The International Journal of Engineering and Science (IJES)
|| Volume || 13 || Issue || 4 || Pages || PP 08-17 || 2024 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Squat Training Assistance System based on MediaPipe

Cheng-Hang Tsai, Ke-Nung Huang*

Department of Electronic Engineering, I-Shou University, Kaohsiung, Taiwan

** Corresponding Author*

----------ABSTRACT----------

It is crucial to maintain good exercise habits for health. Home workouts have always been a choice for many people. Although many home workout training equipment are effective, their high prices hinder their widespread adoption and promotion to the public. To make home workouts more widely adopted among the public, this study will design a squat training assistance system. This system will take the real-time training images captured by the video camera, use the MediaPipe pose estimation module to obtain the coordinates of 33 three-dimensional skeletal landmarks of the whole body, and by calculating the vector relationships between these landmarks, obtain the angle changes between different parts of the human body under different planes. In order to improve accuracy, methods such as smoothing landmark coordinate, visibility condition setting, and initial setting of the angle between coordinate vectors are used in reducing the errors in the coordinate points output by the model during exercise caused by factors such as environment, body movement, and occlusion, and overcoming the shooting angle problem caused by single-camera shooting. Finally, the current exercise image and some information will be displayed on the screen, and relevant exercise data will be provided subsequently.

KEYWORDS;- *MediaPipe, landmark, Pose estimation, Exercise training*

Date of Submission: 02-04-2024 Date of acceptance: 13-04-2024

I. INTRODUCTION

Due to the global pandemic of Covid-19, people have been unable to go out for exercise, and the original gym coaching courses have also been completely suspended or changed to online teaching. This has led to a rapid rise in the trend of home workouts. Unlike other exercises that require various professional large-scale equipment, the characteristics of home workouts are simplicity and convenience. By adhering to the relevant exercise rules, you can achieve efficient exercise. For example, squatting exercise [1] is a kind of exercise that can simply train the core and lower body muscles by using your own body weight.

There are several ways to capture human posture, such as using wearable devices [2], which wear sensors on the body to quantitatively grade human posture. This method has high accuracy, but it is not suitable for long-term use due to maintenance and comfort issues. There are also studies using Kinect [3]. Kinect, in addition to a general camera, is also equipped with a depth sensor composed of infrared transceivers. From this image, 3D coordinate data can be obtained, and corresponding coordinates can be given to various parts of the body to achieve posture estimation. It has higher accuracy than a general camera, but the cost is relatively higher. Finally, the human key-point coordinates from open-source libraries such as OpenPose [4], MediaPipe [5] can be used to estimate the posture from the 2D image captured by a general camera. This method is not as accurate as the first two, but there are no special requirements for the quality of the input image. The computer's video camera or a microcontroller equipped with a small lens can also be used to achieve the function of posture estimation. Although the quality of the image will affect the accuracy of the results, considering that not all postures that need to be recognized require a high degree of accuracy, this study will use this method, which is most suitable for the general public.

There are two methods for posture estimation using the posture estimation module. The first one is to use the KNN (K-Nearest Neighbors) algorithm [6]. For example, when counting the number of pull-ups, two types of pull-up exercise states, pull-up and put-down, will be introduced first. These samples are referred to as classified samples, and new exercise postures will be input into the system as unclassified samples. Under the processing of the algorithm, these new exercise postures can be classified according to the classified samples. But this method is only suitable for exercises with fewer action states, such as squats, sit-ups, pull-ups, etc. The second method involves calculating the angle between the vectors of each part to judge whether the action is completed. Although squatting is also an exercise with fewer actions, the goal of this study is not just to count the number of exercises. Instead, it seeks to be able to more accurately judge whether the squat is correct and to achieve other functions, so this study adopts the second method. In a single lens, the most common practice is to

preset the angle of the target posture [7]. For example, in the squat exercise, the knee angle on the 2D screen will be set to 180 degrees when standing and 90 degrees when squatting to judge the exercise state of the subject. But this method of presetting angles on a 2D plane will cause serious errors due to the shooting angle. This study is based on the three-dimensional coordinates of MediaPipe. It will be observed from three different planes *xy, xz, yz* when analyzing the angle changes in the squat exercise, and the relative exercise state will be calculated from the relative vectors.

II. MATERIAL AND METHOD

The system flowchart of this study is shown in Figure 1. When the image is input, the visibility of the landmark coordinates of the head, knees, and soles of the feet will be set separately. The purpose is to ensure that the person and related parts are clearly present in the picture. After confirmation, the landmark coordinates are smoothed. Then, the knee angle and sole vector when standing are stored as the basis for exercise judgment. Finally, the results of the exercise data analysis will be displayed on the screen, including exercise status, number of exercises, exercise rate, and the number of times the left and right soles are lifted, as well as the generation of exercise-related charts, Excel files, and exercise videos.

