Construction of 3D Human Motion Capture Model Driven by Multimodal Sensor Fusion

YuanMing YOU[1], Rui DANG[1]
{651554243@qq.com, drui1@cdtu.edu.cn}
[1]School of Computer Engineering, Chengdu Technological University,
No. 1, Section 2, Zhongxin Avenue, Pidu District, Chengdu 611730, Sichuan, China

Abstract. The construction of 3D human motion capture models often relies on a single sensor, resulting in incomplete data and low motion capture accuracy. Therefore, this paper studies a multimodal sensor fusion-driven method for constructing 3D human motion capture models. Data from different types of sensors is integrated, acquisition points are rationally arranged, and Kalman filtering is used to fuse the data to accurately estimate the human motion state. Human posture matching is performed by segmenting the fused data sequence into frames to enhance the capture capability of key spatiotemporal features. The DTW algorithm is then used to accurately measure the similarity of posture sequences to achieve matching. Using the fused data and posture information, a 3D human motion capture model is constructed to accurately capture joint movement changes. Experimental results show that under normal and low light conditions, this method correctly captures more samples than the comparison method for seven types of movements, including grasping. In joint trajectory comparison, the simulated hip, knee, and ankle joint angles are close to the actual values, with an average absolute error of less than 1°. On the IMHD2 multimodal dataset, the accuracy of this method exceeds 97% for movements such as punching, fully demonstrating that the method performs better in motion capture accuracy, joint trajectory simulation, and movement classification, and has high application value.

Keywords: Multimodal sensor fusion driven, 3D model, Human motion capture, 3D technology, Model construction

1 Introduction

Film and animation production, sports training analysis and other fields require accurate capture and modeling of human motion. Traditional methods are limited by their own limitations and cannot fully and accurately acquire information, resulting in many challenges in the construction of 3D human motion capture models.

In existing research, Chi C et al. marked key human body locations, segmented body parts and built models based on this, but relied on predefined methods, and the accuracy was limited due to inaccurate capture of complex and subtle movements[1]. Qiu S et al. used sensor networks to collect

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365270

data and process and analyze it, but the accuracy and stability of the sensors were affected by factors such as installation location and human body shaking, and the collected data had errors, resulting in poor capture accuracy[2]. Moniruzzaman M et al. collected sensor data and used machine learning algorithms to reconstruct and predict human posture, but due to the limitation of training data, the model prediction accuracy dropped significantly when new movements appeared[3]. Niu Z et al. reviewed deep three-dimensional human motion capture methods. Although they introduced a variety of advanced methods, in practical applications, deep learning models faced accuracy problems due to the difficulty in obtaining high-quality labeled data and the black box characteristics that made debugging and optimization difficult[4].

To this end, this article proposes a multimodal sensing fusion driven 3D human motion capture model construction method, which constructs a complete and advanced method system from data fusion, precise pose matching to efficient model construction, significantly improving capture accuracy and effectively solving many key problems of existing methods in this area, providing strong support and promotion for the development and application of related fields.

2 3D Human Motion Capture Model Construction Method

2.1 Multimodal Sensor Motion Capture Data Fusion

In the process of capturing 3D human motion, a single sensor often can only obtain local information of human motion, making it difficult to fully and accurately reflect the complete characteristics of human motion. To solve this problem, this article integrates data from different types of sensors, fully explores the advantages of each sensor data, and obtains information related to human movements. In order to ensure accurate data acquisition, this article reasonably arranges human motion capture data collection points to ensure accurate data acquisition. The specific point arrangement is shown in Table 1.

