Computer Systems Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.037865

ARTICLE

Movement Function Assessment Based on Human Pose Estimation from Multi-View

Lingling Chen[1,2,*], Tong Liu[1], Zhuo Gong[1] and Ding Wang[1]

[1]School of Artificial Intelligence, Hebei University of Technology, Tianjin, 300400, China

[2]Intelligent Rehabilitation Device and Detection Technology Engineering Research Centre of the Ministry of Education, Tianjin, 300400, China

*Corresponding Author: Lingling Chen. Email: chenling@hebut.edu.cn

Received: 18 November 2022 Accepted: 17 February 2023 Published: 19 March 2024

ABSTRACT

Human pose estimation is a basic and critical task in the field of computer vision that involves determining the position (or spatial coordinates) of the joints of the human body in a given image or video. It is widely used in motion analysis, medical evaluation, and behavior monitoring. In this paper, the authors propose a method for multi-view human pose estimation. Two image sensors were placed orthogonally with respect to each other to capture the pose of the subject as they moved, and this yielded accurate and comprehensive results of three-dimensional (3D) motion reconstruction that helped capture their multi-directional poses. Following this, we propose a method based on 3D pose estimation to assess the similarity of the features of motion of patients with motor dysfunction by comparing differences between their range of motion and that of normal subjects. We converted these differences into Fugl–Meyer assessment (FMA) scores in order to quantify them. Finally, we implemented the proposed method in the Unity framework, and built a Virtual Reality platform that provides users with human–computer interaction to make the task more enjoyable for them and ensure their active participation in the assessment process. The goal is to provide a suitable means of assessing movement disorders without requiring the immediate supervision of a physician.

KEYWORDS

Human pose estimation; 3D pose reconstruction; assessment of movement function; plane of features of human motion

1 Introduction

The number of people suffering from motor dysfunction due to traffic accidents, strokes, and cerebral thrombosis has increased significantly in recent years. The mobility of the human body in general declines with age, and this can lead to a variety of diseases. According to one survey, neurological disorders pose a significant complication and this context, and are a major cause of disabilities among middle-aged and elderly people [1]. Movement disorders are dominant among such disabilities, and affect the patient's ability to perform daily activities such that they cannot live on their own. In light of this, providing automated methods of assessment for the increasing number of patients

with motor dysfunctions and reducing the burden on medical personnel has become an important area of research in modern medicine.

With rapid advances in research on the problems of classification [2,3] and optimization [4] in computer vision in recent years, human pose estimation has been widely used for medical assistance and motion analysis as well as in Virtual Reality technology [5–7]. The main objective is to recover the parameters of pose of the target body and analyze its motion based on the relevant sequences of images. Early work in the area focused on two-dimensional (2D) human pose estimation, and involved recovering 2D poses from images or videos by using either of two general approaches: top-down and bottom-up methods. The top-down method of identifying pose involves first identifying the position of each person in the given image through a target detection network and then estimating their pose. This method is highly accurate because it can leverage the poses of a single subject, but its speed of inference is low because it relies on the target detection network [8,9]. The bottom-up method of pose estimation involves first identifying the joints of all people in the given image through a detection network and then linking the joints belonging to the same person by using a clustering algorithm [10]. The network can directly estimate the joints of all people in the image. Although the accuracy of the bottom-up approach is lower than that of the top-down approach, it can make faster inferences [11]. Networks used to identify 2D poses that are based on deep learning have delivered good performance [12,13]. Subsequent research has focused on 3D pose estimation.

Research on 3D pose estimation can be classified into two types: single-view and multi-view pose estimation. Single-view pose estimation is generally used to locate the 2D joints of the human body in cropped images to convert the detected 2D pose into three dimensions through a learning-based approach [14,15]. While this method is accurate, it is too heavily reliant on the 2D pose detector, and can regress to the 3D pose of the body in the image [16,17]. It is a simple and fast end-to-end approach that can, however, suffer from the problem of ambiguities in the obtained pose. Moreover, the accuracy of single-view 3D pose estimation is far lower than that of multi-view estimation.

Initial research on multi-view pose estimation was based on using 2D features obtained from multi-view image sequences to reconstruct 3D poses [18,19]. Deep learning-based 2D detectors combined with statistical models of human motion have recently been developed, and have delivered impressive results. Joo et al. compared fitted 3D models with the true values of the corresponding images in the Carnegie Mellon University's (CMU) Panoptic Studio dataset [20] and achieved a model overlap of 87.7% [21]. Dong et al. used the multiplexed matching of 2D poses from multiple views by simply combining the appearance-related and geometric information of the people featured in the images to compare their similarity relationships in 2D [22]. This can significantly reduce the size of the state space and improve the speed of computation, but this method uses an off-the-shelf feature extraction network to simply combine appearance-related information with geometric information in the given image. Its accuracy thus decreases when few cameras are available and it is slow at making inferences, which renders it unsuitable for use. The time taken by the model to make inferences is a key consideration in multi-view human pose estimation. The computational complexity of the model for all views increases exponentially with the number of cameras. Chen et al. used an iterative processing strategy to obtain video frames in chronological order and used them as an iterative frame-by-frame input [23]. This leads to a linear relation between the computational cost and the number of cameras, but the high speed of inference of this method makes it difficult to guarantee its accuracy.

A considerable amount of promising research on the automatic analysis of human behaviors based on deep learning has emerged in recent years. Ullah et al. proposed a long short-term memory (LSTM) network for the automatic recognition of six types of behaviors by using a smartphone that yielded an

average improvement of 0.93 in accuracy compared to previous methods. It provides a new idea for the automatic analysis of human behaviors [24]. In this paper, we propose a framework for automated multi-view human pose estimation that can be applied to overcome the drawbacks of the above models. Unlike traditional methods, the proposed method does not require that the subject wear a motor aid or a sensor system [25–27]. This study makes the following contributions to research in the area:

- We propose an architecture for multi-view human pose estimation that delivers a high accuracy and stability at a high frame rate.
- We propose a method to assess the similarity of motion based on the plane of motion-related features. The latter are quantified by being converted into FMA scores.
- We build a Virtual Reality-based evaluation platform, and implement the proposed method on the Unity framework to realistically reflect human motion through the skinned multi-person linear (SMPL) model [28].

2 Materials and Methods

2.1 Experimental Hardware Support

The hardware used in the experiments consisted of two Kinect image sensors, a computer, and a large-screen display (integrated into the computer end), as shown in Fig. 1a. The Kinect sensors were placed orthogonally with respect to each other to simultaneously capture images, as shown in Fig. 1b. This setup eliminated the problem of data loss caused by self-occlusion such that more accurate data on human poses could be obtained [29,30]. The large-screen display was used to show the results of reconstruction and the scenarios of virtual assessment.

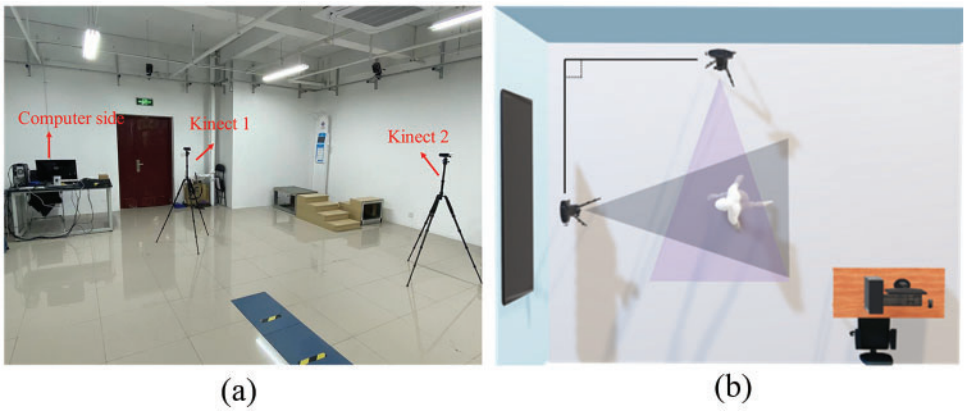

Figure 1: The proposed system. (a) Operational setup. (b) Positions of the cameras

2.2 Human Pose Reconstruction and Motion Data Acquisition

2.2.1 Overall Network Architecture for Multi-View Pose Estimation

The overall architecture of the multi-view pose estimation network is shown in Fig. 2. The You Only Look Once (YOLO) network was used to train the backbone network and design a unique feature loss function to extract the bounding box and appearance-related features from different views. We used a high-resolution network (HRNet), a 2D pose estimation network of the top-down type, to accurately identify the points representing the 2D joints of the human body in the images. Following 2D pose estimation, the 3D pose of the body needed to be reconstructed. Triangulation is the most

direct and commonly used method to quickly reconstruct the 3D pose. However, errors in the estimated 2D pose from any given view may seriously degrade the accuracy of 3D pose estimation. We introduced the 3D pictorial structure (3DPS) model and added temporal information to it to compensate for the drawbacks of the above-mentioned methods.