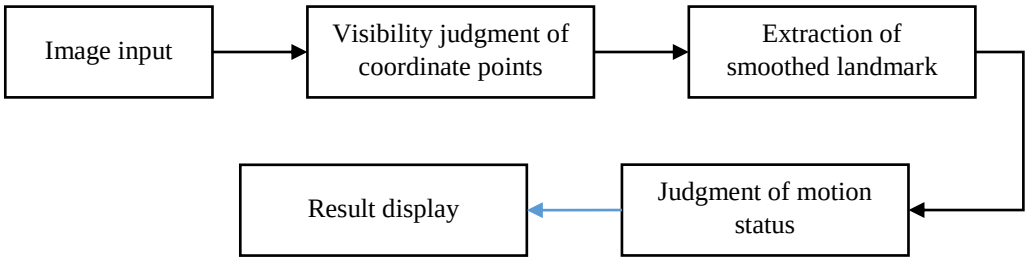

Figure 1 System flow chart

Visibility judgment of coordinate points.

This system divides the visibility judgment into three blocks. The first block is the head. The MediaPipe pose estimation module determines if there is a person in the picture based on whether the head (11 landmarks numbered 0~10 in Figure 2) exists, because it is the strongest feature of the whole body. That is, the clearer the head image, the more accurate the landmark coordinates obtained. To ensure the accuracy of the overall system, the average visibility of the 11 landmarks on the head must be greater than 0.98. The second block is the knee. Using the six-point coordinates of the left and right buttocks, knees, and ankles of the body (points numbered 23~28 in Figure 2) for judgment, their individual visibility must be greater than 0.9 to calculate the angle. If the number of coordinate points on the right foot that meet the threshold is less than the left foot, the right knee angle is replaced by the left foot, and vice versa. If all coordinate points do not meet the set threshold, the exercise state will not be evaluated. The third block is the sole of the foot. Using the six-point coordinates of the left and right soles of the foot (points 27~32 in Figure 2) for judgment, their individual visibility must be greater than the set threshold, otherwise the judgment of whether the sole is lifted will not be performed. Finally, the system will display corresponding prompts on the screen for unclear parts.

Smoothing

Before implementing the squat system, this study will smooth all landmark coordinates. The purpose is to eliminate the extremely unreasonable situations caused by some external factors in the coordinate data output by MediaPipe. This involves reducing the impact of short-term data by the long-term characteristics of coordinate data. This system utilizes the Exponential Moving Average (EMA) as the smoothing method. The weights of each value decrease exponentially with time (frame), and the most recent data has the highest weight. $EMA_f(n)$is the average coordinate value of the current screen, where f is the current screen, P_f is the landmark coordinate value, n is the rolling window, and $0 < \alpha < 1$ is the smoothing factor. The value of α determines the weight distribution of new data and old data. When α is larger, the model reacts faster to new data and relies more on recent data. When the value of α is smaller, the model reacts slower to new data and relies more on distant data. In this study, the default value is $n = 30$ ، $\alpha = 2 / (n + 1)$, so you can determine the weight distribution of new data and old data based on the selected window size.

$$EMA_f(n) = \frac{P_f+(1-\alpha)P_{f-1}+(1-\alpha)^2P_{f-2}+\cdots+(1-\alpha)^{n-1}P_{f-n+1}}{1+(1-\alpha)+(1-\alpha)^2+\cdots+(1-\alpha)^{n-1}} \qquad (1)$$

Judgment of motion status

Using the horizontal coordinates of the left and right shoulders and the left and right hips in the current frame, namely x_{sl}, x_{sr}, x_{pl}, x_{pr} in

Figure **3**, and comparing them with the four-point coordinate values of the previous frame, it can be determined whether the subject is standing steady in the screen. After standing steady, the vertical coordinates y_{sl}, y_{sr}, y_{pl}, y_{pr} are judged. If the values of these four points are all greater than the values of the previous screen, it can be known that the subject is moving downwards, and vice versa.

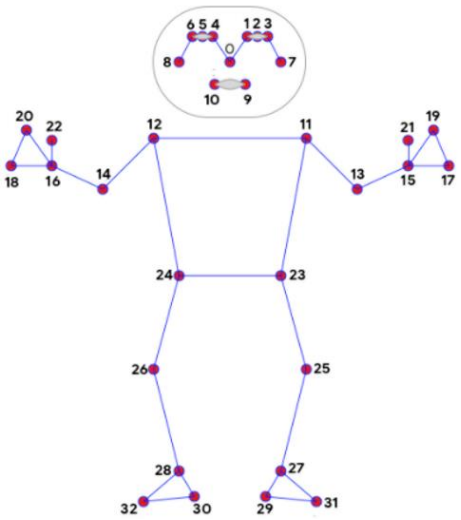

Figure 2 Topology of 33 landmark coordinates of MediaPipe Pose[8]

