulation platform. *International Journal of Electrical Engineering & Education*. https://doi.org/10.1177/0020720920953434.

[8] Li B L, Peng L Y, Gong C B, Chen J R, Zou H L, Luo H Q, & Li N B (2022). Virtual simulation guiding high-risk undergraduate experiments about chemical synthesis of MoS_2 monolayers via a schlenk line. *Journal of Chemical Education*, pp. 3124-3132. https://doi.org/10.1021/acs.jchemed.2c00252.

[9] Hu M Y, Ji J P, Duan J D, & Wang Q (2021). Distributed wind power virtual simulation experiment system for cultivating the ability to solve complex engineering problems. *Computer Applications in Engineering Education*, pp. 1441-1452. https://doi.org/10.1002/cae.22396.

[10] Warland A, Paraskevopoulos I, Tsekleves E, Ryan J, Nowicky A, Griscti J, Levings H, & Kilbride C (2019). The feasibility, acceptability and preliminary efficacy of a low-cost, virtual-reality based, upper-limb stroke rehabilitation device: A mixed methods study. *Disabil Rehabil*, pp. 2119-2134. https://doi.org/10.1080/09638288.2018.1459881.

[11] Maciejewski M, Piszczek M, Pomianek M, & Palka N (2020). Design and evaluation of a steamvr tracker for training applications-simulations and measurements. *Metrology & Measurement Systems*, pp. 601-614. http://dx.doi.org/10.24425/mms.2020.134841.

[12] Yang D, Kim D, & Lee S H (2021). LoBSTr: Real-time lower-body pose prediction from sparse upper-body tracking signals. *Computer Graphics Forum*, pp. 265-275. https://doi.org/10.48550/arXiv.2103.01500.

[13] Qiu S, Zhao H K, Jiang N, Wu D H, Song G C, Zhao H Y, & Wang Z L (2022). Sensor network oriented human motion capture via wearable intelligent system. *International Journal of Intelligent Systems*, pp. 1646-1673. http://dx.doi.org/10.1002/int.22689.

[14] Kumar M S, & Mohan S (2023). Selective fruit harvesting: Research, trends and developments towards fruit detection and localization-A review. *The Journal of Mechanical Engineering Science*, pp. 1405-1444. https://doi.org/10.1177/09544062221128443.

[15] Lin J C, Li S, Qin H, Wang H C, Cui N, Jiang Q, Jian H F, Wang G M (2023). Overview of 3d human pose estimation. *CMES-Computer Modeling in Engineering & Sciences*, pp. 1621-1651. https://doi.org/10.32604/cmes.2022.020857.

[16] Binsch O, Oudejans N, van der Kuil M N A, Landman A, Smeets M M J, Leers M P G, & Smit A S (2022). The effect of virtual reality simulation on police officers' performance and recovery from a real-life surveillance task. *Multimed Tools & Applications*, pp. 17471-17492. https://doi.org/10.1007/s11042-022-14110-5.

[17] Lovanshi M, & Tiwari V (2024). Human skeleton pose and spatio-temporal feature-based activity recognition using ST-GCN. *Multimed Tools & Applications*, pp. 12705-12730. https://doi.org/10.1007/s11042-023-16001-9.

[18] Brock H, Law F, Nakadai K, Nagashima Y (2020). Learning three-dimensional skeleton data from sign language video. *ACM Transactions on Intelligent Systems & Technology*, pp. 30. https://doi.org/10.1145/3377552.

[19] Walters R K, Gale E M, Barnoud J, Glowacki D R, & Mulholland A J (2022). The emerging potential of interactive virtual reality in drug discovery. *Expert Opinion on Drug Discovery*, pp. 685-698. https://doi.org/10.1080/17460441.2022.2079632.

[20] Zheng C, Wu W H, Chen C, Yang T J N, Zhu S J, Shen J, Kehtarnavaz N, & Shah M (2024). Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, pp. 11. https://doi.org/10.1145/3603618.