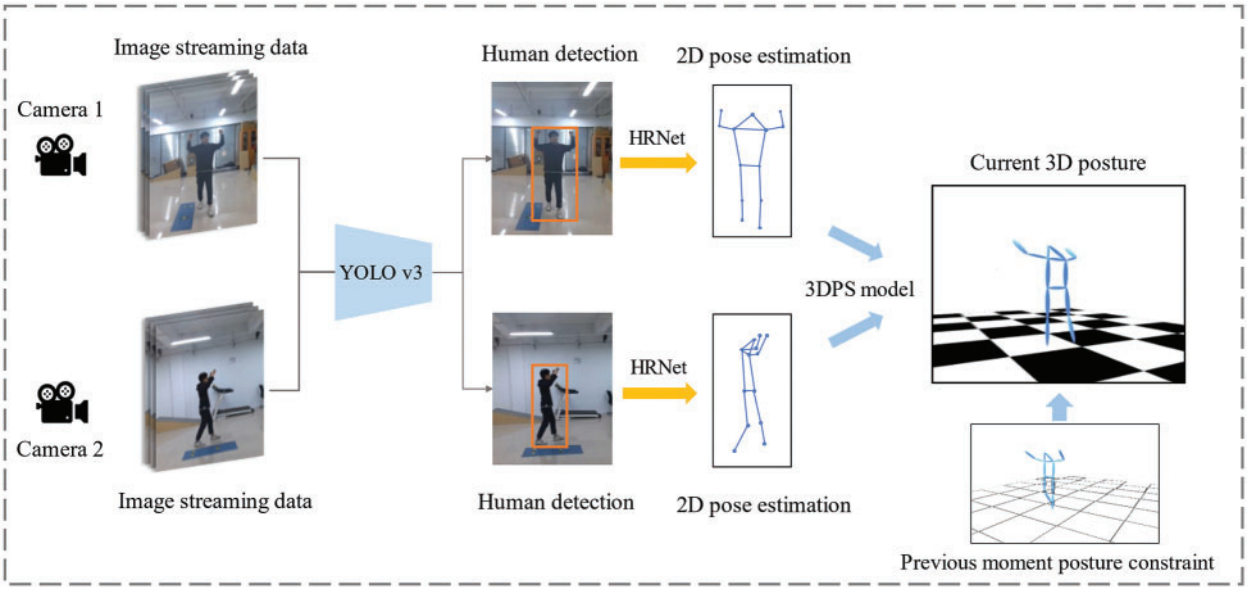

Figure 2: Overall architecture of multi-view pose estimation

2.2.2 Detection Network

The loss function of the YOLO v3 network consists of three components. We removed the loss function for target classification from the network because the task at hand involved the identification of only people in images. To locate the target, the network uses the sum of the squared error (SSE) as the loss function, and constructs it by detecting the error in position between the tensor and the real tensor. The loss function L_p is as follows:

$$L_p = L_{coo} + L_{iou} \tag{1}$$

where L_{coo} represents error in the coordinates of the target between the predicted value and the actual value, and L_{iou} denotes error in the intersection over union (IOU). The function representing error in the coordinates is defined as follows:

$$L_{coo} = \lambda_{coo} \sum_{i=o}^{S^2} \sum_{j=0}^{B} 1_{\substack{obj \\ j=0}} \left[\left(x_i - \hat{x}_i\right)^2 + \left(y_i - \hat{y}_i\right)^2 + \left(w_i - \hat{w}_i\right)^2 + \left(h_i - \hat{h}_i\right)^2 \right] \tag{2}$$

where x_i, y_i denote coordinates of the center of the bounding box, w_i and h_i denote the width and the height of the bounding box, $\hat{x}_i$, $\hat{y}_i$, $\hat{w}_i$, and $\hat{h}_i$ denote the true values of its coordinates and size, and λ_{coo} is the parameter used to weigh the loss of coordinates of the bounding box. It regulates the weight of the error in the coordinates in the final result.

The loss function of the IOU is defined by Eqs. (3) and (4):

$$L_{iou} = -\lambda_{obj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} F - \lambda_{nbj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{nbj} F \tag{3}$$

$$F = \left[\hat{C}_i^j \log\left(C_i^j\right) + \left(1 - \hat{C}_i^j\right) \log\left(1 - C_i^j\right) \right] \tag{4}$$

2.2.3 Human Pose Reconstruction

During 3D human posture reconstruction, the posture in the previous instance can be used to estimate that in the current instance. We add information on the time series to the 3DPS model to constrain the posture and improve the accuracy of 3D pose estimation. We modify the structure of the 3DPS model as follows:

$$p(J|V) \propto \prod_{m=1}^{M} \prod_{i=1}^{N} p\left(V_m | proj_t^m(j_j)\right) \prod_{(i,j)\in\varepsilon} p(j_i, j_j) \prod_{i=1}^{N} p^{\omega_t}(j_i^{t-1} | j_i^t) \tag{5}$$

where $proj_t^m(*)$ represents the projection of the 3D pose at time t, j_i^{t-1} represents the points representing the joints at time $t-1$, and ω_t represents the temporal tracking weights. The tracking term $p(j_i^{t-1}|j_i^t)$ constrains the current frame with respect to the positions of the joints in the previous frame:

$$p^{\omega_t}\left(j_i^{t-1} | j_i^t\right) = 1 - \frac{1}{T} dis\left(j_i^{t-1}, j_i^t\right) \tag{6}$$

where $dis(*)$ represents the Euclidean distance (ED) between joints and T is the normalization factor. The results of reconstruction are shown in Fig. 3.

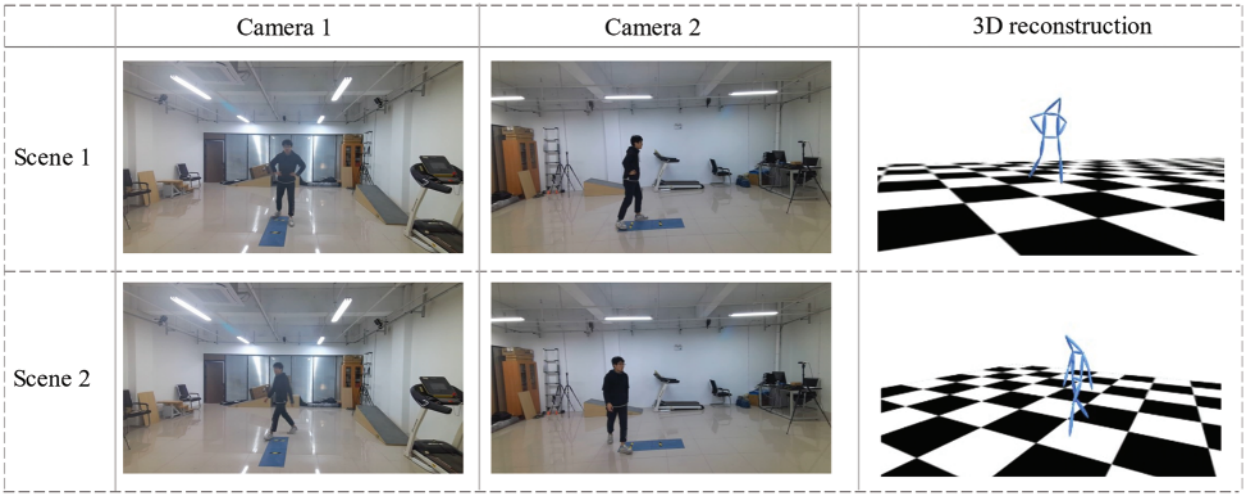

Figure 3: Results of 3D human pose reconstruction

2.3 Application of Motion Detection: Assessment of Movement Function

2.3.1 Joint Data Filtering Based on Kalman Filter

The reconstructed human pose can be used to obtain the coordinates of points representing the joints. The initial trajectory of these points may be missing or jagged, however, such that they do not match the actual trajectory of the physical motion of the person represented. Given that the points

