Journal of Computer and Communications, 2023, 11, 102-121
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Detection of 3D Human Posture Based on Improved Mediapipe

Yiqiao Lin, Xueyan Jiao, Lei Zhao

School of Computer Science and Technology, Shandong University of Technology, Zibo, China
Email: lyq2462975232@126.com

How to cite this paper: Lin, Y.Q., Jiao, X.Y. and Zhao, L. (2023) Detection of 3D Human Posture Based on Improved Mediapipe. *Journal of Computer and Communications*, **11**, 102-121.
https://doi.org/10.4236/jcc.2023.112008

Received: February 3, 2023
Accepted: February 25, 2023
Published: February 28, 2023

Abstract

Based on the continuous development of motion capture technology for ordinary video images, unmarked optical motion capture has become the fastest human posture recognition technology. Compared with other technical products, Google's 3D human body recognition framework—Mediapipe is the most mature representative in this field. However, Mediapipe also has many defects in the detection of 3D human posture. In this paper, firstly, to solve the problem of inaccurate detection of human posture by Mediapipe, the accuracy of 2D human posture detection is improved through the speed threshold correction method for each joint; According to the problem that the monocular camera can not detect the depth *Z* value in the human posture data accurately, the *Z* value of the joint point is corrected for the human tilt angle through statistics; Then, according to the inaccurate recognition of *Z* value caused by large body posture, the accurate correction of *Z* value of human posture under different body posture is realized by normalizing the simulation proportion of each body limb; Finally, in order to solve the problem of jitter, lag problem and periodic noise in multiple frames caused by the speed change of human joints, one euro filtering and mean filtering of joint data are carried out. This paper verifies that the accuracy of 3D human posture detection based on the improved Mediapipe is more than 90% through the multi-pose recognition test for people of different heights, weights, ages and gender.

Keywords

Mediapipe, Inaccurate Recognition of *Z* Value, Speed Threshold Method, Statistical Method, Limb Simulation, One Euro Filtering, Mean Filtering

1. Introduction

In today's field of computer vision, human posture recognition is a hot topic of research work. Its main task is to detect the position of key points of human body in video or image. By inferring the key feature points of the whole body from the RGB video frame, the function of corresponding expansion based on some key feature points is realized. This technology has a broad application prospect in human-computer interaction, posture tracking, behavior recognition and other fields. Therefore, it is very important to realize accurate recognition of human posture, especially to detect the position of key points on the human body.

2. Current Situation Analysis

2.1. Development of Human Posture Technology

Motion capture technology is a technology that records the relevant data and restores the posture of human body structure through external devices [1]. As early as the 1980s and 1990s, motion capture technology has gradually been active in all major film industries. With the development of motion capture technology, motion capture technology has gradually evolved from traditional wearable device [2] technology to many optical motion capture system technologies based on computer vision principles [3].

In recent years, with the rapid development of the field of computer vision and the gradual improvement of artificial intelligence technology, in the optical motion capture system, the human posture recognition technology based on motion capture of ordinary video images [4] has been widely and profoundly applied in the relevant fields of computer vision.

2.2. Human Posture Recognition Category

According to the differences in human posture dimensions, human posture estimation tasks can be divided into 2D human posture estimation and 3D human posture estimation [5]. The goal of 2D human pose estimation (2D HPE) [6] is to locate and identify the key points of the human body, and connect these key points according to the joint sequence to form a projection on the two-dimensional plane of the image, so as to obtain the human skeleton consistent with the real person [7].

The main task of 3D human pose estimation [8] (3D human pose estimation, referred to as 3D HPE) is to predict the spatial position of human joint points in 3D space, which provides X, Y and Z coordinates for each landmark.

2.3. Human Pose Recognition Algorithms at Home and Abroad

In recent years, with the successful application of deep learning in the field of human posture estimation, the accuracy and generalization ability of 2D HPE have been significantly improved. The more advanced is the open-source library Openpose [5] developed by Carnegie Mellon University (CMU) based on convolutional neural network and supervised learning, which is the most popular 2D human posture estimation method at present, and it is an open-source real-time multi-person detection with high accuracy key points, However, 3D human posture data cannot be recognized, and the robustness is poor, and the requirements for computer graphics card equipment are high [8].

PoseNet [9] is a technical solution for posture estimation released by TensorFlow and Google Creative Lab. PoseNet can also be used to estimate a single pose or multiple poses, and it is not too dependent on the performance acceleration of GPU, but it is also unable to recognize 3D human posture data.

In practical applications, because 3D HPE adds depth information on the basis of 2D HPE, its expression of human posture is more accurate than 2D HPE, so its application scope and research value are higher than 2D HPE. Through 3D posture estimation, we can determine the angle of each joint of human skeleton, which can be widely used in human-computer interaction, motion analysis, rehabilitation training, etc. The mainstream 3D HPE includes the following:

MoveNet [10] is an ultra-fast and accurate model that can detect 17 key points of the body. However, this model is a lightweight 3D HPE framework that mainly supports mobile phones. Compared with PC, the accuracy of human posture recognition is not high.

Baidu's 3D limb key point SDK technology inputs RGB images through ordinary monocular cameras, outputs the 3D coordinate information of the 16 core key points of the human body, detects, tracks and accurately estimates the 3D posture of the human body in real time, and is compatible with iOS, Android and embedded platforms. However, it has a high occupancy rate of GPU, and has a big problem in the prediction of the depth of the human body.

Mediapipe, a data stream processing machine learning application development framework developed and open source by Google, is a learning development framework with built-in fast machine learning reasoning and processing, and can also achieve end-to-end acceleration on common hardware. Compared with the above human posture algorithms, Mediapipe not only reduces the dependence on the performance of graphics cards, but also can realize 3D human posture recognition, where the X and Y values are more accurate, But there is a big error in the depth Z value [11].

Based on many problems of Mediapipe in 3D HPE, firstly, according to the unstable problem of human posture recognition caused by light environment, the accuracy of recognition of X and Y values in human posture coordinates in 2D is improved by joint velocity threshold method; According to the problem of inaccurate detection of depth Z value in 3D human body posture by monocular camera, through the test of Mediapipe human body detection, it is found that the inaccurate prediction of Z value is mainly related to two factors, namely, human body tilt and limb large motion posture, which will lead to the overall tilt of the human body and limb length difference, respectively, The statistical method of human body tilt and the method of limb simulation scale normalization are used to correct the Z value; Finally, according to the situation that the speed of a joint changes too fast in the short term and there is pulse interference and uniform noise in the long term, the one-euro filter and the mean filter are used to smooth the data.