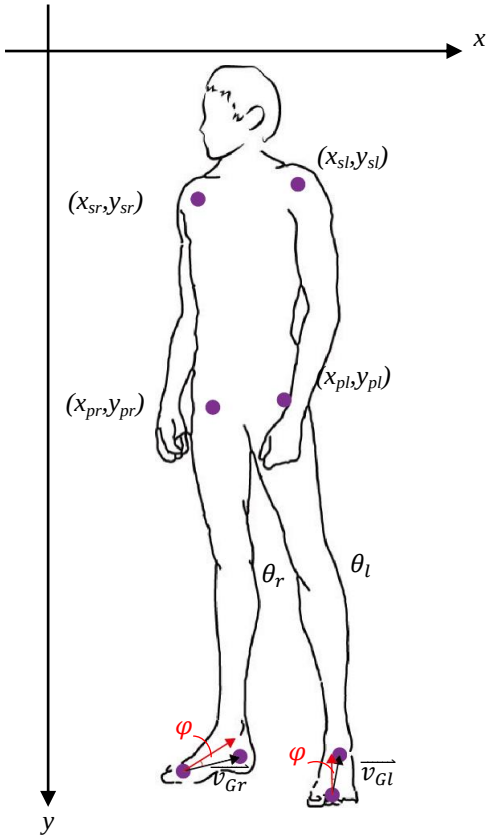

Figure 3 Standing state

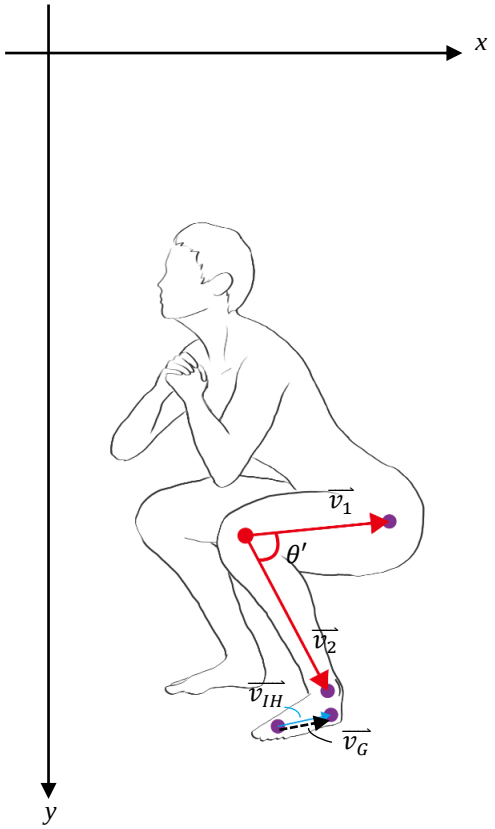

Figure 4 Squatting stat

The determination of the standing state is based on the knee angle θ (

Figure **3**) of the subject before each squat. That is, during the upward movement of the subject, as long as the current knee angle θ '(

Figure **4**) reaches the knee angle θ before each squat, it is judged as a standing state. The calculation of the knee angle is as shown in formula (2), where $\overrightarrow{v_1}$ is the vector from the knee to the hip, that is, the thigh vector, and $\overrightarrow{v_2}$ is the vector from the knee to the ankle. When the thigh is almost parallel to the ground, it is the squatting state in the squat exercise. The ground is represented by the vector $\overrightarrow{v_G}$ from the toe to the heel when standing, and the angle between $\overrightarrow{v_1}$ and $\overrightarrow{v_G}$ is calculated in the same way as formula (2). When the obtained angle is almost 0 degrees, it is judged as a squatting state. During the squatting process, because the center of gravity is too inclined to the toes, the heel is lifted off the ground. Therefore, the calculation of the angle of the heel lift is to use $\overrightarrow{v_G}$ and the toe to the heel $\overrightarrow{v_{IH}}$ during the movement, and the angle φ between the two vectors is calculated in the same way as formula (2). This angle is the angle of the sole lift.

$$\theta' = arccos\frac{\overrightarrow{v_1}\cdot\overrightarrow{v_2}}{|\overrightarrow{v_1}|\times|\overrightarrow{v_2}|} \qquad (2)$$

MediaPipe also provides information on the z-coordinate, which represents the distance between the human body and the camera. The aforementioned angle calculations are all based on the *xy* plane, but the angle between two vectors in three dimensions must be observed from different planes. For example, if you want to calculate the angle between the thigh and the ground, as shown in Figure 5a~d. (the red arrow in Figure 5c is set out of the paper), you need to consider the relationship between the two vectors in three planes at the same time. The same goes for the knee angle and the angle of foot lift.

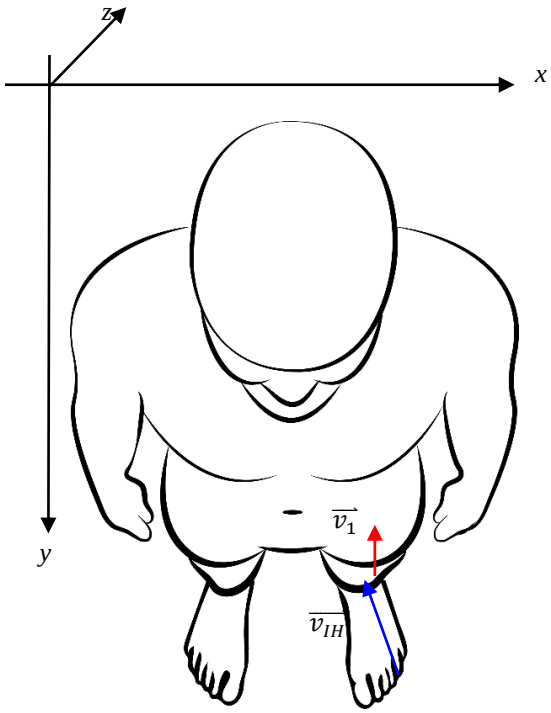

Figure 5a Top view of the human body

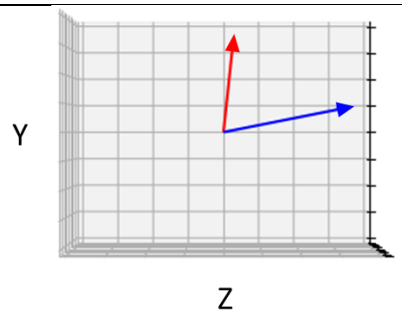

Figure 5b Angle between the thigh and the ground (yz plane)