Table 1: Human motion capture data point arrangement

Sensor Type	**Installation Location**	**Data Collected**
Inertial Measurement Unit (IMU-Accelerometer)	Human limbs and trunk	Acceleration in three directions
Inertial Measurement Unit (IMU-Gyroscope)	Same as accelerometer installation location	Angular velocity in three directions
Optical Sensor Markers	Human joints and key body parts	Three-dimensional spatial coordinates
Electromyography Sensor (EMG)	Surface of major muscle groups	Muscle electrical signal amplitude
Pressure Sensor	Different areas of the sole	Pressure values

Different sensors can capture different aspects of human motion information due to their different working principles and characteristics. However, these data may have noise and error, and

the format and scale of different sensor data may be inconsistent. In order to make full use of the complementarity of these data and improve the accuracy and reliability of the data, this paper uses the correlation and complementarity between different sensor data to carry out Kalman filter fusion. Kalman filter is a commonly used method for dynamic system state estimation, which can obtain more accurate state estimation by predicting and updating the system state in the presence of noise and uncertainty. In multi-mode transmission and capture data fusion, Kalman filter can be used to fuse measurement data of different sensors to obtain more accurate estimation of human motion state [5]. The process is as follows:

Let the state vector collected by the sensor be X_k, representing the human body's motion state at time k, such as position and velocity; the state transition matrix of the system is $F_{k,k-1}$, describing the relationship between state changes from time $k-1$ to k ; the process noise covariance matrix is Q_{k-1}, reflecting the uncertainty in the state transition process [6]. The measurement vector comprises data from various sensors. The measurement matrix Z_k establishes the linear relationship between the state vector and the measurement vector H_k, while the measurement noise covariance matrix R_k describes the noise characteristics during the measurement process. Prediction step of the Kalman filter:

(1)State prediction

$$\hat{X}_{k|k-1} = F_{k,k-1}\hat{X}_{k-1} \tag{1}$$

(2)Covariance prediction

$$P_{k|k-1} = F_{k,k-1}P_{k-1}F_{k,k-1}^{\top} + Q_{k-1} \tag{2}$$

(3)Status update

$$\hat{X}_k = \hat{X}_{k|k-1} + K_k(Z_k - H_k\hat{X}_{k|k-1}) \tag{3}$$

(4)Covariance update

$$P_k = (I - K_kH_k)P_{k|k-1} \tag{4}$$

Through iterative updates in the above steps, more accurate estimates of human motion states are achieved, completing the fusion of multimodal motion capture data.

2.2 Human Pose Matching

Using the fused data alongside specific algorithms and models, the captured human motion data is matched against known human posture templates to accurately identify the current posture. This ensures the model better aligns with actual human movement.

The fused data sequence $\hat{X}_k$ is divided into multiple short-term frames, each with a length of N and a frame shift of M . Each frame undergoes FFT to obtain the time-frequency representation $S(t,f)$, where t denotes time and f denotes frequency. Calculating the energy at each point in the time-frequency plane $|S(t,f)|^2$ reveals the distribution of energy across time and frequency [7]. This

time-frequency analysis method can simultaneously capture the characteristics of human movements in both time and frequency domains, providing richer information for subsequent pose matching.

Self-attention Transformers enhance the model's ability to capture key spatio-temporal features by precisely calculating the correlation between embedded features and assigning different weights to each feature. The specific process can be described by formula (5):

$$Z = Soft\ \max\left(\frac{QK^T}{\sqrt{d_k}}\right)\left|S(t,f)\right|^2 \tag{5}$$

Here, Q denotes the query matrix, and K denotes the key matrix, yielding the weighted representation of embedded features. Subsequently, the DTW algorithm is employed to match collected human motion data with known human pose templates. A distance matrix D is defined for $N_1 \times N_2$, where $D(i,j)$ represents the angular difference between Z_i and G_j.Then, using the Dynamic Time Warping (DTW) algorithm, the collected human motion data is matched with known human posture templates. The DTW algorithm is an algorithm used to measure the similarity between two time series, which can handle time series of different lengths and velocities. The cumulative difference is then computed via dynamic programming:

$$DTW(i,j) = D(i,j) + \min\begin{cases} DTW(i-1,j) \\ DTW(i,j-1) \\ DTW(i-1,j-1) \end{cases} \tag{6}$$

The final similarity between the two pose features can be defined as:

$$S_{DTW}(Z_i,G_j) = \frac{1}{1+DTW(i,j)} \tag{7}$$

Through the DTW algorithm, it is possible to more accurately measure the similarity between different posture sequences, enabling human posture matching. This algorithm is not limited by the length and speed of time series, and can effectively handle the changes in human actions at different time scales, improving the accuracy and robustness of pose matching.