3. Improvements to Mediapipe

Mediapipe is a framework for Google to deploy the model forward after training the human posture data. Due to the influence of the human posture data set when labeling, the trained model has many defects in the detection of 3D human posture. The deep learning algorithm relies on a large number of training data. However, due to the high difficulty and cost of 3D pose annotation, many data sets are collected in a specific laboratory environment, which leads to many inaccuracies in the accuracy of Mediapipe detection.

It is very difficult to label the 3D human posture data set, which requires a variety of lasers, cameras, wearable infrared receivers and other devices. Due to the difficulty of data collection, Google has not disclosed the precious data set. Therefore, this article cannot retrain according to the defects of Mediapipe, so it is necessary to carry out special treatment according to the bad detection effect of 3D human posture based on Mediapipe. For example, the results show that the accuracy of the detection of 2D joint points of Mediapipe will decrease due to the influence of the light environment. In this paper, we can use the speed threshold method mentioned below to perform special processing on the video in different light environments.

Secondly, when the 3D human pose is detected under the monocular camera, the depth Z value of the human joint will have the problem of inaccurate data. Therefore, this paper starts with the multi-solution problem of the mapping from 2D image to 3D pose when predicting the human pose through monocular images, because this is an important reason for the accuracy of the monocular camera's prediction of the depth Z value on the human pose to decline. The specific method can be used to correct the Z value through the human tilt statistical method and the human limb proportion simulation in the following.

3.1. Abbreviations and Acronyms Introduction to Mediapipe

The Mediapipe developed by Google was originally designed to analyze audio and video on YouTube in real time. Later, with the gradual development of computer vision technology, Google has widely used many internal products and services in many aspects such as camera target detection, augmented reality advertising and cloud vision API application. From 2012 to now, Mediapipe has become the most popular open source learning and development framework for human posture detection.

Mediapipe can provide a cross-platform, customizable machine learning solution for many live and streaming media. As a learning framework built by the multimedia machine learning model pipeline, Mediapipe can process the data related to time series such as video and audio in the multimedia, and has less resource consumption. It can run across platforms on mobile devices, workstations and servers, and supports GPU acceleration for different devices (**Figure 1**).

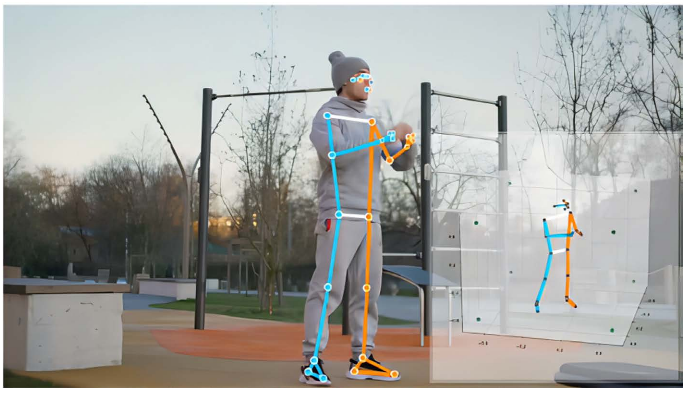

Figure 1. Mediapipe Pose's position detection of 33 posture joints.

Mediapipe has many functions, including face recognition, iris detection, posture, hands, hair and other target detection. Users can develop applications for different functions according to their business needs. Taking pose detection as an example, Mediapipe Pose is a machine learning solution for high-fidelity body pose tracking. BlazePose research is used to infer the 3D coordinates of 33 joint points on the whole body in real time from each frame of RGB video, where Z represents the coordinate depth. The depth of the hip midpoint is the origin. The smaller the value, the closer the coordinate is to the camera.

3.2. Speed Threshold Correction Method

Mediapipe's detection of human posture is not accurate. In some frames, when the ambient light changes, it will show that the previous frame is still the normal detection of human posture, but the next frame will show false detection. For example, in a continuous frame of video, the human body in a certain frame is doing normal actions. At this time, Mediapipe's detection of the human body in the video also presents a normal 2D human posture; However, when the light or surrounding objects change in the next frame, Mediapipe will immediately shift from the normal detection of human posture in the previous frame to the detection of surrounding objects, which leads to a decline in the recognition rate of human posture detection. As shown in Figure 2 and Figure 3, the left image shows the correct detection results of the human posture of Mediapipe at frame 134 of the figure skating video, and the right image shows the error detection results of the human posture of Mediapipe at frame 135 of the figure skating video.

In order to improve the accuracy of the recognition of 2D human posture when Mediapipe detects the human body, this paper collects standard motion videos of various professional sports athletes and fitness coaches, such as 12 kinds of motion videos, such as basketball, figure skating, tai chi, aerobics, martial arts, yoga, gymnastics, etc. These videos have stable environment and standardized human movements, and can be used as standard human posture videos. As shown in Figure 4 and Figure 5, the competition action of Chinese national skater Jin Boyang and the indoor fitness action of Pamela, the hottest female fitness coach on YouTube, take these official standard video actions as the change threshold of the standard joint point rate (the ratio of the change of 2D coordinates of the joint point in two consecutive frames to time). When the change rate of human joint point detected by Mediapipe between two consecutive frames is lower than this threshold, the joint point will not be corrected; If the change rate of detected human joint points between two consecutive frames is higher than this threshold, the joint point coordinates can be corrected through this threshold, thus reducing the false detection to a certain extent.

The specific method is to run the above 12 standard human motion videos through Mediapipe. Each frame of these standard human motion videos is basically accurate in the detection of human posture (the 2D human skeleton detected in each frame is compared with the real human skeleton points by the tester, and only a few frames in each segment of thousands of frames have defects), and then sample the 2D coordinates of each joint point of human posture, At this time, the sampling is to record the coordinates of each joint point in each frame.