representing the joints are in the form of 3D coordinates with a time series, we use the Kalman filter to correct the initially reconstructed points representing the joints. According to the equation of state of the system, its state at the current moment can be estimated as follows:

$$\overline{x}_k = Fx_{k-1} + Bu_{k-1} \tag{7}$$

where $\overline{x}_k$ denotes the predicted value of the current state and x_{k-1} denotes the optimal estimate of the system in the previous moment. The state of the system is $\left[x, v_x, y, v_y, z, v_z\right]$, the positions of the joints are denoted by $[x, y, z]$, $\left[v_x, v_y, v_z\right]$ are the velocities of the joints, and F is the state transfer matrix:

$$F = \begin{bmatrix} 1 & dt & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & dt & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & dt \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{8}$$

Following the prediction of the system's state at the current moment k, its covariance P needs to be updated. The initial value of P is set to a 6×6 zero matrix, and it is updated as follows:

$$\overline{P}_k = FP_{k-1}F^T + Q \tag{9}$$

where $\overline{P}_k$ denotes the covariance of error in the prior estimation, P denotes the covariance of error in the posterior estimation at moment $k-1$, and Q denotes uncertainty in the change in state. It is set as:

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.1 \end{bmatrix} \tag{10}$$

To obtain the predicted value of the system's state, the observed and the predicted values are combined to obtain the optimal value x_k in the current state.

$$\mathrm{x}_k = \overline{x}_k + K_k\left(z_k - H\overline{x}_k\right) \tag{11}$$

where x_k denotes the optimal estimate at the current moment, $\overline{x}_k$ denotes the value predicted by the system, z_k denotes the system's observation, H denotes the observation matrix, and K_k denotes the gain due to the Kalman filter. The solution is as follows:

$$R = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix} \tag{12}$$

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{13}$$

The corresponding covariance of error is calculated as follows:

$$P_k = \left(I - K_k H\right)\overline{P}_k \tag{14}$$

where I denotes the unit matrix and P_k denotes the covariance of the a priori estimation error.

2.3.2 *Dynamic Time Warping (DTW)-Based Alignment of a Series of Movements*

There is a temporal difference between the test movement, and the standard movement, and direct similarity matching between image frames yields significant errors. We use the DTW algorithm to align two sets of time series of actions. Given two sequences of motion data, X and Y, each frame is a 3D pose. We convert the point coordinates of 3D joints into a 1D sequence of angles to avoid the influence of individual differences, as shown in Fig. 4.

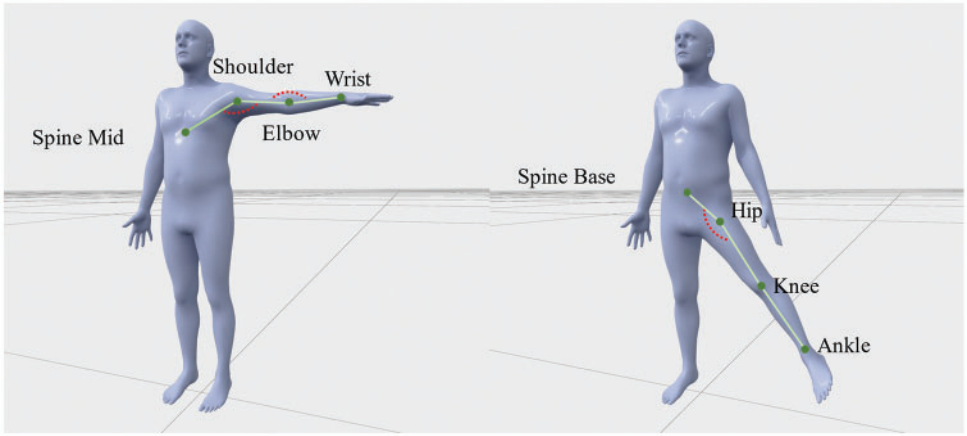

Figure 4: Conversion of the angles of the joints

We thus obtain two sets of angular time series $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_i)$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_j)$. The distance function between points in the given sequence is as follows:

$$\varphi(k) = (\varphi_\alpha(k), \varphi_\beta(k)) \tag{15}$$

where $\varphi_\alpha(k) \in [1, i]$, $\varphi_\beta(k) \in [1, j]$, and $k \in [1, T]$. Given $\varphi(k)$, the cumulative distance of the sequences can be obtained as follows:

$$d_\varphi(\alpha, \beta) = \sum_{k=1}^{T} d(\varphi_\alpha(k), \varphi_\beta(k)) \tag{16}$$

The final output of the DTW is an optimal curve of the twist in $\varphi(k)$ that minimizes the cumulative distance:

$$DTW(\alpha, \beta) = \min d_\varphi(\alpha, \beta) \tag{17}$$

2.3.3 *Matching Features on Plane*

The method to assess functions of the movements of the limbs was developed by comparing the mobility of the joints of subjects with data on healthy humans. Data on the latter were collected as a reference and analyzed by using a method of similarity assessment. The standard paradigm of movement of the FMA was used. Most researchers have used correlation coefficients or the ED to assess similarity in movements, but this yields inaccurate results due to variations in the shapes of human bodies.

We used the proposed method to extract the image of the human skeleton, and transformed the points representing joints during motion into seven feature planes as the basic planes of calculation, as shown in Fig. 5. Each feature plane represented each part of the human body. Seven normal vectors ($V_1 - V_7$) were extracted from these feature planes to determine the difference in the overall direction of the posture of the subject during motion. The local normalcy of the plane of posture was judged by the angle between the edge vectors of the feature plane ($\theta_1 - \theta_7$). Moreover, the angle between the edge vectors of the feature plane and the vertical direction of the torso ($\theta_8 - \theta_{12}$) was used to judge the local relationship between the joints of the limbs and the torso. In this way, a binary group $\langle V, \theta \rangle$ was obtained and used as input to the model for calculating similarity, and the relevant parameters of $\langle V, \theta \rangle$ were obtained as the output. This method overcomes the difficulty posed by the inherent characteristics of the object to be measured, and reduces the complexity of the calculation to improve the efficiency and stability of human pose analysis.