2.3 Building a 3D Human Motion Capture Model

Leveraging fused multimodal data and pose information obtained through human pose matching, a complete, continuous, and realistic 3D human motion model is constructed. This enables comprehensive representation and in-depth analysis of human movements. Compared to two-dimensional representations, three-dimensional human models deliver more realistic visual outcomes, aligning with the visualization requirements of motion capture systems. To enhance the realism of motion simulation, this paper proposes a three-dimensional human motion capture model grounded in three-dimensional modeling techniques [8, 9]. First, human joints are represented as a graph $G = (Z_i, V, E)$, where V denotes nodes (joints) and E denotes edges (skeletal connections). Node features $h_v^{(1)}$ are updated via graph convolutions:

$$h_v^{(l+1)} = \sigma\left(\sum_{u\in\mathcal{N}(v)} \frac{1}{|\mathcal{N}(v)|} h_u^l W^{(l)} + b^{(l)}\right) \tag{8}$$

Where, $\mathcal{N}(v)$ denotes the neighbors of node v, $W^{(l)}$ represents learnable parameters, and $\sigma(\cdot)$ denotes a nonlinear activation function [10]. Through graph convolution operations, nodes can fuse the information of their neighboring nodes, thereby better representing the spatial relationships and motion associations between human joints. The motion state F is modeled via an encoder $q_\phi(z|x)$ and a decoder $p_\theta(x|z)$. The encoder maps the input motion data to a low dimensional latent space and extracts key features of the motion; The decoder remaps the features of the latent space back to the original motion space, generating a continuous motion sequence. The specific formula is as follows :

$$F = -E_{q_\phi(z|x)}[\log p_\theta(x|z)] + h_v^{(l+1)} q_\phi(z|x)\|p(z) \tag{9}$$

Where,$p(z)$ represents the standard normal distribution. Through the aforementioned graph convolutional updates to node features and the encoder-decoder modeling of motion states, this 3D human motion capture model can accurately capture subtle changes in human joint positions during motion sequences.

3 Experiment

3.1 Dataset

To train and validate the model, the IMHD2 multimodal dataset was constructed, which included the following:

(1) Human interaction action annotation: covering 27 types of daily actions (such as grasping, placing, and operating tools), repeated 4 times by 8 subjects (4 males and 4 females), with a total of 861 valid sequences. These daily actions cover various common interaction scenarios of the human body in daily life, and can comprehensively reflect the diversity and complexity of human movements. By repeatedly collecting data from multiple subjects, the diversity and generalization of the data have been increased, which helps to improve the robustness and accuracy of the model.

(2) Multi-view RGB video: 1080P resolution video acquired synchronously to provide visual features. Multi perspective videos can record human movements from different angles, providing richer visual information for models and helping to more accurately capture the posture and motion trajectory of human movements.

(3) IMU measurement data: IMU sensors (sampling rate 100Hz) fixed on the object record rotation and acceleration for motion tracking. IMU sensors can measure the motion state of objects in real-time and accurately, providing important motion data support for analyzing the interaction between the human body and objects.

(4) Object geometry 3D scanning: High-precision laser scanning was used to obtain object models and assist in interactive action modeling. 3D scanning data of object geometry can provide accurate shape and size information of objects, enabling models to simulate the interaction process

between the human body and objects more realistically, improving the accuracy and realism of motion capture.