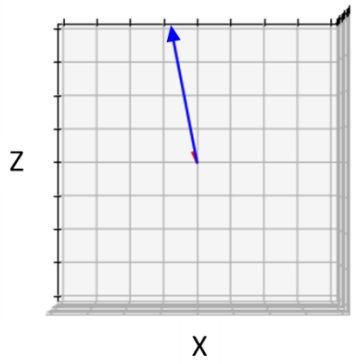

Figure 5c Angle between the thigh and the ground (xz plane)

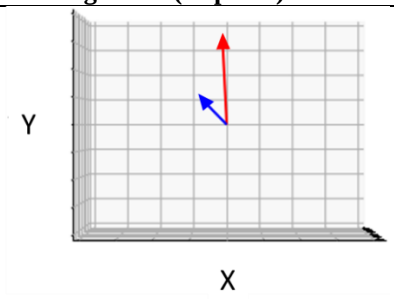

Figure 5d Angle between the thigh and the ground (xy plane)

III. RESULT and DISCUSSION

The hardware and software configuration of this system is an Intel® Core™ i7-8750H CPU @ 2.20GHz processor, 16GB of memory, and PyCharm (Python 3.10) is used as the development environment. The resolution of the test images inputted is 1080 x 1920 pixels. The settings of MediaPipe are all set to default.

The image source can use the computer's webcam as a real-time input (Figure 6), or other external devices, such as using an iPhone 13 (front camera) as a wireless camera in this study (Figure 7). Apart from the difference in image color refinement, at the same resolution settings, the number of frames that can be processed per second is between 18~22 frames. If a webcam is used as the image input, the wiring issue needs to be considered separately, but its cost is relatively low. On the other hand, using a mobile phone as a wireless camera has higher convenience.

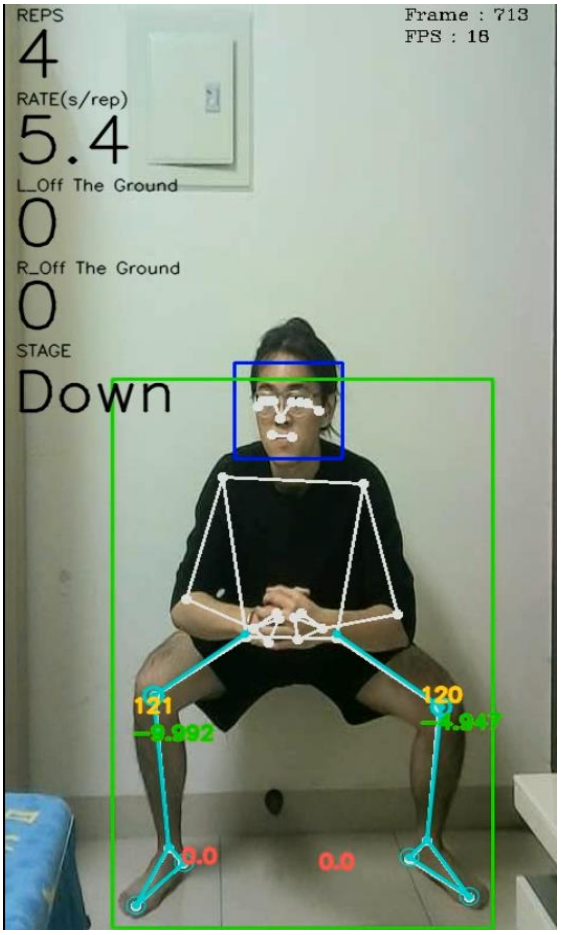

Figure 6 Real-time image from the computer webcam.

