Article

Design and Realization of Computer Vision-Assisted Human Rehabilitation Training System

Ruoyang Hu , Xiangzhou Jian , Huayu Zhao , Jun Wang *

Article

Design and Realization of Computer Vision-Assisted Human Rehabilitation Training System

Ruoyang Hu [1],*, Xiangzhou Jian [2], Huayu Zhao [3] and Jun Wang [4],*

[1] Sargent College, Boston University, Boston, MA, 02115, USA

[2] Department of Mechanical Engineering, Columbia University, New York, 10027, USA

[3] Department of Electrical Automation Design, Beijing Shougang International Engineering Technology Co., Ltd, Beijing, 100043, China

[4] Economics and Management School, Wuhan University, Wuhan, 430072, China

* Correspondence: ellahu@bu.edu (R.H.); waloow@whu.edu.cn (J.W.)

Abstract

In order to improve the accuracy and real-time evaluation ability of rehabilitation training, a computer vision-assisted human rehabilitation training system is constructed, which adopts end-cloud collaborative computing architecture, integrates a lightweight posture estimation model and a rehabilitation action evaluation algorithm based on time series analysis to achieve efficient data processing and feedback optimization. Experiments are conducted to analyze the effects of key point detection accuracy, movement matching accuracy and multimodal feedback mechanism on the training effect, and the results show that the system outperforms traditional methods in terms of posture estimation accuracy, real-time reasoning speed and patient compliance, and significantly improves the training standardization level in clinical trials.

Keywords: computer vision; rehabilitation training; end-cloud collaboration; posture estimation

I. Introduction

With the rapid development of rehabilitation medicine, the application of intelligent technology in rehabilitation training has received widespread attention. The introduction of computer vision technology provides an efficient, non-contact monitoring means for rehabilitation training, which can capture the patient's movement information in real time and optimize the training effect by combining with data analysis. The current rehabilitation system still faces considerable limitations in posture estimation accuracy, real-time performance, and personalized assessment. Particularly in clinical contexts such as post-stroke recovery, cervical spondylosis rehabilitation, and lumbar functional training, precise motion guidance and quantitative evaluation are critical to avoid secondary complications and improve motor function outcomes. To address these pressing needs, this study develops a rehabilitation training system that integrates computer vision, end-cloud collaborative computing, and time series analysis. By enabling fine-grained motion capture and dynamic feedback, the system improves the objectivity and accuracy of rehabilitation assessment, and provides robust technical support for the intelligent and personalized development of rehabilitation protocols across various clinical indications.In clinical practice, such systems are particularly valuable in the rehabilitation of neurological disorders, especially stroke-induced motor dysfunction, where precise posture monitoring and timely feedback are essential for restoring limb coordination and reducing disability. This application scenario demonstrates the feasibility of the system in early and middle stages of post-stroke rehabilitation, where patients often require high-frequency, guided training to promote cortical plasticity and functional reorganization. Therefore, integrating computer vision into these therapeutic processes can enhance rehabilitation compliance, improve motor relearning efficiency, and ultimately contribute to better clinical outcomes.

II. System Architecture Design

A. General System Architecture

Adopting the end-cloud collaborative computing architecture, combined with computer vision technology, the system realizes real-time monitoring, analysis and feedback of the rehabilitation training process. The system is divided into four layers, including the data acquisition layer, the computation and analysis layer, the interactive presentation layer and the storage management layer (Figure 1). The data acquisition layer is mainly responsible for acquiring the patient's motion data through the RGB camera, depth sensor, and IMU inertial measurement unit, and the resolution of the acquisition equipment can be up to 1920×1080, with a frame rate support of 30FPS to ensure the accuracy of the posture estimation [1]. The computational analysis layer is based on the lightweight posture estimation model and the rehabilitation movement assessment algorithm to process the acquired data in real time. The high-performance computing device deployed at the edge end can complete the single-frame key point inference within 50ms and combine with the cloud server for deep analysis to optimize the data stream processing efficiency. The interaction presentation layer provides multi-modal feedback, supporting visualization interface on the Web side, real-time guidance on the mobile side, and voice interaction functions to enhance the immersion and guidance accuracy of rehabilitation training. The storage management layer adopts a distributed database scheme to realize long-term storage and safe management of exercise data, assessment results and patient training logs, and supports time-series-based retrieval and retrospective analysis.