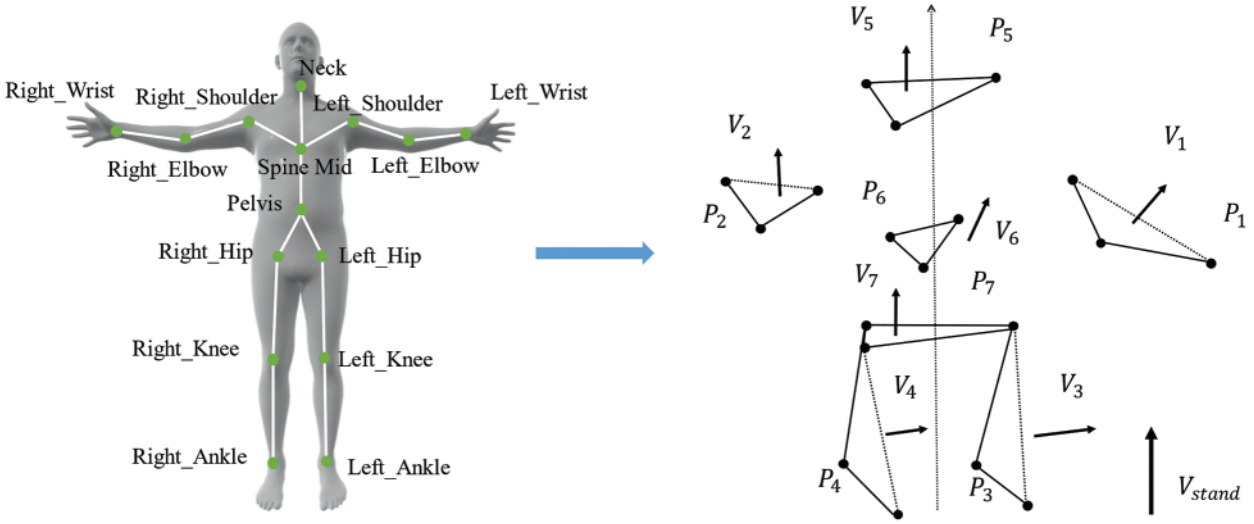

Figure 5: Simplifying the model of the human body

(1) Limb movement

The movement of the limbs can be determined by the inner product of the characteristic plane P_m $(m = 1, 2, 3, 4)$, the normal vector V_m $= (m = 1, 2, 3, 4)$, and the vector in the vertical direction V_{stand} with respect to the spine. The range of motion (ROM) can be accurately determined by the angle θ_m $(m = 1, 2, 3, 4)$. The feature vector is shown in Table 1.

Table 1: Feature vectors of the postures of the limb

Limb posture	Feature vector
Upper-left arm	$V_{LUarm} = R_{LElbow} - R_{LShoulder}$
Lower-left arm	$V_{LLarm} = R_{LElbow} - R_{LWrist}$
Upper-right arm	$V_{RUarm} = R_{RElbow} - R_{RShoulder}$
Lower-right arm	$V_{RLarm} = R_{RElbow} - R_{RWrist}$

(Continued)

Table 1 (continued)

Limb posture	Feature vector
Left thigh	$V_{LThight} = R_{LKnee} - R_{LHip}$
Left crus	$V_{LCrus} = R_{LKnee} - R_{LAnkle}$
Right thigh	$V_{RThight} = R_{RKnee} - R_{RHip}$
Right crus	$V_{RCrus} = R_{RKnee} - R_{RAnkle}$
Feature plane of left arm	$V_1 = V_{LLarm} \times V_{LUarm}$
Feature plane of right arm	$V_2 = V_{RLarm} \times V_{RUarm}$
Feature plane of left leg	$V_3 = V_{LThight} \times V_{LCrus}$
Feature plane of right leg	$V_4 = V_{RThight} \times V_{RCrus}$

(2) Head movement

This can be obtained by comparing the normal vector V_5 of plane P_5 with the vector in vertical direction V_{stand}.

(3) Spine movement

This can be obtained by comparing the angle of transformation between the vector in the direction of the spine V_6 and that vertical to it V_{stand} during rotation.

(4) Hip movement

When the body is vertical, the plane of the hip P_7 remains horizontal, and the normal vector V_7 is parallel to the vector in the vertical direction V_{stand}, as shown in Table 2.

Table 2: Feature vectors of the postures of the trunk

Trunk posture	Feature vector
Head	$V_5 = R_{Neek} \times R_{Head}$
Feature plane of spine	$V_6 = (R_{Neek} - R_{Spine}) \times (R_{Hip} - R_{Spine})$
Feature plane of hip	$V_7 = (R_{Spine} - R_{LHip}) \times (R_{Spine} - R_{RHip})$

The cosine similarity function is used as the basic metric function. Compared with the ED, it focuses more on the difference in direction between vectors. It is calculated as follows:

$$similarity\,(A_t, B_t) = \frac{A_t \times B_t}{\sqrt{(A_t)^2} \times \sqrt{(B_t)^2}} \tag{18}$$

where θ_i is the angle of the joint, and A_t and B_t are the edge vectors of the feature plane. The range of values of [0, 1] indicates closeness to the standard data.

Cosine similarity can measure the difference in direction between vectors as well as the difference in angles, which is expressed as the magnitude of the ROM:

$$corr\,(A_t, B_t) = 1 - \left(\frac{\arccos(similarity(A_t, B_t))}{\pi}\right) \tag{19}$$

3 Experiments and Evaluation

3.1 Datasets

To train an efficient and robust human detection network, we captured a large number of images from multiple cameras to form a dataset. To quickly obtain a large amount of image-related data, we used four cameras to synchronously take images from different perspectives, and obtained over 70,000 images in total as shown in Fig. 6. We fully considered the light intensity, background, and other factors of the experimental environment in the acquisition process to enrich the image-related data. We considered nine scenes, including corridors, stairs, halls, and laboratories, when capturing images.

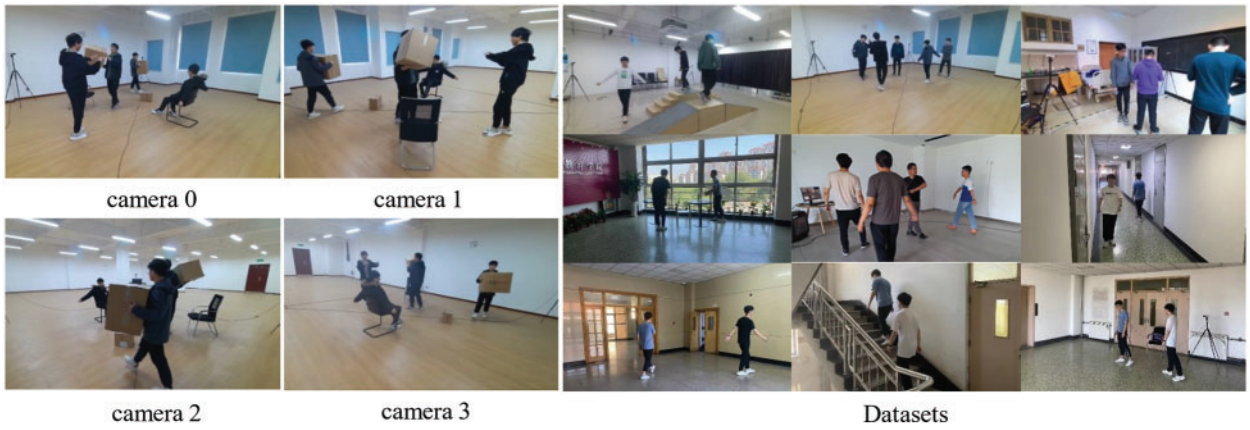

Figure 6: Dataset

Subjects in the experiments were asked to simulate actions involved daily activities, including walking, moving objects, talking, and going up and down stairs, while their images were captured to train the recognition network such that it had a strong capability for generalization.

3.2 Quantitative Analysis

We quantitatively compared the proposed method with prevalent techniques in the area on the publicly available Shelf dataset [19]. This dataset consists of images captured by five cameras at a resolution of 1032 $\times$ 776. "Actor 1–Actor 3" in Table 3 represent the three persons depicted in images in the Shelf dataset and "Avg." represents the average accuracy. The percentage of correctly estimated parts (PCP) criterion was used to assess the performance of the methods.

Table 3 shows the accuracies of detection of the proposed method and prevalent methods in the area on all individual frames in the Shelf dataset. Because the proposed method contains a constraint on the temporal information provided to the 3DPS model, and as the accurate detection of static images does not require this feature, it did not exhibit an absolute advantage over the other methods. In fact, its average accuracy was lower by 0.2% compared with that of the method proposed by Zhang et al. However, a practical model needs to reason over a range of dynamic video frames and deliver excellent performance even at high frame rates. We thus designed and conducted a comparative experiment to assess the accuracy of the methods listed in Table 3 on images captured under different frames per second (FPS) to further test the performance of the proposed method. We still used the PCP as the indicator for evaluation, and the results are shown in Table 4.