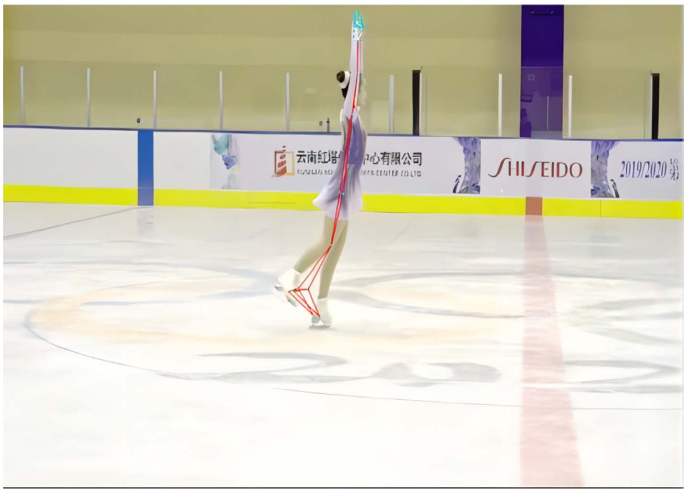

Figure 2. Correct detection of human posture.

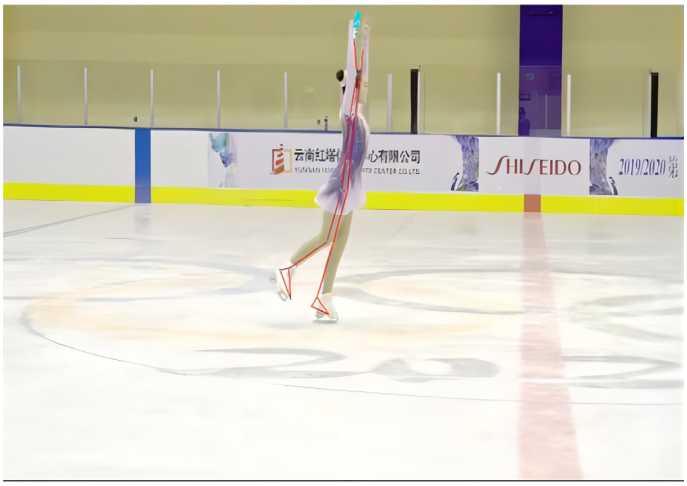

Figure 3. Error detection of human posture.

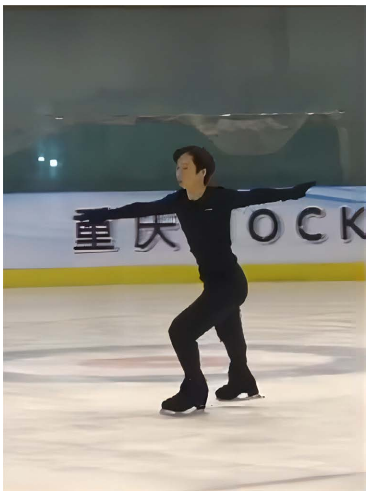

Figure 4. Standard action of skating.

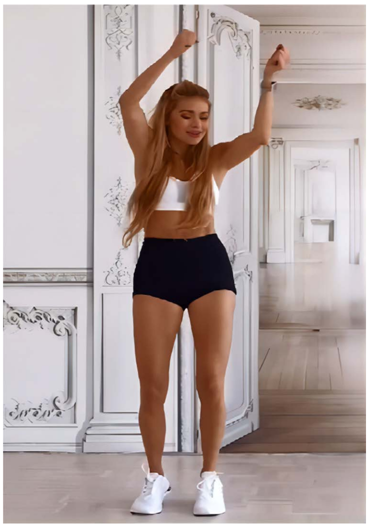

Figure 5. Standard action of indoor fitness.

The coordinate displacement of each joint point j($j \in (1,33)$) is achieved by the 2D coordinate offset distance (P_{i-1}, P_i) of the i − 1 frame data $P_{i-1}(X_{i-1}, Y_{i-1})$ and i frame data $P_i(X_i, Y_i)$:

$$\text{Distance}\left(P_{i-1}, P_i\right) = \sqrt{\left(X_i - X_{i-1}\right)^2 + \left(Y_i - Y_{i-1}\right)^2} \tag{1}$$

The ratio of Distance (P_{i-1}, P_i) to the current height can be calculated to ensure that the human body in the video at different resolutions will not be affected by the height difference:

$$\text{NewDistance}\left(P_{i-1}, P_i\right) = \text{Distance}\left(P_{i-1}, P_i\right) / \text{Height} \tag{2}$$

The time difference between the latest coordinate offset New Distance(P_{i-1}, P_i) and the two adjacent frames ΔT, At the i-th frame of the k-action video, each joint point j will generate each rate change value V_{kji} (in the following formula, k, j and i respectively represent the video action sequence, joint point number and frame number):

$$V_{kji} = \text{NewDistance}\left(P_{i-1}, P_i\right) / \Delta T \tag{3}$$

In this *N*-frames video, the average velocity V_{kj} of each joint point j is:

$$V_{kj} = \left(\sum_{i=2}^{N} V_{kji}\right) / \left(N-1\right) \tag{4}$$

Then 12 tables were generated from 12 action videos, each table recorded the speed average of each joint point *j*, and each joint point calculated the speed average of the 12 tables to generate the speed threshold that can take into account all actions:

$$V_j\left(\text{Threshold}\right) = \left(\sum_{k=1}^{12} V_{kj}\right) / 12 \tag{5}$$

Finally, a speed threshold correction table with 33 joint point speed thresholds was generated.

3.3. Human Body Tilt Statistics

When Mediapipe predicts the human posture of a monocular image, the mapping from a 2D image to a 3D posture is a multi-solution problem, which results in that the detection of 3D human posture by Mediapipe always presents different degrees of tilt (reflected in the fact that the human head is close to the camera and the foot is far away from the camera in the 3D space) regardless of the camera's shooting angle. As shown in **Figure 6** and **Figure 7**, when a person is standing upright, the 3D posture data predicted by Mediapipe for the human body will have the upper body tilt towards the direction of the camera and the lower body tilt towards the opposite direction; As shown in **Figure 8** and **Figure 9**, when the human hand is at a higher height, the height of the hand will cause the human body to tilt more. The higher the height of the hand, the more serious the tilt of the 3D posture data of the human body will be predicted by Mediapipe. This is mainly because Google lacks data collection of different hand heights when training the Mediapipe model.