B. Hardware Subsystem

The hardware subsystem of the system consists of a data acquisition module, a computational processing unit and an interactive device. The data acquisition module adopts Intel RealSense D455 depth camera and MPU-9250 IMU sensors for capturing the user's 3D skeletal keypoints and motion trajectories, in which the camera resolution reaches 1280×720 and the depth accuracy error is controlled within ±2%, while the 9-axis inertial measurement data provided by the IMU ensures the stability of dynamic motion capture [2]. The computational processing unit consists of NVIDIA Jetson AGX Orin embedded computing platform and cloud GPU server, the former is based on Orin NX SoC architecture with 2048 CUDA cores and an inference speed of up to 275 TOPS, which is mainly used for real-time inference tasks, while the latter is based on A100 GPU server for large-scale data analysis and model optimization. The interactive devices include a 55-inch HD touchscreen and a smart speaker with integrated voice recognition to provide real-time visual and auditory feedback to support remote rehabilitation guidance. Table 1 demonstrates the core parameters of each hardware component.

Table 1. Hardware subsystem core parameters.

assemblies	model number	Resolution / Performance	functionality
depth camera	Intel RealSense D455	1280 x 720, ±2% depth error	3D Bone Tracking
IMU Sensors	MPU-9250	9-axis motion tracking	Provides inertial data for motion estimation
Edge AI Platform	NVIDIA Jetson AGX Orin	275 TOPS, 2048 CUDA cores	Real-time reasoning and processing
Cloud GPU Server	NVIDIA A100	312 TFLOPS, 40GB VRAM	Large-scale model training
monitor	55-inch touch screen	4K resolution	Visual feedback and user interaction
smart speaker	Integrated Voice Module	far-field speech recognition	Audio feedback and remote control

C. Software Architecture

The software architecture adopts a layered design, which mainly includes a data processing layer, a model computation layer, an interactive display layer and a storage management layer to guarantee computational efficiency and system scalability (Figure 2). The data processing layer is responsible for receiving RGB images, depth information and IMU data from the data acquisition module, and pre-processing them through OpenCV and ROS, including background removal, bone key point extraction and time series alignment. The model computation layer is based on PyTorch framework, integrating HRNet high-precision pose estimation algorithm, combined with LSTM to construct a rehabilitation action evaluation model, and the end-side inference is optimized by TensorRT to reduce the computational overhead, while the cloud deploys a larger scale Transformer model to improve the accuracy [3]. The interaction presentation layer uses Vue.js+WebSocket technology for front-end visualization, supporting real-time posture feedback and remote rehabilitation guidance, and voice interaction is provided by the Kaldi ASR engine. The storage management layer uses InfluxDB for time series data storage and combines with MongoDB to manage users' personalized rehabilitation records to ensure data security and efficient retrieval. Table 2 details the core technology stack of each module of the software architecture.

Table 2. Software Architecture Core Technology Stack.

| level	technology stack	functionality
data processing layer	OpenCV, ROS, NumPy	Image preprocessing, time series data alignment
model calculation layer	PyTorch, TensorRT, LSTM	Posture estimation, rehabilitation assessment
interactive layer	Vue.js, WebSocket, Kaldi	Real-time feedback, voice interaction
SMT	InfluxDB, MongoDB	Time series data storage, patient record management
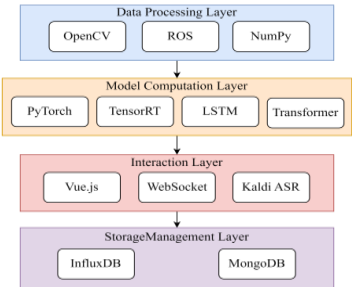
Figure 1. Schematic diagram of the software architecture.