In the IMHD2 multimodal dataset, 7 types of human actions are set, and 30 sets of sample data are collected for each action type. The samples are evenly divided into training and testing samples, with a total of 15 sets each, for model training and performance evaluation. The complete calculation process for joint angle error is as follows: first, align the reference system, align the joint angle data generated by the model with the high-precision motion capture system data used as a reference in space and time, and ensure that the two are compared in the same coordinate system and time sequence; Next, calibrate the measurement system using known standard action data to eliminate system errors; In the measurement process, the sources of uncertainty mainly include sensor accuracy limitations, such as the existence of certain errors in IMU sensors when measuring acceleration and angular velocity; Environmental factors such as temperature and electromagnetic interference may affect sensor performance; The uncertainty of human movement and the differences in exercise habits and styles among different subjects can also introduce errors. By following the above process and considering relevant uncertainty factors, calculate the error between the simulated and actual values of each joint angle.

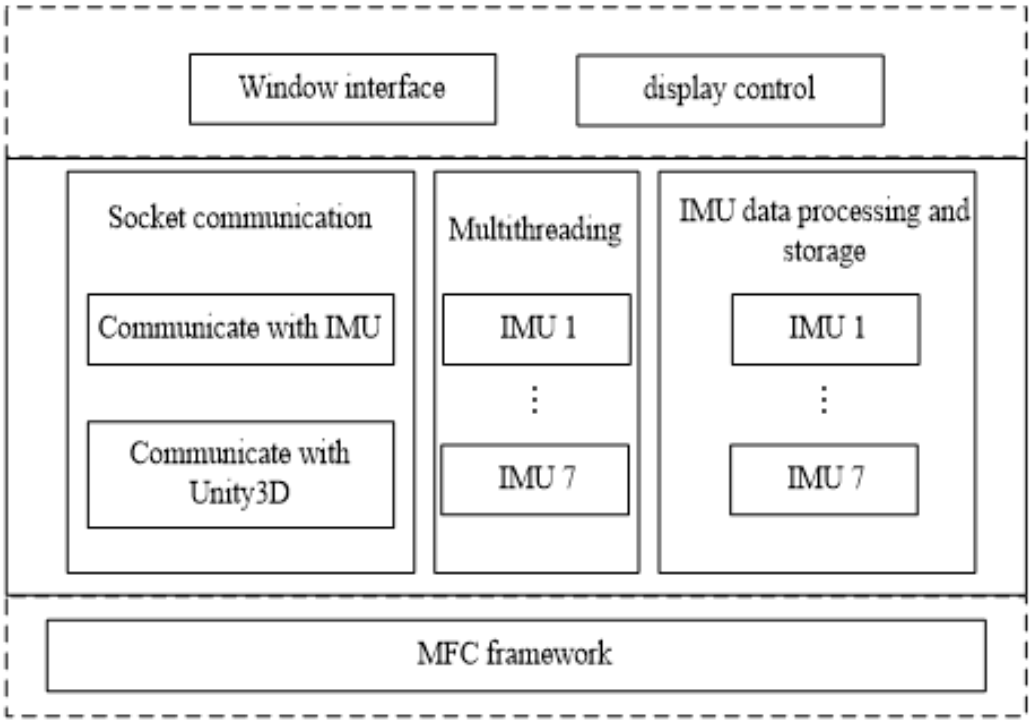
3.2 Experiment Preparation

In order to comprehensively evaluate the performance of the three-dimensional human motion capture system based on multimodal sensor fusion technology, a systematic experimental scheme was designed for the key modules in the complete processing flow, and a dedicated hardware test platform was built. The specific experimental configuration is shown in Table 2:

MFC (Microsoft Foundation Classes), as an object-oriented framework based on C++, encapsulates low-level Windows system calls into a high-order class library, providing predefined window components, dialog controls, and message handling mechanisms. This study uses the Visual Studio 2019 Enterprise Edition development environment to build a TCP communication module for a motion capture system that supports multimodal data access. The software architecture adopts a layered design pattern (as shown in Figure 1):

