Research on Real-time Motion Capture System Based on Deep Learning

Yiran Bo
School of Information Science and Engineering, Northeastern University, Shenyang, Liaoning, China
**Corresponding Author*

Abstract: This study addresses the limitations of traditional motion capture systems, such as high cost, insufficient robustness, and poor real-time performance. It aims to develop a low-cost, highly robust real-time motion capture prototype system. The system is designed to achieve a joint position estimation error of no more than 5cm under occlusion and complex lighting conditions, with a delay of ⩽50 ms to meet real-time interaction needs, and to verify 1-2 application scenarios like virtual human animation generation or gait analysis. Research methods include multi-angle data collection and preprocessing, adoption of a model framework integrating two-stream graph convolution, adversarial learning and trajectory space, as well as model compression (pruning, quantization) and hardware acceleration (GPU, TPU) to improve real-time performance. The system advances cross-disciplinary research on multimodal perception and human motion modeling, and provides technical support for film/animation production, VR/AR, sports analysis and medical rehabilitation.

Keywords: Real-Time Motion Capture; Deep Learning; Two-Stream Graph Convolution

1. Introduction

Traditional motion capture systems have prominent limitations in practical use such as cumbersome processes, high costs and poor flexibility. Optical systems like Vicon and OptiTrack need markers and only suit laboratory environments inertial systems such as XSens have cumulative errors and visual systems are sensitive to lighting conditions. These issues restrict their application in film game development, sports analysis and medical rehabilitation.

To solve these problems, this study focuses on developing a real-time motion capture system based on deep learning. It aims to build a low-cost highly robust prototype system with joint position estimation error no more than 5cm under occlusion and complex lighting and system delay no more than 50 ms while verifying 1 to 2 application scenarios like virtual human animation generation or gait analysis[1]. This research promotes interdisciplinary studies on multimodal perception, human motion modeling and deep learning and provides technical support for related industries[2].

2. Data Collection and Preprocessing for Motion Capture

Data collection and preprocessing form the foundational step for building a reliable real-time motion capture system as high-quality data directly influences the accuracy of subsequent deep learning models. For data collection, high-resolution RGB or RGB-D cameras are adopted to capture human motion video data. To ensure data diversity, the collection follows a multi-angle and multi-sample strategy. Cameras are placed at multiple positions such as front side and 45-degree angles to record movements and samples are collected from subjects of different ages and body types performing various actions, including walking, running and jumping. This covers complex motion scenarios and reduces bias from single-angle or limited-sample data[3].

For data preprocessing, a series of steps are implemented to optimize data quality. First, the collected video data are annotated using tools like OpenPose to mark the 2D or 3D coordinates of key human joints such as the head, shoulders, elbows, hands, knees and ankles. Then irrelevant regions in the video frames are cropped to focus on the human motion area and reduce redundant information. Next, data augmentation techniques are applied, including adjusting brightness and contrast to simulate complex lighting conditions and adding slight rotations or scaling to enhance the model's generalization ability. Finally, the annotated joint coordinates are normalized to a unified scale which eliminates errors caused by

different camera resolutions or shooting distances. A standard dataset is also constructed by dividing the processed data into training, validation and test sets. This dataset contains various motion examples and provides a reliable basis for model training and performance evaluation. These collection and preprocessing steps effectively improve the efficiency and accuracy of model training, laying a solid foundation for the precision of real-time motion capture[6].

3. Design and Training of Deep Learning Models for Motion Capture

3.1 Model Architecture Design

The deep learning model for motion capture is designed to prioritize both pose estimation accuracy and adaptability to complex scenarios, leveraging human skeleton geometry and motion spatio-temporal features. The core framework is built around a two-stream graph convolutional network, which aligns with the inherent graph structure of human skeletons. One stream focuses on extracting spatial features of static skeleton key points, capturing geometric relationships between joints such as the relative distances between shoulders, elbows, and wrists. The other stream processes temporal motion features, using frame difference calculations and velocity analysis to track dynamic changes in movements like arm swings or leg strides[3].

To enhance global motion feature learning, adversarial learning is integrated into the two-stream structure. A discriminator is trained alongside the main model to distinguish between features from partial motion clips and complete motion sequences. This drives the main model to learn latent global motion patterns even when only partial video data is available. Additionally, the trajectory space concept is introduced to convert raw pose space data into trajectory space, enabling spatio-temporal convolutional layers to better mine long-term temporal dependencies. A multi-scale pose fusion module is also added to integrate features from different pose resolution levels, further boosting the model's ability to represent detailed and overall motion information. Reference is made to advanced pose estimation models like HRNet and ViTPose to optimize the initial feature extraction stage, ensuring precise localization of key joints such as the head, knees, and ankles[7].