In order to solve the problem of different degrees of tilt between the 3D human posture predicted by Medipiepe and the real human posture under different hand heights, this paper uses tables to record the relationship between the different heights of human hands and the different tilt degrees of the body in the real world to correct the wrong tilt of the 3D data predicted by Medipiepe under the corresponding tilt degrees.

Specifically, in the real world, when a person maintains a straight standing posture, the upper body tilt tan value is 0, and the upper body tilt tan value predicted by Mediapipe should also be 0. The accurate Z value of the shoulder center point can be obtained by formula (1), ΔY is the Y value difference between the shoulder center point and the hip center point, and Tan(Reality) is the inclination of the human body in reality:

$$Z\left(\text{Reality}\right) = \text{Tan}\left(\text{Reality}\right) * \Delta Y \tag{1}$$

The scaling factor between the real Z value and the original Z value predicted by Mediapipe can be calculated according to formula (2) α:

$$\alpha = Z(\text{Reality})/Z(\text{Mdp}) \tag{2}$$

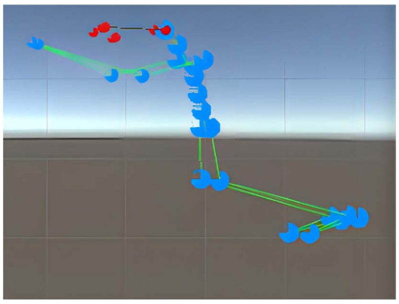

Figure 6. Data of hand lift in 3D.

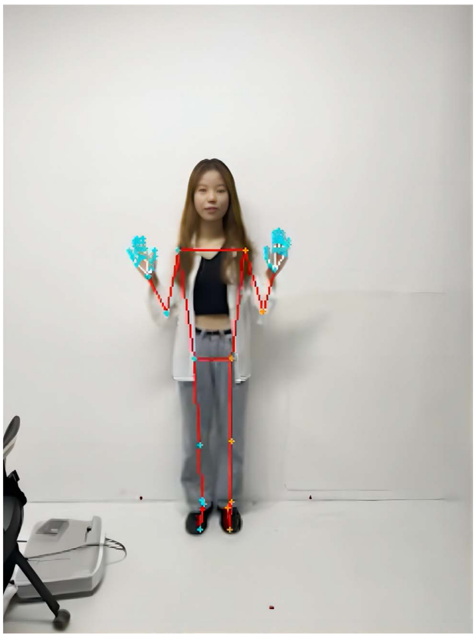

Figure 7. Hands lift in reality.

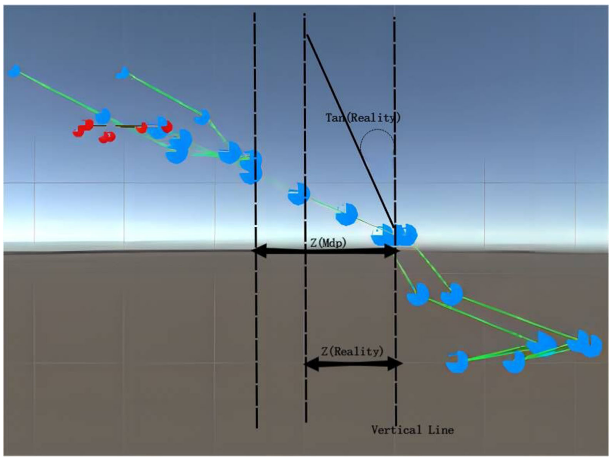

Figure 8. Data of 3D middle hand lifting.

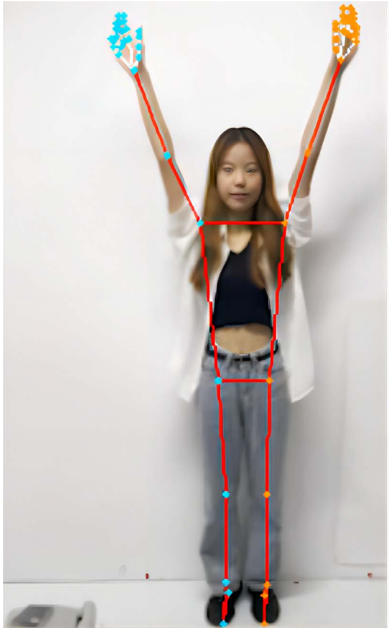

Figure 9. Hands lifting in reality.

Since the inclination of all joint points is consistent, the scaling factor α To find the true Z value of other joint points in turn:

$$Z(\text{Others Reality}) = \alpha * Z(\text{Others Mdp}) \quad (3)$$

The lack of data sets of different hand heights during the training of the Mediapipe model results in the influence of the height of the hand on the inclination of the human body. Therefore, we should also take the height of the hand and the inclination of the upper body as variables, record the true human body inclination and the predicted inclination of Mediapipe under different hand height postures except for the upright standing posture, and obtain the scaling coefficient under different conditions.

As shown in Figure 10, the true inclination when the human hand is raised and tilted 15 degrees and the predicted inclination of the corresponding Mediapipe are respectively. The tilt is corrected by the above formula to make the predicted Z value of Mediapipe more accurate and effective.

According to observation, when the human body leans 90 degrees, the human body posture predicted by Mediapipe is similar to the real human body posture, so the inclination of the upper body is divided into five parts from 0 degrees to 90 degrees. According to the characteristics of the human body, the height of the hand is also divided into five parts from below the crotch to above the head, and other critical values adopt linear interpolation.

3.4. Simulation of Human Limb Proportion

In Mediapipe, when the human body presents some movement postures, such as kicking and extending the arm, the 3D human body data will have limb deformation (the arm or leg will become excessively longer or shorter compared with the real human body limb), which is caused by the inconsistency between the Z value of some joint points predicted by Mediapipe and the real Z value under the complex limb posture of the human body. To solve this problem, the method of human limb simulation [12] can be used to realize the normalization of limb proportion.

Specific implementation: firstly, the eight-headed body (body height = 8 * head length) in the 3D model is used as the human body standard. As shown in Figure 11, a standard human body body proportion table is generated by collecting the proportion of each limb part to the height. When Mediapipe predicts the posture of the human body at each frame, the corresponding Z value is corrected by calculating the proportion of the length of each limb of the current human body to the height.