Table 3: Results on the shelf dataset in terms of accuracy

References	Actor 1	Actor 2	Actor 3	Avg
Burenius et al. [18]	66.1	65.0	83.2	71.4
Belagiannis et al. [19]	75.3	69.7	87.6	77.5
Ershadi-Nasab et al. [31]	93.3	75.9	94.8	88.0
Dong et al. [22]	97.2	79.5	96.5	91.1
Zhang et al. [32]	**99.0**	**96.2**	97.6	**97.6**
Proposed method	98.6	95.8	**97.9**	97.4

Table 4: Accuracy of pose estimation on images captured at different frame rates

References	30 FPS	15 FPS	10 FPS
Burenius et al. [18]	71.2	71.4	71.4
Belagiannis et al. [19]	77.5	77.5	77.5
Ershadi-Nasab et al. [31]	88.0	88.0	88.0
Dong et al. [22]	96.9	96.9	96.9
Zhang et al. [32]	81.5	81.5	81.5
Proposed method	**97.1**	**97.1**	**97.0**

The proposed method obtained better results than the traditional 3DPS method, and attained results similar to those of the methods developed by Zhang et al. [32] and Dong et al. [22] owing to deep learning-based optimization and because it was trained on a large dataset. However, in terms of speed of reasoning, it was superior to the methods proposed by Zhang et al. [32] and Dong et al. [22] It had an average accuracy of dynamic detection that was higher by 15.6% (30 FPS) than that of Zhang et al.'s method. The latter method did not leverage temporal information, because of which changes in the frame rate of the camera did not have a significant impact on the final results. Temporal information was added to the 3DPS model to optimize the captured pose at any given moment by the pose captured at the previous moment, because of which it yielded more reliable results in scenes captured at a low frame rate or those featuring people moving quickly. Therefore, the proposed method can guarantee a high speed of inference and stable performance while minimizing the loss of accuracy.

3.3 Results of Data Filtering

We chose two actions from the Clinical Movement Assessment Scale for experiments to illustrate the effectiveness of filtering the points representing the joints in the image. The trajectories of the joints of the shoulder and the elbow were plotted to represent movements of the upper limbs, and those of the joints of the hip and the knee were plotted for movements of the lower limbs. The results are shown in Fig. 7.

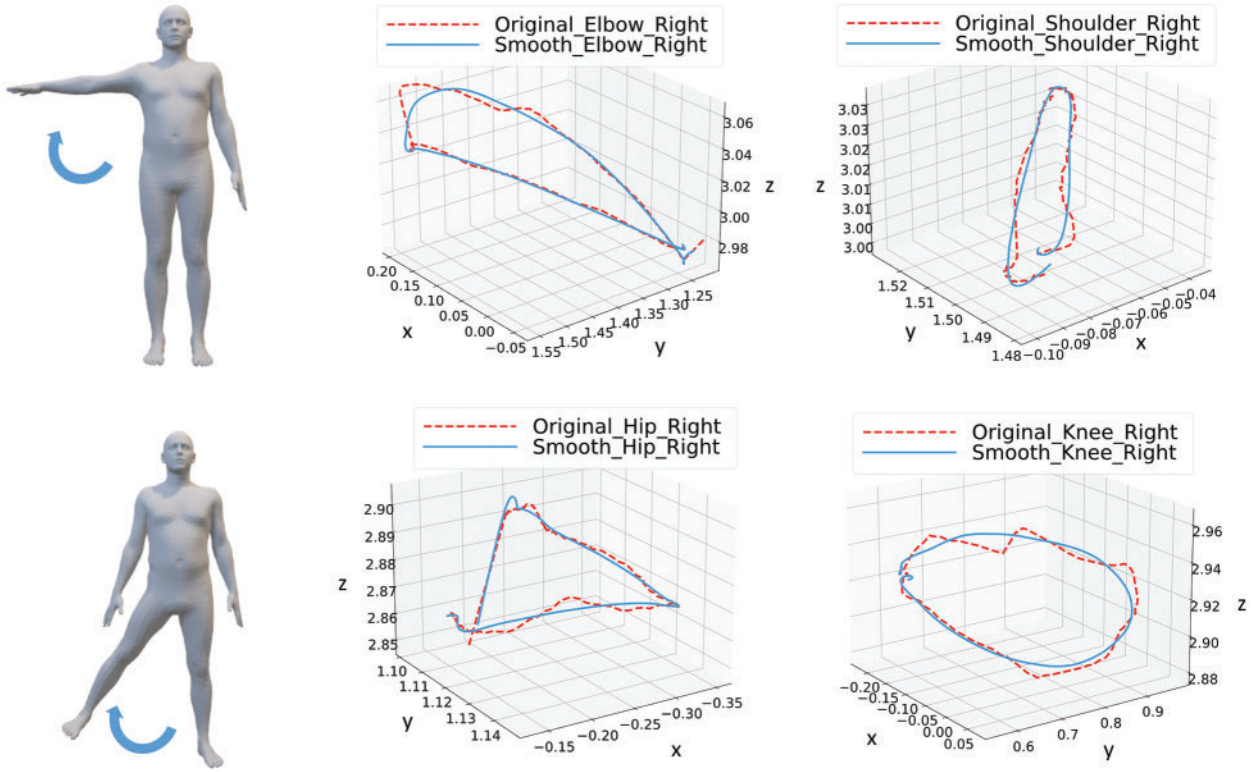

Figure 7: Filtering the points representing the joints by using the Kalman filter

3.4 Results of DTW

Two sets of angular sequences of abduction of the shoulder (90 degrees) were used as input, one for the group of healthy subjects and the other for the patient's movement sequence, and the results of alignment are shown in Fig. 8. The DTW algorithm was able to eliminate error in time between the sets of sequential movements.

3.5 Analysis of Assessment of Movement Functions

In the experiment, the subjects performed shoulder abduction movements. Their feature planes during the movements were plotted and qualitatively analyzed in comparison with those of the normal group. The results are shown in Fig. 9.

The results in Fig. 9 show that the area of the feature plane of healthy subjects was larger than that of the subjects with motor dysfunction, where this reflects the poorer accessibility of the movements of limbs of the latter compared with the former.

We performed quantitative calculations to further illustrate the variation between the subjects' movements and standard movements. The 15 feature indicators listed in Tables 1 and 2 were calculated, and the results are shown in Fig. 10. The quantitative results yielded the similarity scores of each part of the human body. As the experimental movement here was the abduction of the right shoulder, the other parts of the body had lower variation and, thus, higher scores.

We set *sim* as the quantitatively calculated directional variation in each part of the body and *corr* as the quantitatively calculated angular variation in it. The two calculations were combined to obtain the final evaluation:

$$s = \left(\frac{\sum_{i=1}^{N} sim(i)}{N} C_1 + \frac{\sum_{i=1}^{N} corr(i)}{N} C_2 \right) \times 100\% \tag{20}$$

where $N = 15$ is the number of parts of the body involved in the similarity calculation, and C_1 and C_2 are proportional parameters. The weight of the direction of motion and its motion was set to 0.5.