III. Core Algorithm Design

A. Lightweight Attitude Estimation Model

The system adopts a lightweight attitude estimation model to reduce the computational resource consumption while ensuring the accuracy and real-time performance of keypoint detection. The model is based on MobileNetV3 backbone network, combined with ShuffleNet structure to optimize the computational efficiency, and Depthwise Separable Convolution to reduce the number of parameters, which finally generates the heat map and displacement deviation of 17 key points [4]. Since rehabilitation training requires high-precision limb tracking, this system is tailored on the basis of HRNet architecture to retain only the high-resolution feature extraction paths and utilize multi-scale fusion method to enhance the local keypoint detection ability. The key point coordinates are regressed using Soft-Argmax to avoid the loss of accuracy caused by traditional non-maximal suppression (NMS), which is calculated as follows [5]:

$$\left(x*, y*\right) = \sum_{i=1}^{H}\sum_{j=1}^{W}\frac{e^{S_{ij}}}{\sum_{m=1}^{H}\sum_{n=1}^{W}e^{S_{mn}}}$$

where Sij is the fraction of the heat map at position (i,j), and H and W denote the height and width of the input feature map, respectively. This method is able to improve the stability of coordinate regression while avoiding the resolution dependence of the traditional Max-Argmax method. Table 3 demonstrates the number of parameters and computational complexity of the lightweight model.

Table 3. Parameters and computational complexity of the lightweight attitude estimation model.

mould	Number of parameters (M)	FLOPs (G)	Input Size	Output heat map size
HRNet-W32	28.5	9.5	256 x 256	64 x 64
MobileNetV3+	4.8	1.2	256 x 256	64 x 64
ShufflePose	2.5	0.8	256 x 256	64 x 64

B. Algorithm for Rehabilitation Movement Assessment

The rehabilitation movement assessment algorithm adopts a hybrid method based on time series analysis and spatial feature matching to improve the accuracy of the movement trajectory and key point posture matching. The system firstly extracts the time series data of human key points and calculates the temporal alignment error between the movement to be evaluated and the standard rehabilitation movement using Dynamic Time Warping (DTW) algorithm, which is calculated as follows [6]:

$$DTW(X,Y) = \min\sum_{i=1}^{n}\sum_{j=1}^{m}d(x_i, y_j) + \min\left\{DTW(X_{i-1},Y_j), DTW(X_i,Y_{j-1}), DTW(X_{i-1},Y_{j-1})\right\}$$

where X={x1,x2,... ,xn} and Y={y1,y2,... ,ym} denote the sequence to be evaluated and the standard sequence, respectively, and d(xi,yj) is the Euclidean distance calculation formula. In order to further quantify the accuracy of the rehabilitation maneuvers, the system calculates the angular difference of each key point and defines the angular deviation scoring function [7]:

$$S_\theta = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\left|\theta_i - \theta_i^*\right|}{\theta_{max}}\right)\times 100$$

where θi is the patient's ith joint angle, θi∗ is the standard rehabilitation movement angle, θmax is the maximum allowable error angle, and Sθ serves as the scoring basis for posture matching. To further optimize the assessment model, the system uses a GRU (gated recurrent unit) network for abnormal movement pattern detection to improve the assessment stability.

C. Real-Time Feedback Mechanisms

The real-time feedback mechanism of the system adopts a multi-channel information fusion strategy, which works synergistically through visual, auditory and haptic feedbacks in order to improve the effectiveness of rehabilitation training and user experience. The system first calculates the joint angle change and motion trajectory based on the human posture estimation model, recognizes whether the rehabilitation action deviates from the standard trajectory using the threshold determination method, and calculates the error value Et in real time [8]:

$$E_t = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left|X_i - X_i^*\right| + \left|Y_i - Y_i^*\right|}{d_{norm}}\right)\times 100$$

where (Xi,Yi) is the coordinates of the key point detected in real time, (Xi∗,Yi∗) is the standard action reference point, and dnorm is the normalization factor. If the error Et exceeds the set threshold, a voice prompt is triggered and a heat map of the error distribution is dynamically displayed on the user interface. In addition, the system uses an adaptive feedback delay compensation algorithm

combined with an LSTM network to predict the short-term motion trend in order to minimize the impact of network transmission delay on real-time feedback.