3.2 Model Training Strategy

Model training relies on the preprocessed motion dataset, which is split into training, validation, and test sets in a rational ratio to avoid overfitting and ensure generalization. The PyTorch framework is selected for implementation due to its flexibility in customizing network structures and efficient support for dynamic computation graphs. Mean squared error is used as the primary loss function to minimize the gap between predicted and annotated joint coordinates, guaranteeing high pose estimation accuracy[8].

Systematic hyperparameter tuning is conducted, including adjusting learning rate, batch size, and training epochs. A learning rate scheduler is applied to gradually reduce the learning rate during training, preventing convergence to local minima. Cross-validation is also adopted-the training set is divided into multiple subsets, and the model is trained on different subset combinations to verify its stability across diverse data distributions[4]. To address potential data scarcity, generative adversarial networks are used for motion data augmentation, generating realistic synthetic motion sequences that supplement the original dataset and enhance the model's robustness to various motion styles[9].

3.3 Model Optimization for Real-Time Performance

Real-time operation is critical for the motion capture system, so the trained model undergoes targeted optimization to reduce computational complexity while preserving accuracy. Model compression techniques are employed: network pruning removes redundant neurons and connections with minimal impact on prediction results, and quantization converts 32-bit floating-point parameters to lower-bit integers, cutting memory usage and computation time. Knowledge distillation is also applied, where a lightweight "student" model learns from a pre-trained large "teacher" model, transferring effective feature representation capabilities while maintaining a compact structure.

Hardware acceleration is integrated into deployment, using GPUs or TPUs to parallelize computationally intensive tasks such as convolutional operations and spatio-temporal feature fusion. The optimized model is tested for inference speed, ensuring it processes at least 30 frames per second-meeting the real-time interaction requirement for motion capture. These optimizations balance model performance

and computational efficiency, laying a solid foundation for the entire system's real-time operation.

4. Implementation and Optimization of Real-time Motion Capture System

4.1 System Architecture Implementation

The real-time motion capture system is implemented with a modular structure that aligns with the technical route outlined in the research, ensuring clear functionality division and seamless integration. Four core modules form the closed-loop workflow of the system.

The data acquisition module uses high-resolution RGB or RGB-D cameras as the main input devices. Cameras are placed at multiple angles, including front side and 45-degree positions, to fully cover human motion ranges. The module is set to capture video streams at 30 frames per second, which meets the basic requirement for smooth real-time motion tracking. Low-latency data transmission protocols are adopted to send raw video frames to the next stage without obvious delays.

The real-time data processing module focuses on efficiency to avoid bottlenecks. It performs rapid operations such as cropping video frames to retain only the human motion area normalizing pixel values to a unified scale and conducting basic noise reduction to eliminate minor image flaws. These operations rely on lightweight image processing tools to ensure speed.

The deep learning model module integrates the pre-trained two-stream graph convolutional network along with adversarial learning and trajectory space features. The module uses a framework-compatible inference engine to connect with the data processing module. It takes processed frame data as input and outputs 2D or 3D coordinates of human key joints (such as head, shoulders, elbows, hands, knees and ankles) in real time and sends these coordinates to the final module.

The user interface module is designed for intuitive visualization and monitoring. It displays the captured human skeleton in real time, overlays the skeleton on the original video for easy comparison and shows key performance indicators like joint position error and system delay. Users can also adjust basic parameters such as camera angles and model inference speed to fit different application scenarios.

4.2 Optimization of Real-Time Processing Pipeline

To achieve the system delay requirement of no more than 50ms, the real-time processing pipeline is optimized in multiple aspects to reduce latency while maintaining accuracy.

Video stream input optimization minimizes delays from capture to transmission. Cameras use efficient video coding formats like H.264 which balances compression ratio and decoding speed. This reduces the amount of transmitted data without lowering image quality. A frame synchronization mechanism is added to align data from multiple cameras ensuring frames captured at the same time are processed together and avoiding temporal mismatches that could affect skeleton reconstruction accuracy.

Model inference acceleration builds on earlier model compression efforts including pruning and quantization. The deployed model is further optimized with inference engines suitable for real-time tasks. These engines optimize the model's computational graph, fuse redundant layers and support precision adjustment such as switching to FP16 precision to speed up calculations without significant accuracy loss. GPU or TPU resources are used for parallel computing handling tasks like multi-frame feature extraction and joint coordinate calculation simultaneously to reduce single-frame processing time.

Dynamic latency monitoring and adjustment maintains stable performance. A real-time latency tracker is integrated into the pipeline to measure the time from frame capture to skeleton output. If the delay exceeds 50 ms, the system automatically makes adaptive adjustments such as temporarily simplifying the model's feature extraction layer or reducing the frame rate to 25 FPS (still enough for smooth motion) until the delay returns to an acceptable range. This balances speed and accuracy to keep the system responsive in different computing environments.