The layered design pattern divides software systems into different levels, each with clear responsibilities and functions. This design pattern has the characteristics of high cohesion and low coupling, which can improve the maintainability, scalability, and reusability of software. In motion capture system software, the layered design pattern can separate modules such as data acquisition, data processing, model inference, and result display, making each module independent of each other and easy to develop and debug. At the same time, when a module needs to be modified or upgraded, it will not have a significant impact on other modules, reducing the maintenance cost of the system. By modular design, high cohesion and low coupling of each functional component are ensured, providing a stable and reliable software foundation for subsequent experimental verification, while facilitating targeted optimization and functional expansion based on experimental feedback.

-80pt-



80pt

Fig. 1. Software receiving program framework for motion capture model

Table 2: Hardware settings

Hardware Type	Model/Specification	Purpose	Quantity
RGB Camera	Intel RealSense D435i (1080P@30fps)	Collect color video and depth information	1
Inertial Measurement Unit (IMU)	Xsens MTi-680 (Sampling rate 100Hz)	Measure the acceleration and angular velocity of objects/limbs	1 (for object) + 4 (for limbs)
Computing Unit	NVIDIA RTX 3090 GPU + Intel i9-12900K	Real-time processing of multi-modal data and model inference	1
Synchronous	Trigger National Instruments NI-9401	Synchronize the timestamps of the camera and IMU	1
Motion Capture Suit	Xsens MVN Link (17 nodes)	Provide high-precision reference motion data (for comparison)	1
Interactive Objects	Customized tools (cups, hammers, etc. with embedded IMUs)	Simulate daily interaction scenarios	4

3.3 Experimental Results and Analysis

3.3.1 Accuracy of Human Motion Capture

The human motion categories are set to include grasping motion, placement motion, tool operation motion, standing posture, lying posture, walking motion and jumping motion. 30 sets of sample data are collected for each motion type, and these sample data are divided into training samples and test samples, with 15 sets each.

Mix various types of human motion data together as a test dataset to evaluate the method proposed in this paper and the approach proposed in reference [1] A new parametric 3D human body modeling approach by using key position labeling and body parts segmentation(Comparison Method 1), Sensor Network Oriented Human Motion Capture via Wearable Intelligent System proposed in Reference [2] (Comparison Method 2), and Reference [3] Wearable motion capture: Reconstructing and predicting 3d human poses from wearable sensors(Compare the performance of method 3). The tests were conducted in both low light and normal light environments to evaluate the recognition ability of these four methods for human movements under different lighting conditions. The specific test results are shown in Table 3.

As shown in Table 3, under normal lighting, our proposed method correctly captured 28 grasping movements, compared to 22, 20, and 18 for methods $1-3$ respectively; for placement movements, our method captured 29, compared to 23,21, and 19 for methods $1-3$ respectively; the number of correctly captured movements for other movements was also higher for our method. Under low light, our proposed method captured 26 grasping movements, compared to 16, 15, and 14 for

methods $1-3$; the number of correctly captured movements for other movements was also higher than for the three comparison methods. Through testing and comparing various human movements in normal and low light environments, the 3D human motion capture model construction method proposed in this paper performs the best in motion capture accuracy. This is because the method in this paper adopts multimodal sensor fusion technology, which can comprehensively utilize the advantages of different sensors, overcome the limitations of a single sensor under the influence of environmental factors such as light changes, and thus more effectively adapt to the motion capture requirements under different light conditions.