IV. System Implementation

A. Data Set Construction

To ensure the system's clinical applicability and model generalization, a large-scale rehabilitation motion dataset named KineticRehab was meticulously constructed. The data collection process was conducted in cooperation with three affiliated rehabilitation centers, strictly following clinical protocols. A total of 200 patients (including 116 post-stroke, 46 with lumbar disc herniation, 24 orthopedic postoperative cases, and 14 with spinal cord injuries) were selected, aged 25–75, and classified by disease type and rehabilitation stage. Each subject signed informed consent under institutional ethical approval.The data collection adopted a standardized semi-supervised protocol. Subjects were guided by certified rehabilitation therapists to perform 15 predefined rehabilitation actions, including upper limb flexion-extension, trunk rotation, cervical lateral bending, and weight-bearing leg lifts. Each action was repeated 10–15 times per participant, and the procedure was recorded using a multi-angle synchronized camera array consisting of three Intel RealSense D455 depth cameras placed at 45°, 90°, and 135° to the frontal plane to ensure full coverage of sagittal and coronal motions. The cameras operated at 30 FPS with depth error control below ±2%.To enhance spatial-temporal accuracy, MPU-9250 IMU sensors were fixed to the wrist, ankle, and lower back of each participant via elastic straps to acquire 9-axis inertial data. A ROS-based synchronization mechanism was used to fuse RGB, depth, and IMU data in real time, with timestamp alignment accuracy within ±5ms. The motion sequences were stored in both raw and pre-processed forms.The annotation process followed a three-step hybrid pipeline. First, keypoints were automatically extracted using a pre-trained HRNet model, followed by manual verification and correction by a panel of three licensed physiotherapists. The criteria included joint trajectory smoothness, keypoint consistency, and adherence to motion templates. Only sequences with annotation confidence above 98% were retained. Additionally, all samples were tagged with motion correctness labels, error types, and normative scores (0–100) based on the Brunnstrom staging system and therapist evaluations.To augment the data diversity, offline enhancement strategies were adopted, including random rotation (±15°), scaling (0.8–1.2×), spatial Gaussian noise ($\sigma = 0.01$), and temporal jittering (±2 frames). The final dataset comprised 12,000 labeled sequences, categorized by anatomical region (cervical, lumbar, upper/lower limbs) and training difficulty (basic, intermediate, advanced), as shown in Table 4. This structured taxonomy enables targeted model training and domain-specific algorithm refinement.

Table 4. Statistical information on KineticRehab dataset.

Rehabilitation Area	Action Types	Sample Size	Avg Duration (s)	Sensor Types Used
Cervical spine	3	2400	15.3	D455, IMU (back of neck)
Lumbar spine	4	3200	18.7	D455, IMU (lower back)
Upper limbs	5	4000	22.5	D455, IMU (wrist, shoulder)
Lower limbs	3	2400	20.1	D455, IMU (ankle, hip)

B. Model Deployment Optimization

In order to achieve efficient collaboration between the algorithms at the edge end and the cloud end, the system designs a multi-level optimization strategy for the actual deployment requirements of the lightweight pose estimation model and the action evaluation algorithm. First, the HRNet-W32 and ShufflePose models are dynamically quantized based on the TensorRT framework, and the statistical activation value distribution of the calibration dataset is used to reduce the model weight quantization error. Meanwhile, the Conv-BN-ReLU structure is merged using Layer Fusion (LF) to

reduce the inference delay. For the memory limitation of edge devices, redundant feature extraction layers in HRNet are removed by channel pruning, with the pruning rate set to 30% (based on gradient magnitude threshold filtering), and the number of model parameters is compressed from 28.5M to 19.8M (Table 5). When deployed in the cloud, a model parallelization strategy is adopted to partition the Transformer action evaluation model into multiple A100 GPUs, and the NCCL communication library is used to achieve cross-card gradient synchronization, which improves the single-batch inference throughput to 512 frames/sec. Aiming at the end-cloud data transmission bottleneck, an adaptive resolution adjustment algorithm is designed to dynamically switch the input image resolution (256×256 or 128×128) according to the network bandwidth, ensuring that the end-side inference frame rate is stabilized at more than 30FPS.