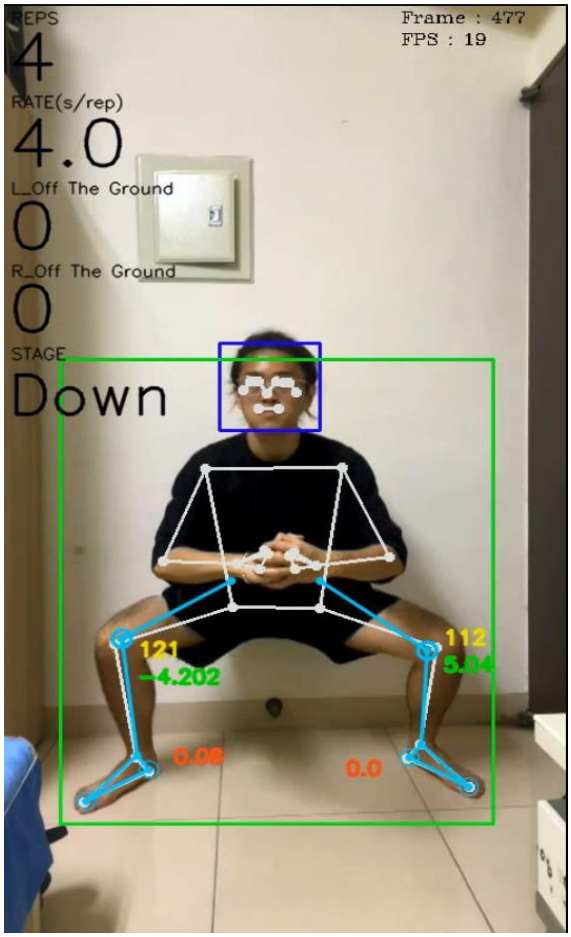

Figure 7 Real-time image from the mobile phone.

In addition to the image input methods mentioned earlier, you can also use pre-recorded videos as input. In this case, the average frame rate is 25 frames per second. Considering the convenience of testing, the screens and data charts presented in this study were all pre-recorded using the front camera of an iPhone 13 at a resolution of 1080 x 1920, with a frame rate of 30 FPS, and then input. The blue lines drawn on the Figure 8 character represent the smoothed results of the lower body coordinate points. The left side of the character represents motion information in the following order: REPS for the number of motions, RATE for the motion rate, L/R_Off. The Ground for the number of times the left and right feet are lifted, and STAGE for the motion state. The three values on the left and right feet of the character form a set. The first value represents the angle when x=0 (yz plane), the second value represents the angle when y=0 (xz plane), and the third value represents the angle when z=0 (xy plane), and they correspond to the angle between the thigh and the ground, the knee angle, and the foot lift angle from top to bottom. When the character is unclear, it will display corresponding prompts, for example, Figure 9 represents the character is still moving, and Figure 10 depicts the character's left foot being obstructed.

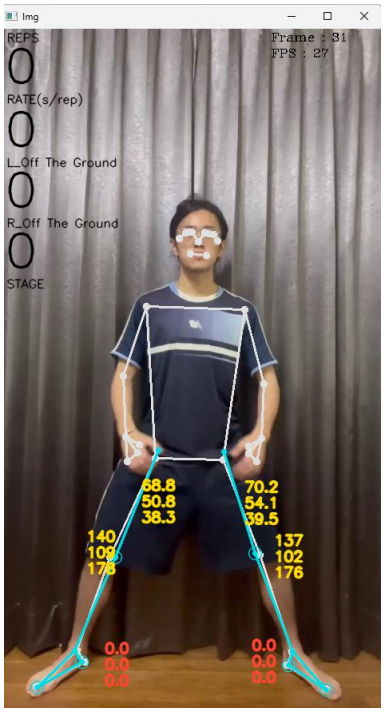

Figure 8 Initial stage of the person. At this time the knee angle at z=0 (xy plane) will be recorded, which is 178 degrees. The reason will be explained in the next paragraph.

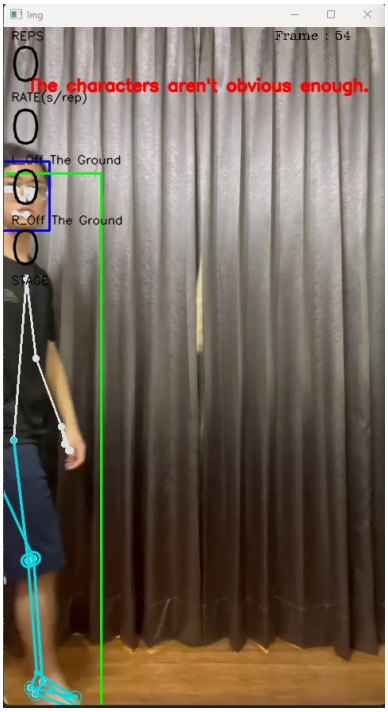

Figure 9 The person is still moving, the screen will display "The characters aren't obvious enough."

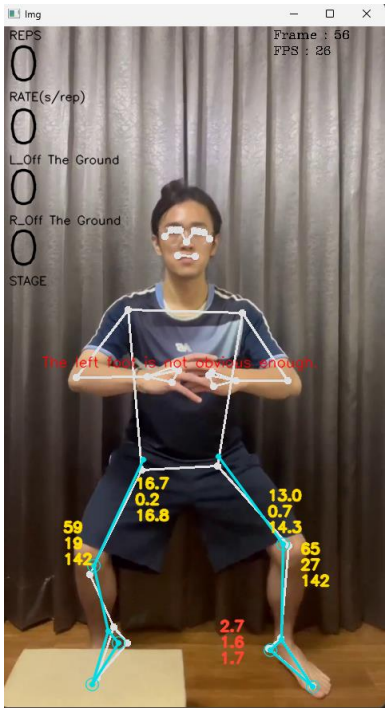

Figure 10 The left foot is obscured, the screen will display "The left foot is not obvious enough."