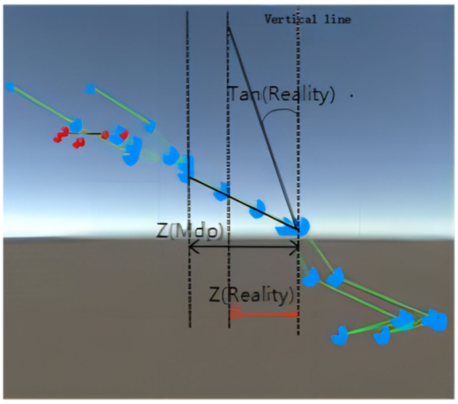

Figure 10. Adjust Z (Mdp) to Z (Reality) when the hand is raised and tilted 15 degrees.

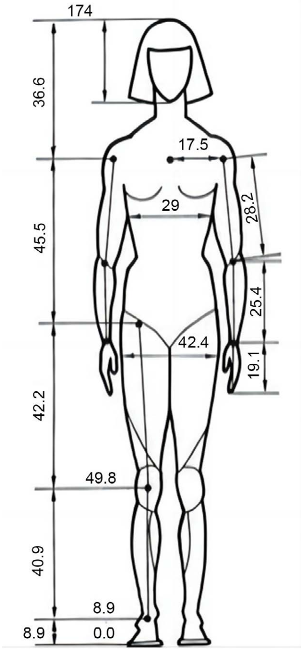

Figure 11. Proportion length of each limb in the eight-headed body.

3.5. Mean Filtering and One-Euro Filtering

When the RGB camera captures the motion, due to hardware problems, the camera itself is out of focus or shooting is unstable, the coordinate values (X, Y, Z values) of the joint point predicted by Mediapipe will be unstable, as shown in the signal diagram in **Figure 12**. Here, the Y value in the hip joint point is taken as an example (X, Y, Z values have the same change effect), and the horizontal axis is time (spacing unit is frame). The vertical axis is the Y value (spacing unit is 1K pixels), and the Y value is 0 when the hip is at the center of the image. The data will have an instantaneous shift from the ideal position in a short time, and there will be the influence of uniform noise in a long time. This paper combines the mean filtering and one-euro filtering algorithm to filter the original data, so as to reduce the dynamic error caused by the data instability.

In the numerical value of human posture movement, the coordinates of the corresponding joint points in the front and back frames will change greatly, so the large numerical fluctuation in a short period is also called pulse interference, which is mainly caused by the noise generated by the speed change of the joint points in a short period of time in numerical value during some actions. To solve this problem, the one-euro filtering algorithm [11] is adopted in this paper, This is a low-pass filter that filters noise signals in real time, mainly to reduce jitter caused by signal change rate. The filter is equipped with two configurable parameters, the minimum cut-off frequency $f_{c\min}$ and speed coefficient β.

$$\hat{X}_i = \alpha * X_i + (1-\alpha) * X_{i-1},\ \ i \geq 2 \tag{1}$$

X is the abscissa value of the joint point (also can be Y, Z), the signal value of T_i at the time of frame i is expressed as X_i, and the filtered signal is expressed as, where the smoothing factor $\alpha \in [0,1]$ is not a constant, but adaptive, that is, it is dynamically calculated using information about signal change rate (speed). The adaptive smoothing factor aims to balance the trade-off between jitter and lag. At low speed, the numerical weight of the joint point at the current time will increase, and the numerical weight of the previous frame will decrease. At low speed, the value on the signal graph will be more stable, thus reducing jitter; At high speed, the numerical weight of the current time will decrease, and the numerical weight of the previous frame will increase. At this time, the lag of the value on the signal graph will be reduced. The smoothing factor is defined as:

$$\alpha = 1\Big/\left(1+\left(1/T_e\right)\right) \tag{2}$$

where T_e is the sampling period calculated according to the time difference between samples:

$$T_e = T_i - T_{i-1} \tag{3}$$

When the video frame rate is 30, the sampling period is usually once for two frames, and T_e is 1/15 second.

τ Is the time constant calculated using the cut-off frequency:

$$\tau = 1/2\pi * f_c \tag{4}$$

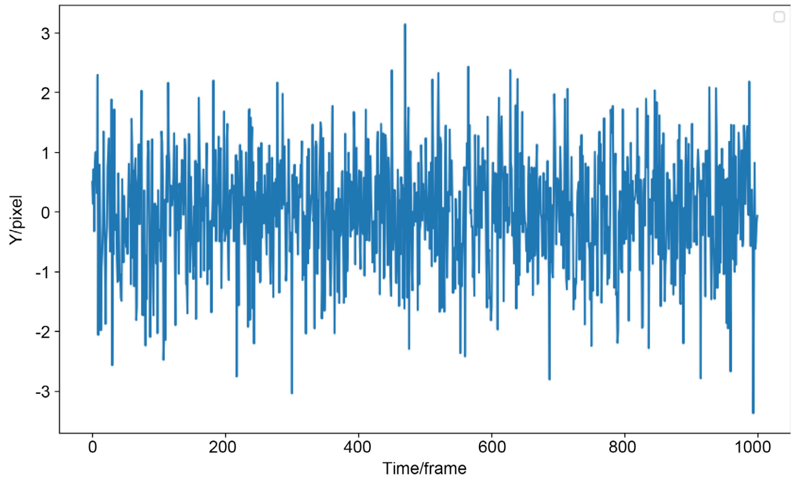

Figure 12. Raw data of head joint point *Y* value.

The cut-off frequency f_c increases linearly with the increase of the rate of change (*i.e.* speed), where $f_{c\min} > 0$ is the minimum cut-off frequency, $\beta > 0$ is the speed coefficient:

$$f_c = f_{c\min} + \beta * |X_i| \tag{5}$$

$X_{i\text{-}1}$ is the unfiltered signal value of the current frame, X_i is the filtered signal value of the previous frame, $X_{i\text{-}1}$ is the rate of change between X_i and $X_{i\text{-}1}$ in the unit time T_e, the rate of change as the discrete derivative of the signal $X_{i\text{-}1}$, $X_{i\text{-}1}$ is calculated as follows:

$$\dot{X}_i = 0$$

$$\dot{X}_i = \left(X_i - \hat{X}_{i-1}\right)\Big/T_e, i \geq 2 \tag{6}$$

There are two configurable parameters in the model, minimum cut-off frequency $f_{c\min}$ and speed coefficient β. At low speed, a low cutoff stabilizes the signal by reducing jitter, but with the increase of speed, the cutoff is increased to reduce delay. In order to minimize jitter and lag when tracking human motion, a simple two-step procedure can be used to set two parameters. When the body parts remain stable or move at a very low speed, reduce $f_{c\min}$ to reduce jitter again, and maintain acceptable delayed movement during these slow processes. Next, the human body moves rapidly in different directions, and β will increase. The key is to minimize delay. If high speed hysteresis is a problem, increase β; If slow jitter is a problem, reduce $f_{c\min}$. Finally, by reducing the minimum cut-off frequency $f_{c\min}$, the numerical jitter at slow speed will be reduced and the data will be more stable; Increase speed coefficient β. It will reduce the lag caused by fast speed and make the information more accurate and effective. Through many tests, the default setting of β is 0, and $f_{c\min}$ is set to 1.