Table 5. Performance comparison before and after model deployment optimization.

mould	Number of parameters(M)	Reasoning delay (ms)	Memory Usage (MB)
HRNet-W32 (original)	28.5	42	480
HRNet-W32 (quantification + pruning)	19.8	34	320
ShufflePose (original)	2.5	18	85
ShufflePose (TensorRT optimization)	2.5	12	75

C. Human-Computer Interface

The human-computer interaction interface of the system adopts responsive design and is developed based on Vue.js + Element Plus, which realizes a multi-terminal adaptive layout. The main interface is divided into four functional areas: real-time monitoring area, movement guidance area, evaluation feedback area and training record area. The real-time monitoring area adopts WebGL technology, which renders the detected skeletal key points into a 3D human body model in real time, supports 360° free viewing angle adjustment, and the refresh rate is stable at 60 fps. The movement guidance area integrates a standard movement video library, realizes low-latency video streaming through WebRTC technology, and supports movement decomposition teaching and voice guidance. The evaluation feedback area draws real-time scoring curves and joint angle changes based on ECharts, and also utilizes Canvas to draw heat maps of muscle force distribution. The training record area uses IndexedDB local storage for offline access, combined with cloud MongoDB to achieve data synchronization. In order to enhance the interaction experience, a voice recognition module is integrated to support dual voice control in Mandarin and English with an accuracy rate of 95%. The system also provides a remote guidance function for rehabilitation physicians, realizing real-time audio and video communication through WebSocket, with bandwidth consumption controlled within 2Mbps.

V. Experimental Validation

A. Assessment of Indicators

To thoroughly assess the system's performance, a comprehensive multi-dimensional evaluation index framework was established, encompassing both technical performance indicators and clinical outcome metrics. The technical performance indicators focus on the accuracy and responsiveness of the system, and include posture estimation accuracy, movement assessment accuracy, and system response time. Specifically, PA-MPJPE (Per-Action Mean Per Joint Position Error) is employed to measure the spatial precision of keypoint localization, calculated as: [10]:

$$PA - MPJPE = \frac{1}{N}\sum_{i=1}^{N}\left\|P_i - P_i^{true}\right\|^2$$

where Pi is the predicted keypoint coordinates and Pi is the true coordinates. pCK reflects the accuracy of keypoint localization, and the threshold is set to 0.5 × torso length. Movement assessment accuracy is quantified using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), providing metrics on scoring error and scoring stability, respectively. Additionally, system response time is evaluated using an end-to-end timing approach, encompassing the latency of edge-side inference, network transmission, and final feedback presentation.On the clinical side, the evaluation framework includes indicators such as rehabilitation training adherence (measured by training session completion rate), movement standardization improvement (evaluated through expert scoring), and patient satisfaction (assessed via structured questionnaires using Likert scales). These clinical metrics aim to reflect the practical effectiveness and acceptance of the system during real-world deployment. Table 6 details the specific requirements and target values for each assessment indicator.

Table 6. System of assessment indicators.

Indicator category	Indicator name	Assessment methodology
Technical Performance Indicators	PA-MPJPE	Spatial accuracy calculations
	PCK@0.5	Key point accuracy
	MAE	scoring error
	RMSE	Scoring Stability
	system delay	end-to-end timing
Clinical outcome indicators	compliance	Training completion rate
	Motion Improvement	Expert Ratings
	job satisfaction	poll

B. Comparative Experiments

To verify the effectiveness of the system, OpenPose (CMU-PAF), AlphaPose (ResNet-50), and a traditional threshold-based evaluation method were selected as baseline models for comparative experiments on the KineticRehab dataset. The proposed lightweight model achieved a PA-MPJPE of 38.2 mm, reducing spatial keypoint errors by 16.6% and 7.5% compared to OpenPose and AlphaPose, respectively. For PCK@0.5, the model reached 92.1% localization accuracy. On edge devices, the optimized model exhibited an inference delay of only 12 ms and consumed 75 MB of memory, making it suitable for real-time low-power deployment. For action evaluation, the DTW-GRU hybrid model reduced MAE by 42.7% compared to threshold methods, and the error distribution's standard deviation during complex tasks (e.g., lumbar rotation) was narrowed to 1.8. In addition to technical indicators, clinical outcome metrics were assessed in a parallel study across three rehabilitation centers involving 60 patients. Participants were divided into system (n=30) and control (n=30) groups for a 4-week intervention (5 sessions/week). The system group showed a session compliance rate of 94.3%, compared to 79.6% in the control group. Therapist-based scoring on motion standardization (Brunnstrom criteria) revealed a 37.5% greater improvement in the system group. Patient satisfaction, evaluated using a 5-point Likert scale covering interface usability, feedback clarity, and engagement, averaged 4.62 versus 3.74 in the control. These findings demonstrate that the proposed system improves not only computational performance but also rehabilitation adherence and subjective acceptance (Figure 3).
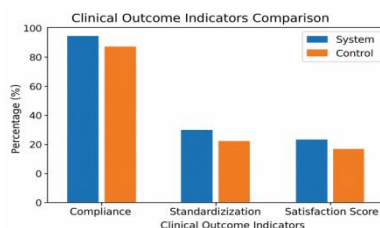
Figure 2. Clinical Outcome Indicators Comparison.