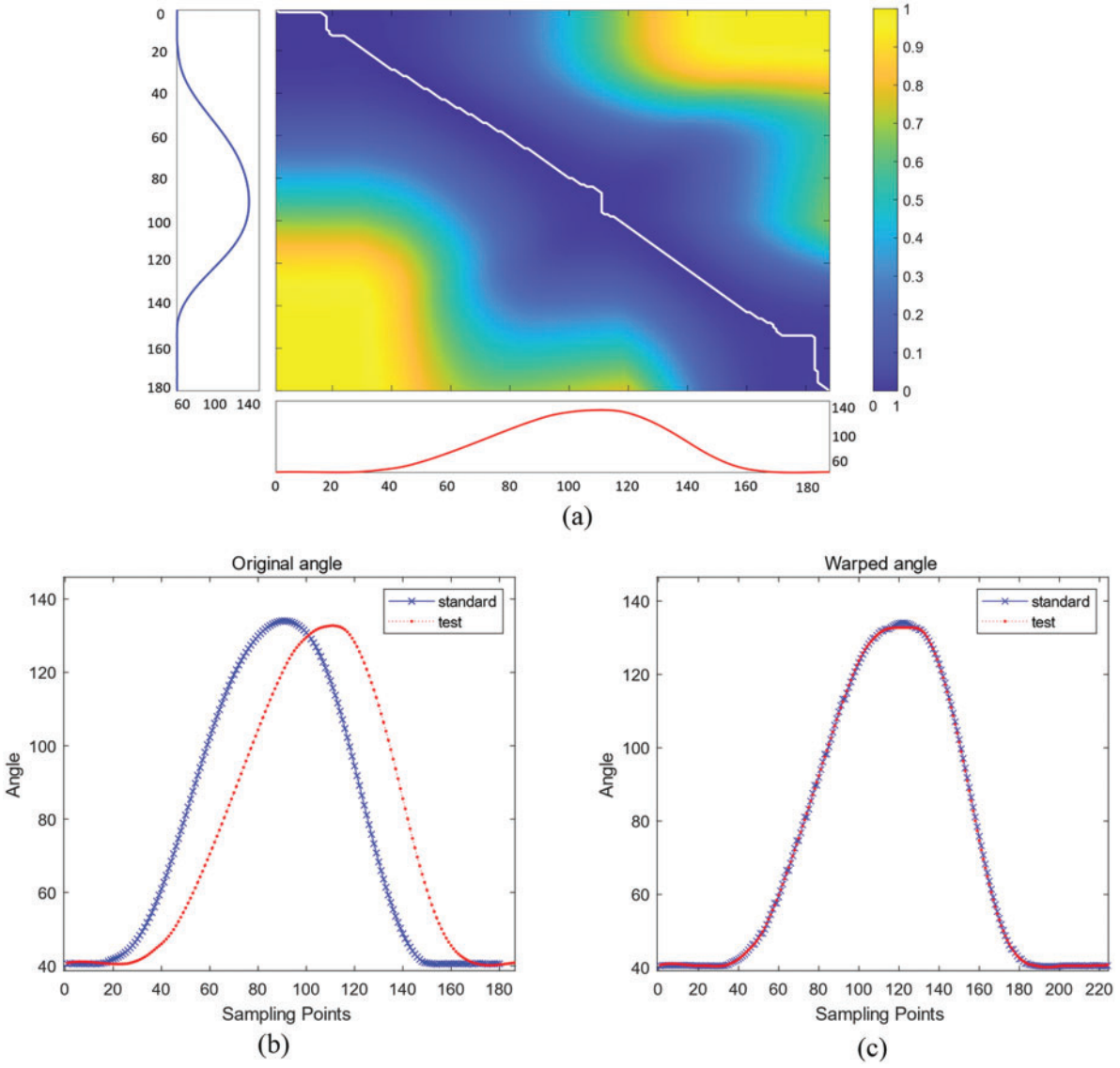

Figure 8: Results of alignment of the DTW. (a) Path of optimization of the DTW. (b) Original sequence of angles of the joints. (c) Aligned sequence of angles of the joints

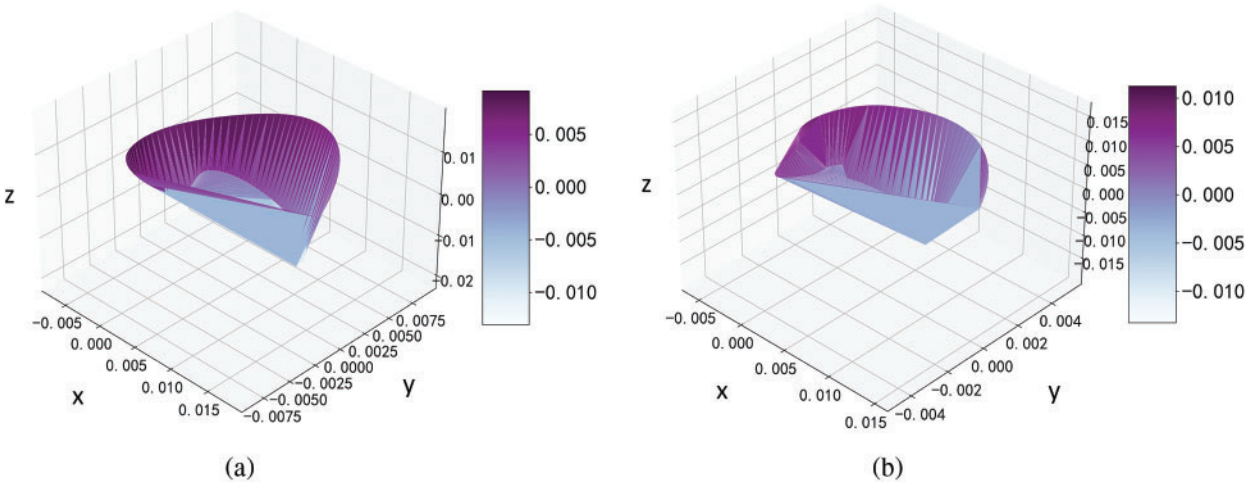

Figure 9: Feature plane during shoulder abduction movement. (a) Standard paradigm of the feature plane of motion. (b) Feature plane of motion of the subject

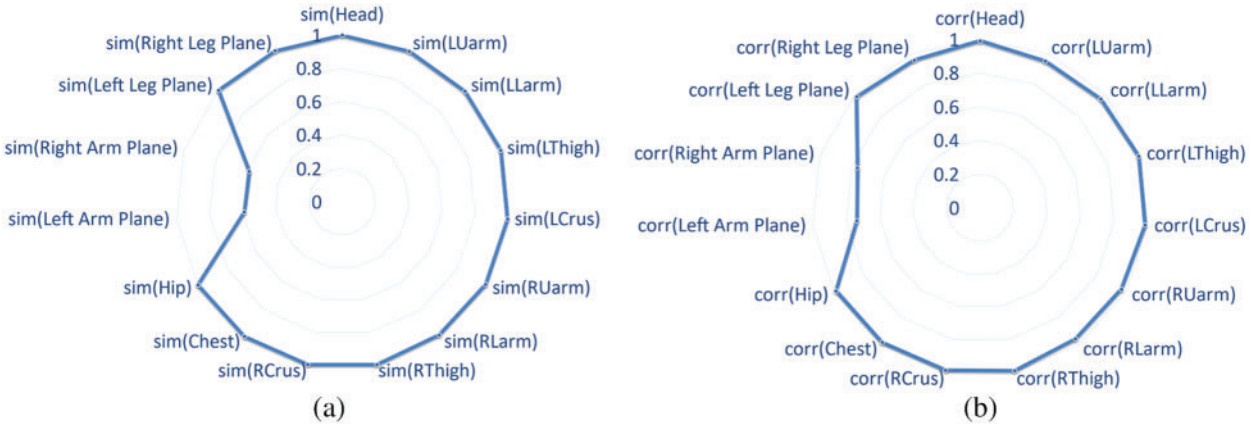

Figure 10: Similarity analysis of the plane of features. (a) Differential scores for the directions of limb movement. (b) Differential scores for the ROM

To further illustrate the effectiveness of the proposed method, 18 subjects were invited to participate in the experiments to assess movement functions. All participants were evaluated on the artificial FMA scale as the control group for the experiment and their results were subjected to a correlation analysis. The subjective ratings of experts were regarded as the gold standard to validate the proposed algorithm. All participants were aware of the experimental procedure, and provided their written informed consent for participation before the experiment. It is hereby declared that all the experiments involved in this paper have passed the local ethical review.