Table 3: Number of Correctly Captured Human Movement Samples

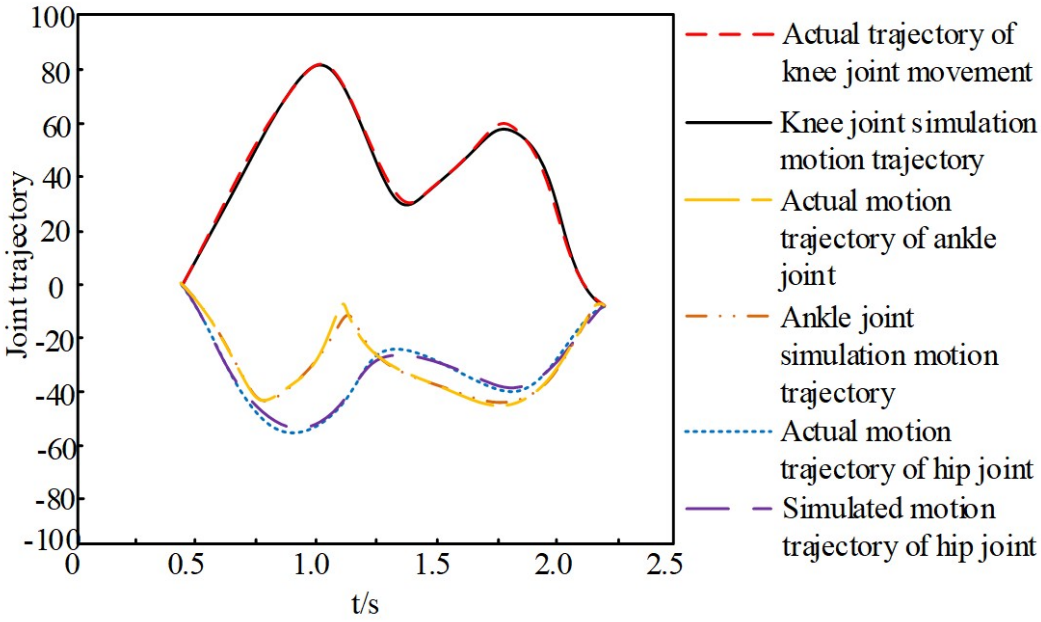
Action Type	Normal Light Environment				Weak Light Environment			
	Proposed Method	Comparison Method 1	Comparison Method 2	Comparison Method 3	Proposed Method	Comparison Method 1	Comparison Method 2	Comparison Method 3
Grasping Action	28	22	20	18	26	16	15	14
Placing Action	29	23	21	19	27	17	16	15
Tool Operation Action	27	20	19	17	25	15	14	13
Standing Posture	30	24	22	20	28	18	17	16
Lying Posture	29	22	20	18	27	16	15	14
Walking Action	28	23	21	19	26	17	16	15
Jumping Action	27	21	19	17	25	15	14	13

3.3.2 Joint Trajectory Comparison

To evaluate the similarity between human movements and the movements executed by the 3D human motion capture model, a comparative analysis of the motion trajectories of each joint was conducted. The relevant joint trajectories are shown in Figure 2.

In the experiment, the angle values of the hip, knee, and ankle joints at different time points were accurately measured and compared. As can be clearly observed from Figure 2, the motion trajectories of the human hip, knee, and ankle joints show a high degree of similarity to the motion trajectories of the corresponding joints in the model, with an average absolute error of less than 1°. This is thanks to the reasonable structural design of the 3D human motion capture model in this article, which uses graph convolution and encoder decoder structure to play an important role. Graph convolution treats joints as graph nodes and connections as edges, allowing nodes to aggregate neighbor information and capture spatial relationships between joints. For example, knee joint motion can also be integrated into feature representations influenced by hip and ankle joints; In the