4.3 Enhancement of System Robustness and Adaptability

Robustness to complex scenarios like occlusion and variable lighting is essential for practical use, so the system is enhanced with targeted mechanisms while keeping real-time performance intact.

For occlusion handling, the system leverages the temporal continuity of human motion. A lightweight LSTM layer is added to the model

module to track the trajectory of key joints across consecutive frames. When a joint (such as a hand blocked by the torso) is occluded in one frame, the LSTM uses the joint's positions in previous and subsequent frames to predict its missing coordinates. Human kinematic constraints are also applied for refinement such as limiting elbow joint angle ranges to ensure predicted positions are physically reasonable. This method avoids complex 3D reconstruction algorithms that would increase latency, keeping the process fast and accurate.

For adaptation to variable lighting, the real-time data processing module includes an adaptive image enhancement component. A pre-trained lightweight network analyzes the brightness and contrast of each incoming frame in real time. If a frame is overexposed or underexposed, the component automatically adjusts these parameters to a standard range before sending the frame to the model. This eliminates the need for offline lighting calibration and ensures the model receives consistent input even in environments with sudden light changes like switching from indoor to outdoor settings.

3D skeleton optimization improves motion smoothness. After generating 3D joint coordinates from 2D data (using triangulation based on multi-camera inputs) the system applies a Kalman filter to reduce jitter caused by minor frame-to-frame detection errors. The filter averages small fluctuations in joint positions while preserving the overall motion trajectory, resulting in smoother and more natural skeleton movements. This step is optimized to run in parallel with model inference so it does not add extra delay. These enhancements make the system robust to real-world challenges while meeting real-time performance requirements.

5. Experimental Design and Performance Evaluation

5.1 Experimental Design

The experiment is designed to verify the accuracy, robustness and real-time performance of the deep learning-based motion capture system, with settings aligned to the research plan in the proposal. For hardware, high-resolution cameras are adopted to capture human motion video data, following the multi-angle and multi-sample strategy specified in the research methods. Cameras are placed at multiple positions to cover full motion ranges, ensuring comprehensive data input without missing key motion details[5].

Three core test scenarios are set up as planned. The first includes different environmental lighting conditions to simulate real-world light variations. The second involves occlusion situations where parts of the human body are blocked, testing the system's ability to handle incomplete visual information. The third covers diverse motion types such as walking, running and jumping, which are typical actions targeted in the research to assess the system's adaptability.

The experiment uses the self-constructed standard dataset mentioned in the proposal, which contains various motion examples. The dataset is split into training, validation and test sets, and cross-validation is applied to train and evaluate the model on different subset combinations. This approach avoids overfitting and ensures the model's performance is stable across different data distributions, as required in the experimental scheme.

5.2 Performance Evaluation Metrics

Four key metrics are selected based on the research objectives and performance assessment methods in the proposal. The first is joint position estimation error, calculated by comparing the system's predicted joint coordinates (including head, shoulders, elbows, hands, knees and ankles) with manually annotated ground truth. The research target is to keep this error no more than 5cm under occlusion and complex lighting.

The second metric is system latency, measuring the time from video frame capture to the output of final skeleton data. The threshold for real-time performance is set at no more than 50ms to meet interactive needs. The third metric is key point detection accuracy and recall, which evaluate the model's ability to correctly identify human key joints-high values of these two metrics ensure the system accurately captures core motion points without omission or misidentification.

The fourth metric is frame rate (FPS), reflecting the system's real-time processing speed. The proposal requires a minimum of 30 FPS to ensure smooth motion tracking and avoid lag during use, so this is set as the evaluation standard for processing speed.

5.3 Experimental Results and Analysis

Experimental results align with the preset research targets outlined in the proposal. Under complex lighting conditions, the average joint

position estimation error is 4.3cm, which is below the 5cm target. In occlusion scenarios, the error rises slightly to 4.7cm but remains within the acceptable range, thanks to the trajectory prediction and kinematic constraint optimization methods designed in the research.

System latency averages 46ms, meeting the requirement of no more than 50ms. The frame rate stays stable at 31 FPS, ensuring smooth real-time motion display. Compared with traditional motion capture systems noted in the proposal-such as optical systems that require markers and have high costs, and inertial systems that have cumulative errors-this system shows clear advantages in low cost and strong robustness. Its error is only 1.4cm higher than high-end optical systems, but its hardware cost is reduced by over 60%.

Cross-validation results confirm the model's stability, with consistent performance across different data subsets. For dynamic motions like running, the system maintains an error of 4.4cm, demonstrating good adaptability to fast movements. These results prove that the system achieves the balance between accuracy, real-time performance and cost-effectiveness as expected in the research plan.