Table 5 compares the results of the proposed method of assessment with those of the ED-based method. The correlation coefficient reflected a strong correlation between the proposed method and the artificial scores, with a value of 0.87 ($p < 0.001$). The traditional ED-based method of assessment directly used spatial distances between 3D joints while ignoring temporal variations in movements and the varying physical characteristics of different people. Its results were thus less accurate. Fig. 11 shows

a scatterplot of the final scores of both methods and the artificial ratings for 18 subjects. Additional analyses of the scores showed that in many cases, the algorithms overestimated the results compared with the artificial ratings. Therefore, we calibrated the scores of the proposed method by using the fitted equation of linear regression so that it generated similar evaluation scores to those assigned by experts.

We applied the proposed method to the Unity framework and built a Virtual Reality-based scene for evaluation by using the SMPL as the models of virtual characters and the proposed method of pose estimation as the basic skeleton to drive the models. The use of SMPL models as avatars for exhibiting movement and control allowed for a more realistic reproduction of the movements of the human limbs, as shown in Fig. 12. The virtual coach (left) model was driven by standard movement-related data to guide the subject by mimicking its movements. The user model (right) was driven in real time by the human skeleton as estimated from the subject's pose to provide immediate feedback to the user on their pose.

Table 5: Quantitative results of the assessment of the capacity for movement

No.	Age (year)	Height (cm)	Weight (kg)	Artificial evaluation	ED evaluation	Evaluation by proposed algorithm
01	25	176	62	76	65.2	85.3
02	27	180	65	80	62.1	85.6
03	22	178	74	75	67.5	70.0
04	29	172	58	72	60.6	81.3
05	31	173	61	70	63.4	78.1
06	45	170	75	72	65.7	80.2
07	28	175	72	73	68.0	78.6
08	26	173	75	66	62.6	72.3
09	35	177	82	85	80.2	91.5
10	24	181	60	83	77.6	87.7
11	23	177	65	72	65.8	77.0
12	22	178	74	74	65.6	77.4
13	46	167	65	68	67.2	65.3
14	30	177	70	63	66.9	68.0
15	52	172	76	73	77.5	80.0
16	26	165	60	78	72.6	82.4
17	19	164	66	72	65.0	75.9
18	20	167	63	88	79.9	92.4
Pearson's correlation coefficient				1.0	0.67	0.87

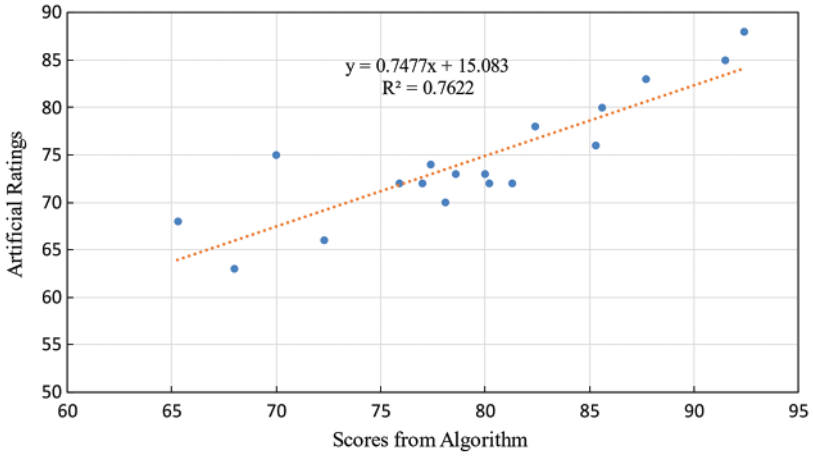

Figure 11: Linear relationship between algorithm scores and artificial ratings

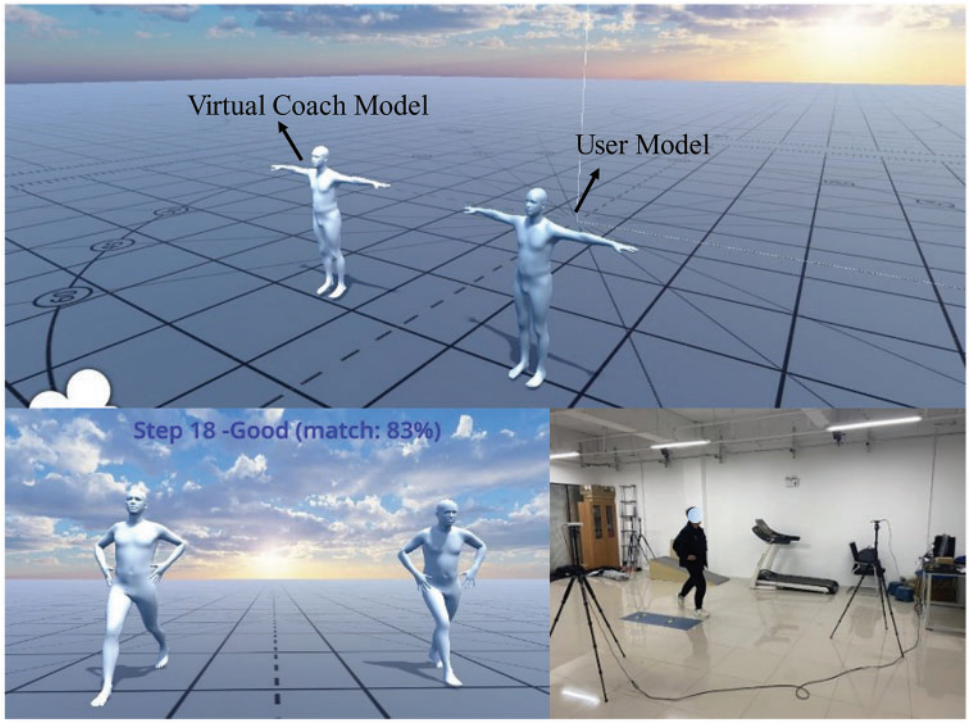

Figure 12: Implementation of the proposed method in the unity framework

4 Conclusions and Future Work

In this study, we proposed a feasible and lightweight method of pose estimation. A detection network was fully trained by building a large dataset of scenes featuring moving people, and yielded an accuracy of 97.1% while delivering stable performance on images captured at 30 FPS. Following this, we used the proposed method to develop an algorithm to assess the similarity of movements on feature planes. By evaluating the motor functions of 18 subjects and comparing them with their manual

scores, we found that the proposed method of evaluation exhibited a strong linear correlation with the manual scores. We used linear regression to fine-tune the model and render its results more realistic. In contrast to past methods that have used performance scores obtained from subjects through serious games to evaluate their motor abilities [33–35], the quantitative scores of movements obtained here are more easily interpretable for use in clinical medicine. Although these scores may be related to the range of motion of certain joints, they are influenced by other factors as well, such as the patient's level of cognition and the difficulty of the game.

When we used the proposed method of pose estimation in the Unity framework to drive the SMPL models, there were deviations in the local pose of the model, and it even yielded inaccurate poses, because the points representing the joints of the hand and parts of the ankle were not identified. This slightly affected the results of evaluation of the proposed model. In future work, we plan to reconstruct the points representing the joints of the hand and parts of the ankle, and will consider embedding the initial model of the human skeleton into the SMPL model to obtain a more accurate 3D model through forward kinematics. In addition, we plan to exploit multi-person posture estimation to extend the proposed method to simultaneously assess the motor functions of several people to enhance its scope of application.

Acknowledgement: The authors thank their institutions for infrastructure support.

Funding Statement: This work was supported by grants from the Natural Science Foundation of Hebei Province, under Grant No. F2021202021, the S&T Program of Hebei, under Grant No. 22375001D, and the National Key R&D Program of China, under Grant No. 2019YFB1312500.

Author Contributions: Lingling Chen performed the supervision; Tong Liu performed the data analyses and wrote the manuscript; Zhuo Gong performed the experiment and software; Ding Wang performed the methodology and visualization.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, L Chen, upon reasonable request.

Ethics Approval: This study was approved by the Biomedical Ethics Committee of Hebei University of Technology (Approval Code: HEBUTHMEC2022018, Approval Date: March 14, 2022). Written informed consent was obtained from all participants.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Schönenberg, U. Teschner and T. Prell, "Expectations and behaviour of older adults with neurological disorders regarding general practitioner consultations: An observational study," *BMC Geriatrics*, vol. 21, no. 1, pp. 1–12, 2021.