The charts generated after the exercise are shown in Figure 11 to Figure 14. The subject is facing the camera head-on. The exercise starts at around frame 100, with a total of about 1600 frames. The number of squats is 10 times, with each time squatting to a position where the thighs are almost parallel to the ground, and also returning to a standing position. During the exercise, the foot is sometimes lifted. The knee angle at x=0 (yz plane), y=0 (xz plane) has a difference of about 20 degrees between the left and right feet at the high point (Figure 11), which is due to the error of MediaPipe itself, rather than the body is not facing the camera. It deviates from the ideal of approaching 180 degrees when standing. Therefore, only the angle at z=0 (xy plane) is used as the basis for deciding whether to return to the standing state (Figure 12), and it is determined by whether the current knee angle (orange circle) reaches the knee angle of the previous standing (orange dashed circle).

As shown in Figure 13, when the angle between the thigh and the ground is less than 0 degrees, it means that the thigh has squatted to be parallel to the ground. At this time, the angle in the block where x=0, y=0, and z=0 will be zero, as shown by the black dashed line in the figure. This means that when the subject squats to the thigh parallel to the ground, regardless of the plane being observed, the two vectors will be parallel. At this time, you can directly judge whether the squat is successful based on whether the angle between the two vectors in space is zero, as indicated by the green circle in the figure. This does not mean that this angle can be directly used as the angle between the thigh and the ground, but only uses whether it is zero as the basis for judgment. Therefore, in the output screen, it still shows the angle when x=0, y=0, and z=0. The program's decision is directly based on whether the angle between the two vectors in space is zero.

As shown in Figure 14, during the exercise, no matter whether it is on the xy, xz or yz plane, there should not be any observed changes in angle. Therefore, in the three-dimensional assessment of foot lift, as long as the angle exceeding 5 degrees on one plane is considered lifted. For example, in the figure, the x=0, y=0 block has a lifting situation (purple circle), but the z=0 block does not. The green circle is the opposite scenario, and the aforementioned lifting situations must be recorded one by one.

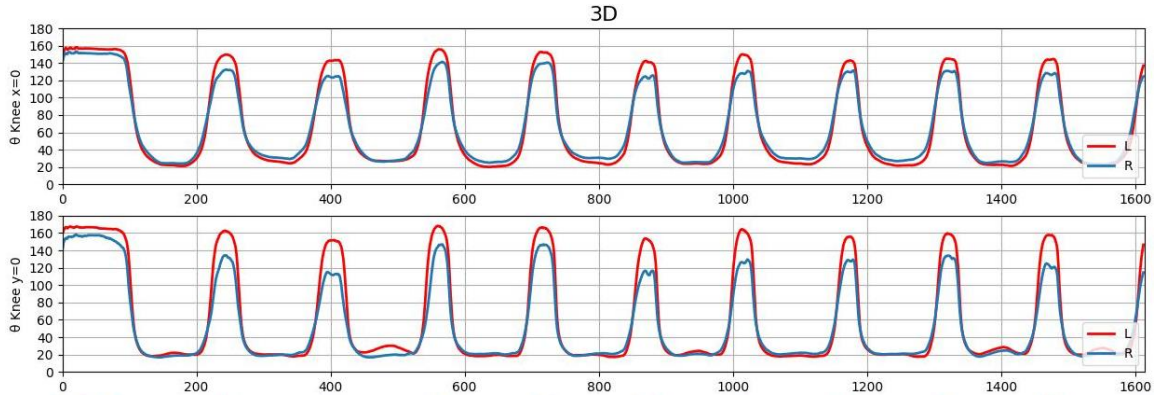

Figure 11 Knee angle at x=0 (yz plane), y=0 (xz plane)

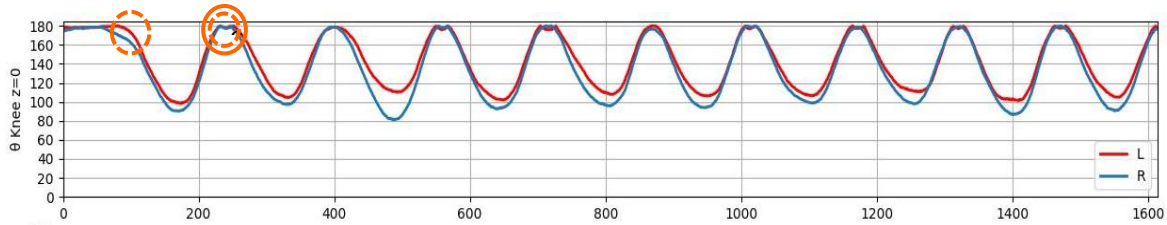

Figure 12 Knee angle at z=0 (xy plane)