-140pt-

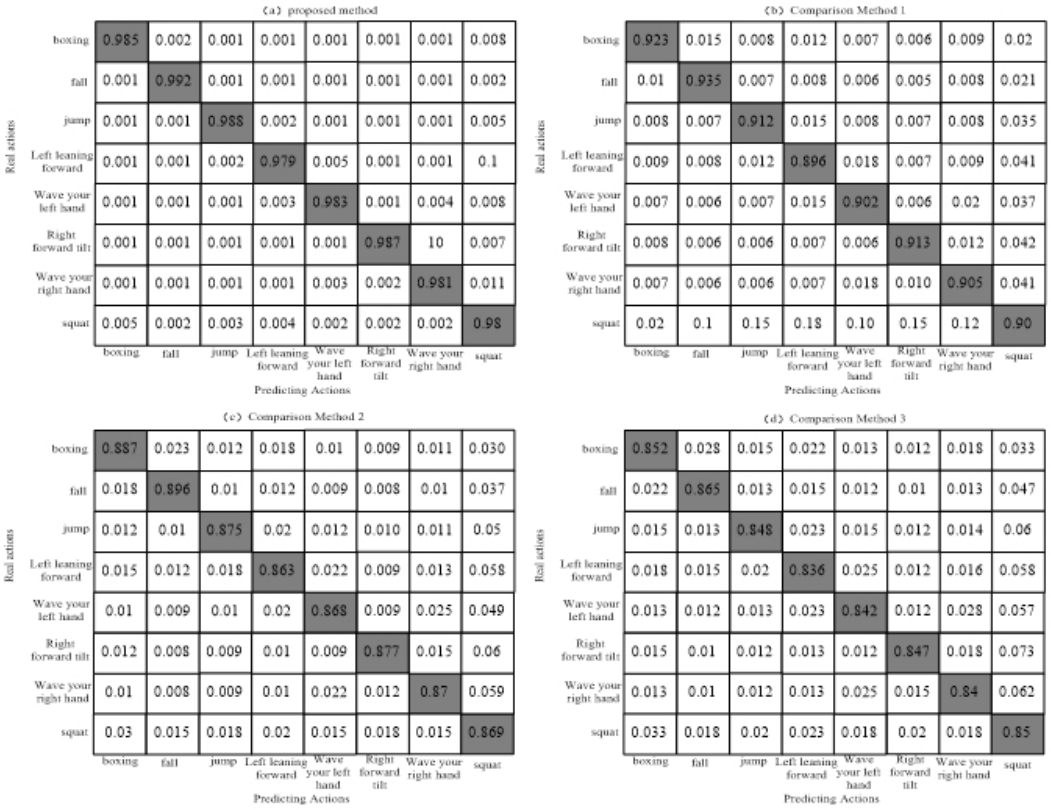


140pt

Fig. 2. Joint Trajectories

encoder decoder structure, the encoder maps motion data to a low dimensional latent space to extract key features and remove redundant information. The decoder then maps it back to the original space to generate a continuous motion sequence, allowing the model to learn human motion patterns and generate realistic trajectories. After a detailed comparative analysis of the trajectories of each joint, the trajectory generated by the model is extremely close to the actual trajectory of the human body, and the simulated values of each joint angle have small errors compared to the real values. This proves that the method is accurate and reliable in simulating human joint motion, and has significant advantages and good prospects in the field of 3D human motion capture.

-150pt-



150pt

Fig. 3. Confusion Matrix

3.3.3 Confusion Matrix

To evaluate the motion capture performance of the model on the IMHD2 multimodal dataset, a series of specific real action categories were determined, including punching, falling, jumping, leaning forward to the left, waving to the left, leaning forward to the right, waving to the right, and squatting. During the experiment, the model was allowed to perform motion capture on the input data, and the captured action categories were compared with the real action categories one by one. The accuracy of each real action being captured as each possible action was calculated, and the cells in the confusion matrix were filled in this way to comprehensively evaluate the model performance. At the same time, the application effects of the method in this paper were compared with those of other methods, and the results are shown in Figure 3.