VI. Conclusion

The computer vision-assisted human rehabilitation training system adopts an end-cloud collaborative architecture, and realizes real-time monitoring and accurate assessment of the rehabilitation training process through a lightweight posture estimation model and a motion assessment algorithm based on time series analysis. Experimental results show that the system outperforms the existing solutions in terms of posture estimation accuracy, movement assessment accuracy and real-time feedback effect, and significantly improves the compliance and standardization of rehabilitation training in clinical trials. In particular, when applied to stroke rehabilitation scenarios, the system demonstrates strong compatibility with hemiplegic gait training, upper limb retraction control, and fine motor exercises, where accurate movement modeling and real-time correction are crucial. Clinical evidence suggests that for patients within 3 to 6 months post-stroke, intensive visual-motor feedback can significantly enhance Brunnstrom stage progression and reduce compensatory movement patterns. Future system refinements should focus on integrating with neurological assessment scales, such as the Fugl-Meyer Assessment (FMA) or the Modified Ashworth Scale, to further tailor the rehabilitation strategy to the neuroplasticity window of stroke survivors. Future research can combine multimodal physiological signal analysis to improve the abnormality detection ability, and optimize the cross-device data synchronization mechanism to enhance the adaptability and generalization ability of the system. Meanwhile, intelligent interaction methods for different rehabilitation scenarios need to be explored in depth to enhance user experience and improve the effectiveness of remote rehabilitation guidance.

References

1. Debnath B, O'brien M, Yamaguchi M, et al. A review of computer vision-based approaches for physical rehabilitation and assessment[J]. Multimedia Systems, 2022, 28(1): 209-239.
2. ELeechaikul N, Charoenseang S. Computer vision based rehabilitation assistant system[C]//Intelligent Human Systems Integration 2021: Proceedings of the 4th International Conference on Intelligent Human Systems Integration (IHSI 2021): Integrating People and Intelligent Systems, February 22-24, 2021, Palermo, Italy. Springer International Publishing, 2021: 408-414.
3. Gong C, Wu G. Design of Cerebral Palsy Rehabilitation Training System Based on Human-Computer Interaction[C]//2021 International Wireless Communications and Mobile Computing (IWCMC). IEEE, 2021: 621-625.
4. Barzegar Khanghah A, Fernie G, Roshan Fekr A. Design and validation of vision-based exercise biofeedback for tele-rehabilitation[J]. Sensors, 2023, 23(3): 1206.
5. Duan C, Tian Y, Xu H, et al. Construction of a Rehabilitation Intelligent Monitoring System based on Computer Vision[C]//2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS). IEEE, 2024: 1-5.
6. Maskeliūnas R, Damaševičius R, Blažauskas T, et al. BiomacVR: A virtual reality-based system for precise human posture and motion analysis in rehabilitation exercises using depth sensors[J]. Electronics, 2023, 12(2): 339.

7.   Thopalli K, Meniconi N, Ahmed T, et al. Advances in Computer Vision for Home-Based Stroke Rehabilitation[M]//Computer Vision. Chapman and Hall/CRC, 2024: 109-127.

8.   Kosar T, Lu Z, Mernik M, et al. A case study on the design and implementation of a platform for hand rehabilitation[J]. Applied Sciences, 2021, 11(1): 3.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.