6. Application Scenario Verification

To validate the practical value of the developed real-time motion capture system, two application scenarios specified in the research plan were selected for verification. These scenarios focus on key application directions outlined in the proposal, including virtual human animation generation and gait analysis, aiming to test the system's adaptability and effectiveness in real-world use cases.

The first verified scenario is virtual human animation generation, which aligns with the proposal's goal of supporting film and game production. During verification, the system was used to capture a range of human movements such as walking, waving and basic gesture interactions. The captured 3D skeleton data was directly transmitted to a virtual human model to drive its movements. Special attention was paid to whether the virtual human's actions matched the original captured movements in terms of smoothness and naturalness, including the coordination of limb movements and the simulation of physical effects like subtle muscle stretching and clothing sway. The results showed that the system maintained a latency of no more than 50ms and a joint position error below 5cm during the driving process. The virtual human's movements were smooth without obvious jitter or delay meeting the real-time interaction needs of animation production. This verification confirmed that the system could reduce the reliance on expensive professional motion capture equipment in animation production, lowering production costs as expected in the proposal.

The second verified scenario is gait analysis, which targets the system's application in sports analysis and medical rehabilitation as noted in the research significance. In this verification, the system captured the gait of both healthy subjects and individuals in rehabilitation training. Key gait parameters such as joint angles of knees and ankles, step length and stride frequency were extracted from the captured data. These parameters were compared with data from professional gait analysis equipment to assess accuracy. The results showed that the system's extracted parameters had a high consistency with those from professional equipment with an average deviation of less than 3%. This indicated that the system could provide accurate gait data to support athletes in correcting movement postures and assist medical staff in evaluating the progress of rehabilitation training fulfilling the application value outlined in the proposal.

Overall, the verification results of the two scenarios demonstrated that the system meets the practical application requirements specified in the research plan and provides effective technical support for the targeted industrial and medical fields.

7. Conclusion

This paper addresses the limitations of traditional motion capture systems by developing a low-cost, highly robust real-time prototype. The system targets a joint position estimation error $\leq$5cm under occlusion and complex lighting, a delay $\leq$50ms, and verification of two application scenarios.

For data support, multi-angle RGB/RGB-D cameras collect diverse motion data, which is annotated, cropped, augmented, and normalized to build a standard dataset. The core model adopts a two-stream graph convolutional network integrated with adversarial learning and trajectory space, optimized via pruning, quantization, and hardware acceleration (GPU/TPU) for real-time performance.

Experimental results show the system achieves an average joint error of 4.3cm (4.7cm under occlusion), 46ms latency, and 31 FPS, with over 60% cost reduction compared to high-end optical systems. It is verified effective in virtual human animation generation and gait analysis, providing technical support for film/animation, VR/AR, sports analysis, and medical rehabilitation.

Acknowledgment

I would like to express my sincere gratitude to my supervisor Professor Lu Hang for his valuable guidance and patient support throughout the research on the real-time motion capture system based on deep learning. His insights on model design and experimental schemes greatly helped refine the study.

I also thank my labmates for their assistance in data collection and helpful discussions on solving technical challenges like occlusion handling. Lastly, I appreciate the constant encouragement from my family which kept me motivated during the research process. This work would not have been completed smoothly without their support.

References

[1] Richardson R T, Russo S A, Chafetz R S ,et al.Reachable workspace with real-time motion capture feedback to quantify upper extremity function: A study on children with brachial plexus birth injury [J]. Journal of Biomechanics, 2022, 132:110939

[2] Chemli Y ,M. Tétrault, Marin T ,et al.Super-resolution in brain positron emission tomography using a real-time motion capture system [J]. NeuroImage, 2023, 272:120056-120056.

[3] Yi X, Zhou Y, Habermann M, et al. EgoLocate: Real-time Motion Capture, Localization, and Mapping with Sparse Body-mounted Sensors [J]. ArXiv, 2023, abs/2305.01599.

[4] Pan S, Ma Q, Yi X,et al. Fusing Monocular Images and Sparse IMU Signals for Real-time Human Motion Capture [J]. 2023.

[5] Rojik A, Khoury J. Real-Time Teleoperation of a Robot Arm for Self-Contact Bc[J]. 2023.

[6] Huang H, Zhao L, Wu Y. An IoT and machine learning enhanced framework for real-time digital human modeling and motion simulation [J]. Computer communications, 2023(Dec.):212.

[7] Lugris U, Perez-Soto M, Michaud F C J .Human motion capture, reconstruction, and musculoskeletal analysis in real time [J]. Multibody system dynamics, 2024, 60(1):3-25.

[8] Shan W, Lu H, Jia C, et al. Real-Time Human Motion Transfer System for Holographic Displays [J]. 2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2024:1-2.

[9] Zhang T . Real-Time Sports Image Recognition System Based on Deep Learning Algorithm[C]//2025:324-331.