[2] W. H. Bangyal, J. Ahmad and H. T. Rauf, "Optimization of neural network using improved bat algorithm for data classification," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 4, pp. 670–681, 2019.

[3] Y. M. Tang, L. Zhang, G. Q. Bao, F. J. Ren and W. Pedrycz, "Symmetric implicational algorithm derived from intuitionistic fuzzy entropy," *Iranian Journal of Fuzzy Systems*, vol. 19, no. 4, pp. 27–44, 2022.

[4] W. H. Bangyal, K. Nisar, A. A. B. Ag, M. R. Haque, J. J. Rodrigues *et al.,* "Comparative analysis of low discrepancy sequence-based initialization approaches using population-based algorithms for solving the global optimization problems," *Applied Sciences*, vol. 11, no. 16, pp. 7591, 2021.

[5] R. Divya and J. D. Peter, "Smart healthcare system-a brain-like computing approach for analyzing the performance of detectron2 and PoseNet models for anomalous action detection in aged people with movement impairments," *Complex & Intelligent Systems*, vol. 8, no. 4, pp. 3021–3040, 2022.
[6] A. Pardos, A. Menychtas and I. Maglogiannis, "On unifying deep learning and edge computing for human motion analysis in exergames development," *Neural Computing and Applications*, vol. 34, no. 2, pp. 951–967, 2022.
[7] M. H. Li, T. A. Mestre, S. H. Fox and B. Taati, "Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation," *Journal of Neuroengineering and Rehabilitation*, vol. 15, no. 1, pp. 1–13, 2018.
[8] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu *et al.,* "Cascaded pyramid network for multi-person pose estimation," in *Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, pp. 7103–7112, 2018.
[9] K. Sun, B. Xiao, D. Liu and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Computer Vision and Pattern Recognition, CVPR 2019*, Long Beach, CA, USA, pp. 5693–5703, 2019.
[10] Y. Tang, Z. Pan, W. Pedrycz, F. Ren and X. Song, "Based kernel fuzzy clustering with weight information granules," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 2, pp. 342–356, 2022.
[11] Z. Cao, T. Simon, S. E. Wei and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, pp. 7291–7299, 2017.
[12] S. Kreiss, L. Bertoni and A. Alahi, "Pifpaf: Composite fields for human pose estimation," in *Computer Vision and Pattern Recognition, CVPR 2019*, Long Beach, CA, USA, pp. 11977–11986, 2019.
[13] D. C. Luvizon, D. Picard and H. Tabia, "2D/3D pose estimation and action recognition using multitask deep learning," in *Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, pp. 5137–5146, 2018.
[14] F. Moreno-Noguer, "3D human pose estimation from a single image via distance matrix regression," in *Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, pp. 2823–2832, 2017.
[15] J. Martinez, R. Hossain, J. Romero and J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Int. Conf. on Computer Vision, ICCV 2017*, Venice, Italy, pp. 2640–2649, 2017.
[16] G. Pavlakos, X. Zhou and K. Daniilidis, "Ordinal depth supervision for 3D human pose estimation," in *Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, pp. 7307–7316, 2018.
[17] B. Tekin, P. Márquez-Neila, M. Salzmann and P. Fua, "Learning to fuse 2D and 3D image cues for monocular body pose estimation," in *Int. Conf. on Computer Vision, ICCV 2017*, Venice, Italy, pp. 3941–3950, 2017.
[18] M. Burenius, J. Sullivan and S. Carlsson, "3D pictorial structures for multiple view articulated pose estimation," in *Computer Vision and Pattern Recognition, CVPR 2013*, Portland, OR, USA, pp. 3618–3625, 2013.
[19] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab *et al.,* "3D pictorial structures revisited: Multiple human pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 1929–1942, 2015.
[20] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe *et al.,* "Panoptic studio: A massively multiview system for social motion capture," in *Int. Conf. on Computer Vision, ICCV 2015*, Santiago, Chile, pp. 3334–3342, 2015.
[21] H. Joo, T. Simon and Y. Sheikh, "A 3D deformation model for tracking faces, hands, and bodies," in *Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, pp. 8320–8329, 2018.
[22] J. Dong, W. Jiang, Q. Huang, H. Bao and X. Zhou, "Fast and robust multi-person 3D pose estimation from multiple views," in *Computer Vision and Pattern Recognition, CVPR 2019*, Long Beach, CA, USA, pp. 7792–7801, 2019.
[23] L. Chen, H. Ai, R. Chen, Z. Zhuang and S. Liu, "Cross-view tracking for multi-human 3D pose estimation at over 100 fps," in *Computer Vision and Pattern Recognition, CVPR 2020*, Seattle, WA, USA, pp. 3279–3288, 2020.

[24] M. Ullah, H. Ullah, S. D. Khan and F. A. Cheikh, "Stacked LSTM network for human activity recognition using smartphone data," in *European Workshop on Visual Information Processing, EUVIP 2019*, Rome, Italy, pp. 175–180, 2019.
[25] A. Leardini, G. Lullini, S. Giannini, L. Berti, M. Ortolani *et al.,* "Validation of the angular measurements of a new inertial-measurement-unit based rehabilitation system: Comparison with state-of-the-art gait analysis," *Journal of Neuroengineering and Rehabilitation*, vol. 11, no. 1, pp. 1–7, 2014.
[26] S. V. Prokopenko, E. Y. Mozheiko, M. L. Abroskina, V. S. Ondar, S. B. Ismailova *et al.,* "Personalized rehabilitation assessment of locomotor functions in Parkinson disease using three-dimensional video analysis of motions," *Russian Neurological Journal*, vol. 26, no. 1, pp. 23–33, 2021.
[27] B. Tran, X. Zhang, A. Modan and C. M. Hughes, "Design and evaluation of an IMU sensor-based system for the rehabilitation of upper limb motor dysfunction," in *Biomedical Robotics and Biomechatronics, BioRob 2022*, Seoul, Korea, pp. 1–6, 2022.
[28] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 1–6, 2015.
[29] H. Du, Y. Zhao, J. Han, Z. Wang and G. Song, "Data fusion of multiple kinect sensors for a rehabilitation system," in *Engineering in Medicine and Biology Society, EMBC 2016*, Orlando, FL, USA, pp. 4869–4872, 2016.
[30] Y. Jiang, K. Song and J. Wang, "Action recognition based on fusion skeleton of two kinect sensors," in *Int. Conf. on Culture-Oriented Science & Technology, ICCST 2020*, Beijing, China, pp. 240–244, 2020.
[31] S. Ershadi-Nasab, E. Noury, S. Kasaei and E. Sanaei, "Multiple human 3D pose estimation from multiview images," *Multimedia Tools and Applications*, vol. 77, no. 12, pp. 15573–15601, 2018.
[32] Y. Zhang, L. An, T. Yu, X. Li, K. Li *et al.,* "4D association graph for realtime multi-person motion capture using multiple video cameras," in *Computer Vision and Pattern Recognition, CVPR 2020*, Seattle, WA, USA, pp. 1324–1333, 2020.
[33] L. Y. Chen, L. Y. Chang, Y. C. Deng and B. C. Hsieh, "The rehabilitation and assessment in virtual reality game for the patient with cognitive impairment," in *Int. Symp. on Computer, Consumer and Control, IS3C 2020*, Taichung, Taiwan, pp. 387–390, 2020.
[34] D. L. Farias, R. F. Souza, P. A. Nardi and E. F. Damasceno, "A motor rehabilitation's motion range assessment with low-cost virtual reality serious game," in *Int. Conf. on E-Health Networking, Application & Services, HEALTHCOM 2021*, Shenzhen, China, pp. 1–5, 2020.
[35] M. E. Gabyzon, B. Engel-Yeger, S. Tresser and S. Springer, "Using a virtual reality game to assess goal-directed hand movements in children: A pilot feasibility study," *Technology and Health Care*, vol. 24, no. 1, pp. 11–19, 2016.