The second kind of filtering is the mean filtering algorithm, which is mainly used to keep the value stable in the overall time period so as to smooth the overall data. During filtering, each coordinate value of each joint point is averaged within N frames. When the value of N is large, the filtered joint point value is relatively smooth, but the sensitivity is poor, that is, the data of the N-th frame node is affected by the data of the previous $N - 1$ frame, which seriously weakens the proportion of the frame node data in this frame, resulting in the reduction of the action amplitude compared with the real degree; On the contrary, when the value of N is small, the filtering and smoothing effect is poor, but the sensitivity is good, and the visual effect will show that the action amplitude has a certain degree of increase. There are two modes for Mediapipe operation: 1) only running CPU 2) CPU and GPU calling at the same time. After testing, the effect of taking 5 frames of N is the best in the mode of running CPU only, and the effect of smoothing is ideal when taking 2 frames of N in the mode of CPU and GPU calling simultaneously.

4. Experiment and Results Analysis

4.1. The Experimental Environment

In order to verify the effectiveness of the improved Mediapipe proposed in this paper for 3D human posture detection, the following comparative experiments were carried out. The experimental environment is CPU Intel Core i7-8750H, main frequency 2.21 GHz, memory 16GB, GPU GeForce GTX 1050 Ti, memory 4096 M, operating system Ubuntu 18.04, Mediapipe version 0.8.6, development environment G++8.40, Bazel 4.20. This paper uses 30 FPS video or 30 frame rate USB camera to shoot in indoor or outdoor scenes. The test population is between the ages of 12 and 50, with a height of 1.5 m - 1.95 m and a weight of 40 kg - 90 kg, covering men, women and children.

4.2. 2D Detection of Human Posture

According to the problem of inaccurate detection of 2D human posture by Mediapipe, the human posture was corrected by the speed threshold correction method, and the recognition test of human posture was carried out through 52 non-standard videos (the video length is within 300 - 1500 frames, and the video is the non-standard video that caused the inaccurate detection of Mediapipe due to the change of light environment).

This paper verifies the feasibility of the detection experiment by evaluating the average accuracy of a single frame (the ratio of the number of accurate joint points to the total number of joint points) and the overall accuracy of the video (the accuracy of the correct frame to all frames). The data is shown in Table 1:

Table 1. Comparison of objective evaluation indexes of 2D human posture detection before and after the improvement of Mediapipe.

Evaluation index/algorithm	The original one, the Mediapipe	Subhead
Average accuracy	71.76%	93.52%
Overall accuracy	75.42%	96.28%

From the experimental data in Table 1, we can see that the average accuracy of Mediapipe's 2D human posture detection data after the speed threshold method has been improved from 71.76% to 93.52% through the test of 52 non-standard videos. It can be seen from the effect pictures in Figure 13 and Figure 14 that the recognition of human skeleton points by the speed threshold method is closer to the real human skeleton points in the 2D plane effect, and is more accurate than the original Mediapipe in the detection of 2D human posture.

4.3. Detection of *Z* Value in 3D Human Posture

Aiming at the problem that Mediapipe can not detect the *Z* value of 3D human posture accurately due to human body tilt, this paper first modifies the human posture through human body tilt statistical method, and basically achieves the consistency of the predicted human posture of Mediapipe and the tilt of human posture in the real world. By calculating the tilt value of the two, the accuracy reaches more than 80%. The corrected effect is as follows (Figure 15):

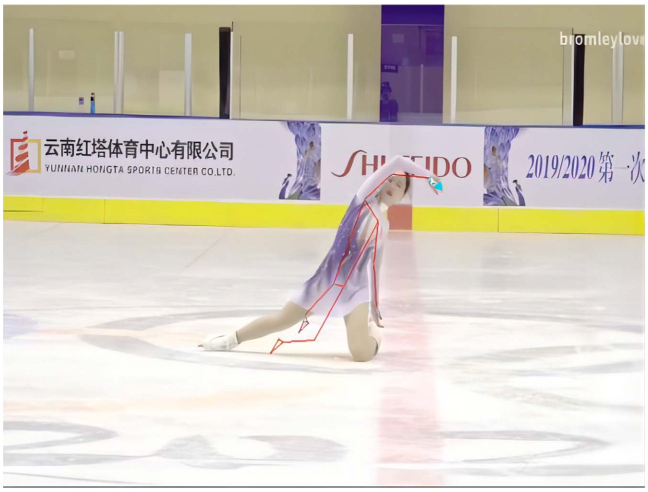

Figure 13. Misrecognition of human posture by Midiapipe.

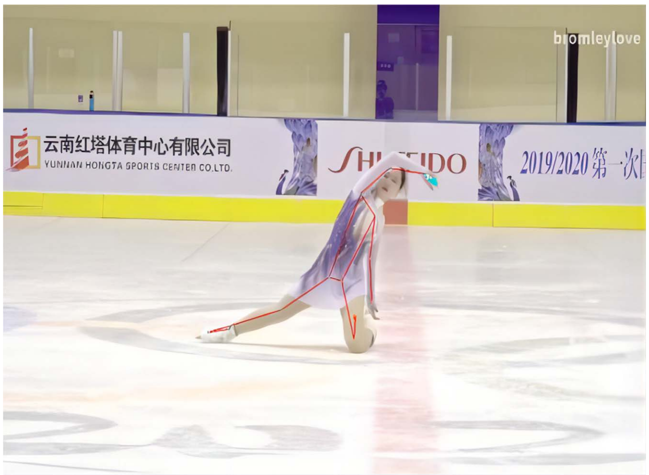

Figure 14. Correction of human posture recognition by velocity threshold method.