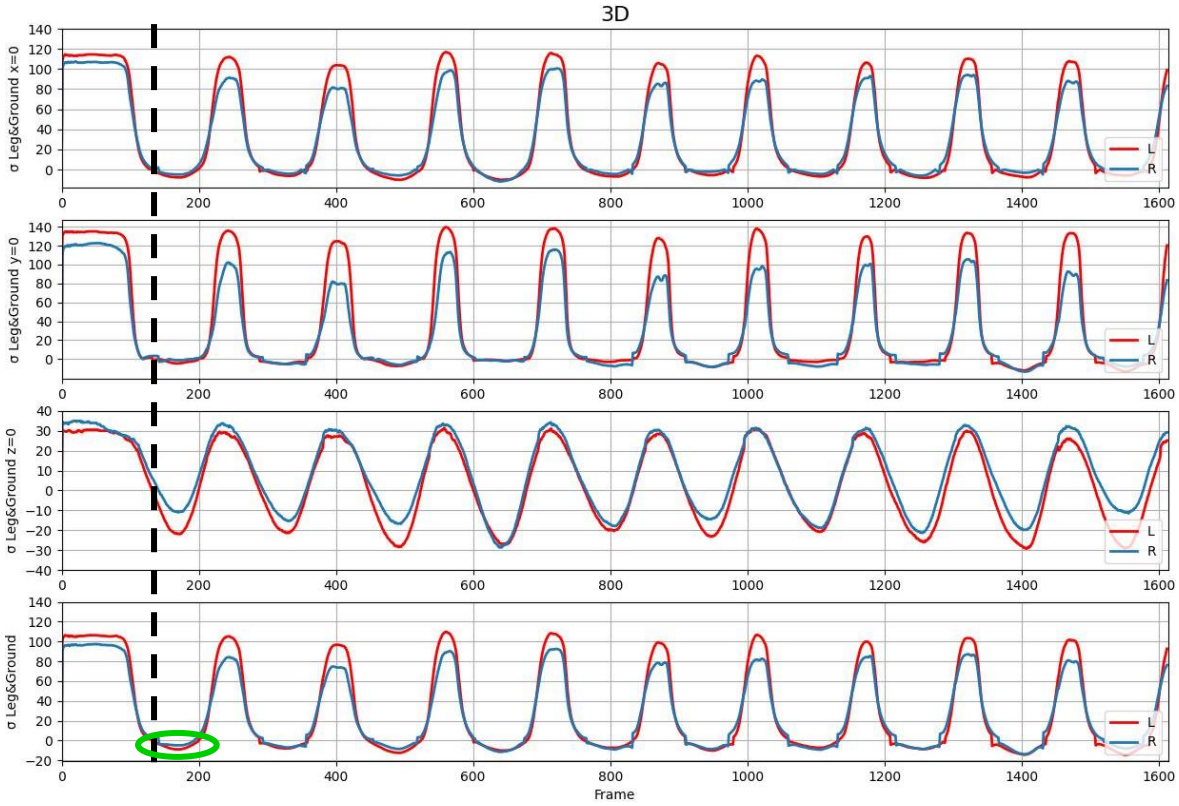

Figure 13 Angle between the thigh and the ground on each plane.

As shown in Figure 14, during the exercise, no matter whether it is on the xy, xz or yz plane, there should not be any observed changes in angle. Therefore, in the three-dimensional assessment of foot lift, as long as the angle exceeding 5 degrees on one plane is considered lifted. For example, in the figure, the x=0, y=0

block has a lifting situation (purple circle), but the z=0 block does not. The green circle is the opposite scenario, and the aforementioned lifting situations must be recorded one by one.

	1	2	3	4	5	6	7	8	9	10
L	✓			✓				✓	✓	
R	✓		✓			✓			✓	✓

Lift up ✓

Table 1 Actual table of foot lift.

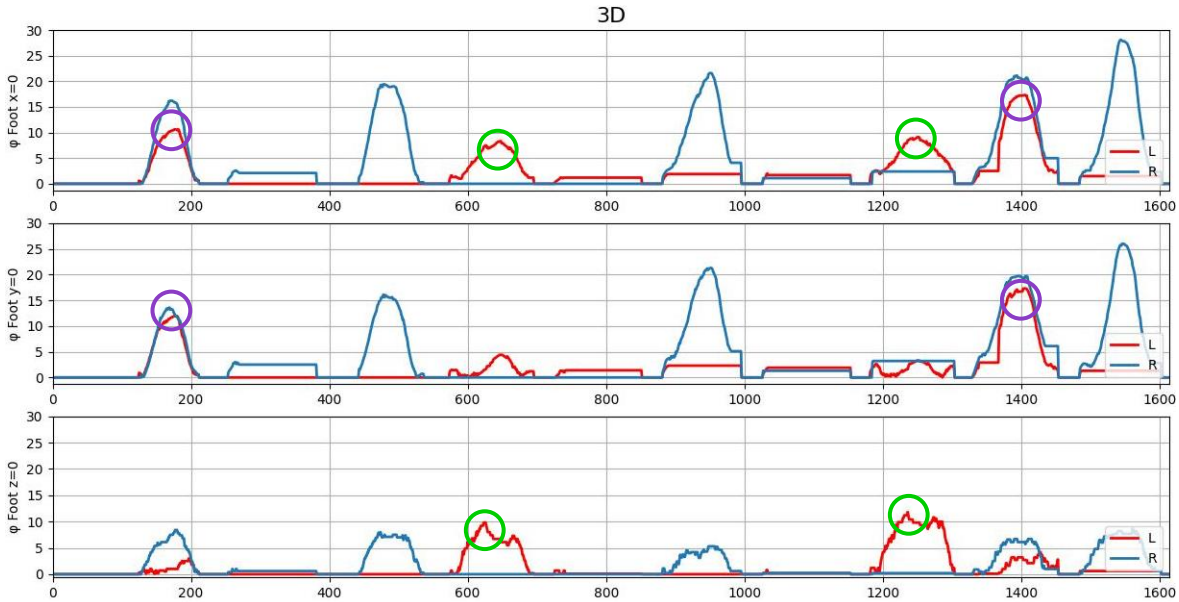

Figure 14 Angle of foot lift on each plane.