As can be seen from the confusion matrix data, on the IMHD2 multimodal dataset, for the punching action, the accuracy of our proposed method reaches 98.5%, while the accuracy of comparison methods $1-3$ are 92.3%, 88.7%, and 85.2%, respectively; for the falling action, the accuracy of our proposed method is 99.2%, while the accuracy of comparison methods $1-3$ are 93.5%, 89.6%, and 86.5%, respectively; for other actions such as jumping and leaning left forward, the accuracy of our proposed method is also higher than that of comparison methods $1-3$. In terms of the accuracy of capturing various actions, our proposed method performs outstandingly overall, with an accuracy of generally above 97% for each action, while the accuracy of comparison methods $1-3$ fluctuates significantly across different actions and is generally lower than that of our proposed method. The above data comparison fully demonstrates that the method proposed in this paper can accurately classify captured actions into real categories through deep mining and fusion of multimodal data, with superior performance and strong reliability in practical applications. It can provide accurate recognition results and is expected to play an important role in intelligent security, sports and fitness, medical rehabilitation and other fields, providing strong technical support and having high practical value and application prospects.

4 Conclusion

In this study, a 3D human motion capture modeling approach driven by multimodal sensor fusion is proposed to address the limitations of conventional single-sensor systems, including incomplete data acquisition, environmental susceptibility, and insufficient capture accuracy. By integrating multi-source sensory data, performing precise pose feature matching, and constructing high-fidelity 3D human models, the proposed method enables robust estimation of human motion states and high-precision reconstruction of joint trajectories. Experimental results demonstrate that this approach exhibits strong environmental adaptability under varying illumination conditions and outperforms comparative methods in motion recognition and pose reproduction, thus showing superior robustness and practicality. It can provide dependable technical support for applications such as film and television production, sports analysis, human–computer interaction, and medical rehabilitation. In future work, we will further optimize algorithm efficiency, explore advanced sensor fusion schemes, enhance the generalization capability in complex scenarios, and promote the development of this technology toward higher precision, real-time performance, and intelligence.

Acknowledgments

The authors acknowledge the Second Batch of Modern Industry School in Sichuan Province: "Industry School of Artificial Intelligence Media and Software" (project No. [2023]263).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Chi C, Zeng X, Bruniaux P, et al. A new parametric 3D human body modeling approach by using key position labeling and body parts segmentation. Textile Research Journal. 2022;92(19-20):3653-79.

[2] Qiu S, Zhao H, Jiang N, Wu D, Song G, Zhao H, et al. Sensor network oriented human motion capture via wearable intelligent system. International Journal of Intelligent Systems. 2022.

[3] Moniruzzaman M, Yin Z, Hossain MSB, Choi H, Guo Z. Wearable Motion Capture: Reconstructing and Predicting 3D Human Poses From Wearable Sensors. Journal on Biomedical and Health Informatics (J-BHI). 2023;27(11):12.

[4] Niu Z, Lu K, Xue J, et al. From methods to applications: A review of deep 3d human motion capture. IEEE Transactions on Circuits and Systems for Video Technology. 2024;34(11):9874-95.

[5] Talha M. Research on the use of 3D modeling and motion capture technologies for making sports training easier. Journal of Sport Psychology / Revista de Psicología del Deporte. 2022;31(3).

[6] Tian Y, Zhang H, Liu Y, Wang L. Recovering 3D Human Mesh From Monocular Images: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023;45(12):1-45.

[7] Lin J, Zeng A, Lu S, Cai Y, Zhang R, Wang H, et al. Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset. arXiv preprint arXiv:230711588. 2023.

[8] SenganSudhakar, KumarKailash, SubramaniyaswamyV, RaviLogesh. Cost-effective and efficient 3D human model creation and re-identification application for human digital twins. Multimedia Tools and Applications. 2021.

[9] Peng J, Zhou Y, Mok PY. EHFusion: an efficient heterogeneous fusion model for group-based 3D human pose estimation. The Visual Computer. 2024.

[10] Park J, Jeon IB, Yoon SE, Woo W. Instant Panoramic Texture Mapping with Semantic Object Matching for Large-Scale Urban Scene Reproduction. IEEE Transactions on Visualization and Computer Graphics. 2021;PP(99):1-1.