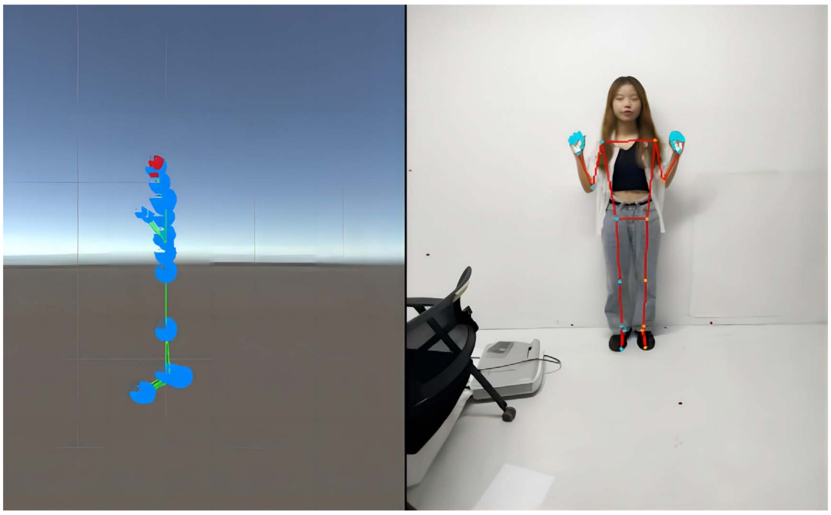

Figure 15. Modified Mediapipe's prediction of 3D human posture.

In view of the problem that Mediapipe can not detect the Z value in 3D human posture accurately due to the large body posture, this paper modifies the human posture through the human simulation scale normalization method, as shown in **Figures 16-18** below. When a person does a leg lift, the corrected limb data is more accurate than the original data. Through the standard human body proportion table, each frame of human body posture data predicted by Mediapipe has been accurately adjusted with the relevant Z value data, effectively eliminating the stretch deformation in the visual effect caused by the excessive amplitude of human body limbs, and further improving the recognition accuracy of human body posture.

4.4. Filtering of 3D Human Posture Data

Finally, according to the problem of pulse noise and uniform noise caused by too fast movement of joint points due to hardware problems, camera itself out of focus or shooting instability when capturing the action of RGB camera, the above video samples are used for one-euro filtering and mean filtering respectively. The effect image intercepted by multiple filters under long video is shown in **Figure 19**. Here, the Y value is taken as an example, and the horizontal axis is time (spacing unit is frame), The vertical axis is the Y value (the spacing unit is 1K pixels), which represents the broken line signal graph of the five filters compared with the original data and the data after adding noise. From the graph, it can be seen that the moving average filter, single exponential filter, double exponential filter and Kalman filter have caused the loss of some peak values to a large extent, and the loss of action detail data is serious; In addition, there is a serious lag problem at high speed, while one-euro filtering has the characteristics of being more sensitive to jitter at low speed of the joint point, and can reduce the lag at high speed, which not only retains the details of the original data to a great extent, but also effectively solves the problem of numerical jitter and numerical lag caused by the speed change of the joint point.

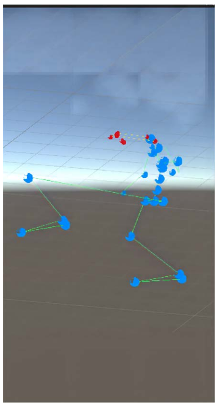

Figure 16. Original data.

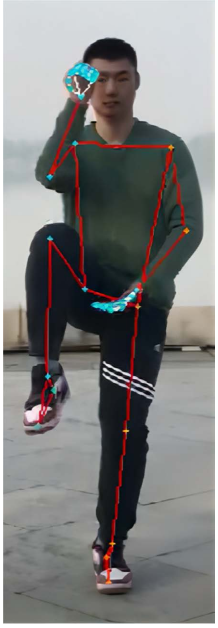

Figure 17. Corrected data of detected human.

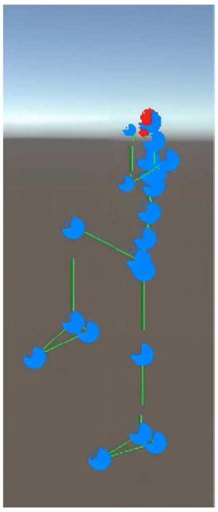

Figure 18. Corrected data.

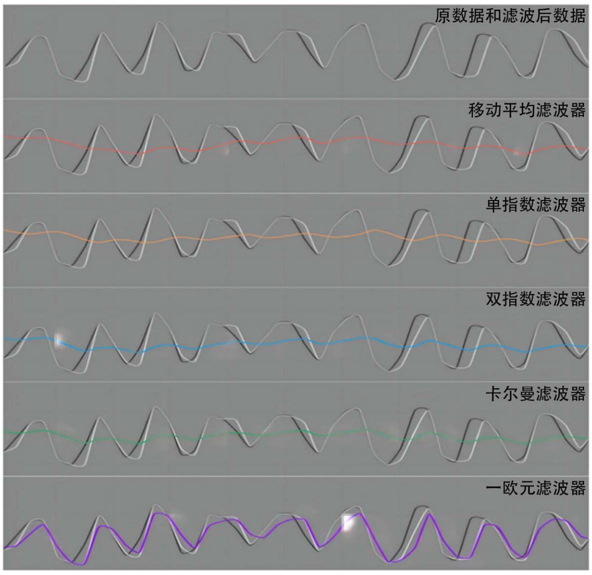

Figure 19. Filtering diagram of human joint point values by different filters at different times.

In the mean filtering, as shown in **Figure 20**, the blue color of the signal graph is the original coordinate value of Mediapipe, and the green color is the coordinate value after the mean value. Here, take the *Z* value as an example, the horizontal axis is the time (spacing unit is frame), and the vertical axis is the *Z* value (spacing unit is 1 K pixel). Compared with the original value, this filter has a good suppression effect on periodic interference, high smoothness, and can effectively remove the impact of noise caused by the rapid change of non-joint points over a long period of time, especially for the noise generated by uniform distribution. It has a very good smoothing effect on the overall value, and the visual effect is smoother.