The system's average FPS is approximately 25. In the overall system, the Landmarks, which is the program for obtaining human body landmark coordinates, takes up approximately 75% of the execution time. Coordinate smoothing accounts for approximately 6%. Main, which is the main program related to image input/output, angle calculation, and data collection, accounts for approximately 19%.

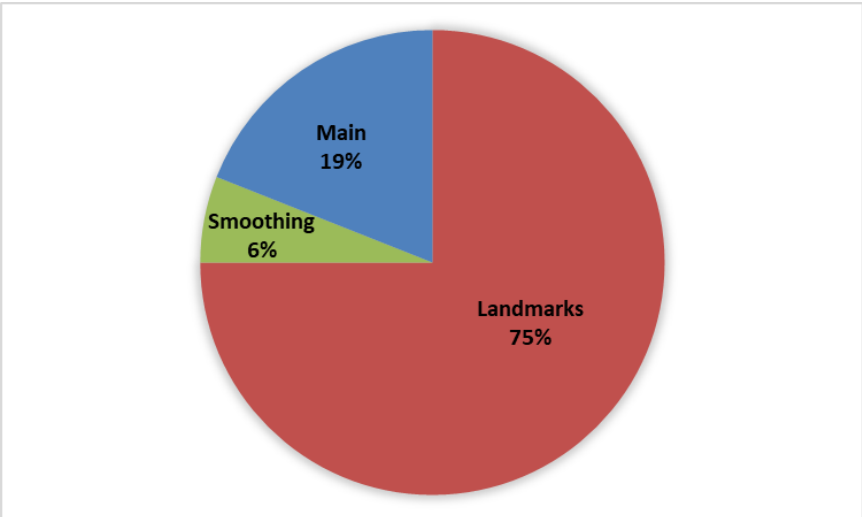

Figure 15 Chart of time consumption ratio.

IV. CONCLUSIONS

This study has implemented the use of images recorded by a single camera, or real-time motion images to assist in squat exercises. MediaPipe was chosen as the core of the overall system because it provides a superior number of human body key point coordinates compared to other posture estimation models, and it also has lower hardware requirements. By fully utilizing the landmark coordinate information, methods were designed to reduce coordinate point errors. Additionally, vector relationships between various parts were derived to expand related functions. Finally, the exercise images and information are displayed on the screen, and relevant exercise data is generated after the image ends for observing the overall exercise situation. Squatting is just one of many types of exercises. As long as the coordinates and vector relationships between various parts are properly utilized, they can be applied to other exercises in a similar manner. Since mobile phones are popular among the public, if this system is implemented as a mobile app that provides text or voice correction prompts when incorrect postures appear, it will be easier to use and promote."

REFERENCES

[1] L. Del Vecchio, "The health and performance benefits of the squat, deadlift, and bench press," *MOJ Yoga & Physical Therapy*, vol. 3, Apr. 2018, doi: 10.15406/mojypt.2018.03.00042.4

[2] Z. Wu, J. Zhang, K. Chen, and C. Fu, "Yoga Posture Recognition and Quantitative Evaluation with Wearable Sensors Based on Two-Stage Classifier and Prior Bayesian Network," *Sensors (Basel, Switzerland)*, vol. 19, 2019.

[3] W. Cao, J. Zhong, G. Cao, and Z. He, "Physiological Function Assessment Based on Kinect V2," *IEEE Access*, vol. 7, pp. 105638-105651, 2019, doi: 10.1109/ACCESS.2019.2932101.

[4] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172-186, 2021, doi: 10.1109/TPAMI.2019.2929257.

[5] C. Lugaresi *et al.*, "MediaPipe: A Framework for Building Perception Pipelines," *arXiv e-prints*, 2019. doi:10.48550/arXiv.1906.08172.

[6] X. Li, M. Zhang, J. Gu, and Z. Zhang, "Fitness Action Counting Based on MediaPipe," in 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Nov. 5-7, 2022, pp. 1-7, doi: 10.1109/CISP-BMEI56279.2022.9980337.

[7] V. S. P. Bhamidipati, I. Saxena, D. Saisanthiya, and M. Retnadhas, "Robust Intelligent Posture Estimation for an AI Gym Trainer using Mediapipe and OpenCV," in *2023 International Conference on Networking and Communications (ICNWC)*, pp. 1-7, April 5-6, 2023, doi: 10.1109/ICNWC57852.2023.10127264.

[8] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking," arXiv e-prints, 2020. doi:10.48550/arXiv.2006.10204.