This paper verifies the feasibility of the detection experiment by evaluating the average accuracy rate of a single frame (the ratio of the number of accurate joint points to the total number of joint points) and the overall accuracy rate of the video (the accuracy rate of the correct frame to all frames) of the effect image of *Z* value detection in 3D human posture. The data is shown in **Table 2**.

From the experimental data in **Table 2**, it can be seen that the average accuracy rate of 2D human posture detection data of Mediapipe through human tilt statistics, limb simulation, one-euro filtering and mean filtering has increased from 32.43% to 95.87%, and the overall accuracy rate has increased from 34.26% to 96.12%, through a total of 500 tests on 20 people of different height, weight, age and sex, Significantly improved Mediapipe's detection of 3D human posture. It can be seen from the effect diagram in **Figures 15-20** that the improved Mediapipe is more accurate than the original Mediapipe in detecting the *Z* value of the 3D human posture. The 3D data of the human body is closer to the real human posture in space, and the noise removal effect is obvious on the digital signal diagram. Finally, the movement of the characters in the overall video is more smooth and smooth.

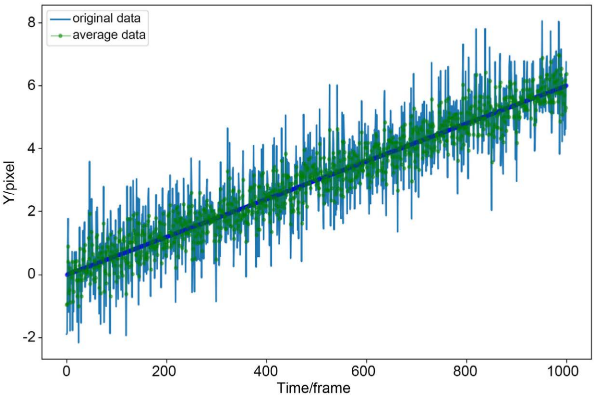

Figure 20. Average filtering.

Table 2. Comparison of objective evaluation indexes of *Z* value in 3D human posture detection before and after improvement of Mediapipe.

Evaluation index/algorithm	The original one, the mediapipe	Human body tilt statistics method	Body simulation method	One euro filte	Mean filter
Average accuracy	32.43%	81.57%	91.54%	93.916%	95.87%
Overall accuracy	34.26%	82.58%	92.47%	93.979%	96.12%

5. Conclusion

According to Mediapipe's inaccurate detection of human posture, The accuracy of 2D human posture detection is improved by correcting the rate of each joint; The inaccurate prediction of depth *Z* value in human posture data by monocular camera, Z-correction of the human tilt angle, Improve the accuracy of the human body on the *Z* value of the joint point under different tilt degrees; According to the problem of inaccurate identification of the limb joint point *Z* value caused by the large limb posture of the human body, By integrating the simulated proportion of the human body, The accurate correction of *Z* value of human posture under different limb posture is realized; Finally, for the periodic noise problem of inaccurate detection and continuous multiple frames due to the excessive speed of human joints, Euro filtering and mean filtering of the data were performed on the joint points, The problem of jitter lag and uniform noise in the point speed change of human joints is solved. Finally, the detection of 3D human posture was improved by the improved Mediapipe. In the future, the research on human posture will be more popular in the computer field. According to the accurate 3D human posture data, corresponding motion recognition, behavior analysis, rehabilitation detection and human-computer information interaction can be carried out, which has a broad development prospect.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Rahul, M. (2018) Review on Motion Capture Technology. *Global Journal of Computer Science and Technology*, **18**, 1-F. https://typeset.io/papers/review-on-motion-capture-technology-3nao8jb6kv

[2] Kim, J., Campbell, A.S., de Ávila, B.E.F., *et al.* (2019) Wearable Biosensors for Healthcare Monitoring. *Nature Biotechnology*, **37**, 389-406. https://www.nature.com/articles/s41587-019-0045-y

[3] González, L., Álvarez, J.C., López, A.M and Diego, Á. (2021) Metrological Evaluation of Human-Robot Collaborative Environments Based on Optical Motion Capture Systems. *Sensors*, **21**, 3748. https://doi.org/10.3390/s21113748

[4] Xiao, W.H. (2017) Research on Label-Free Surgical Navigation Technology Based on Structural Light. Master's Thesis, South China University of Technology.

[5] Nguyen, H.C., Nguyen, T.H., Scherer, R. and Le, V.H. (2022) Unified End-to-End YOLOv5-HR-TCM Framework for Automatic 2D/3D Human Pose Estimation for Real-Time Applications. *Sensors*, **22**, 5419. https://www.mdpi.com/1424-8220/22/14/5419/xml

[6] Liu, Y.X. (2021) 2D Human Pose Estimation Based on Self-Distillation. Master's Thesis, University of ESTC. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202201&filename=1021748638.nh

[7] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. and Sheikh, Y. (2021) OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 172-186. https://ieeexplore.ieee.org/document/8765346

[8] Wang, J. (2021) 3D Human Pose Estimation and Human Image Generation Application. Doctor's Thesis, National University of Defense Technology. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CDFDTEMP&filename=1022809939.nh

[9] Kwon, H.M. and Seo, J. (2022) Effect of Compressed Sensing Rates and Video Resolutions on a PoseNet Model in an AIoT System. *Applied Sciences*, **12**, 9938. https://www.x-mol.com/paper/1576984976760037376

[10] Zeadally, S., Zomaya, A. and Chao, H.-C. (2009) Editorial for Special Issue of Telecommunication Systems on "Mobility Management and Wireless Access". *Telecommunication Systems*, **42**, 163-164. https://schlr.cnki.net/zn/Detail/index/GARJ0010_5/SPQD00002265649

[11] Gökhan, G., Jansen, T.S., *et al.* (2022) Video-Based Hand Movement Analysis of Parkinson Patients before and after Medication Using High-Frame-Rate Videos and MediaPipe. *Sensors*, **22**, 7992. https://www.mdpi.com/1424-8220/22/20/7992/html

[12] Wang, W.Y., Fu, C. and Cao, F. (2021) Kinematic Analysis and Gait Simulation of Humanoid Robots Based on Motion Capture Technology. *Mechanical Transmission*, **45**, 110-117. https://www.cnki.com.cn/Article/CJFDTOTAL-JXCD202108016.htm