Advanced Project for AI Convergence

Title: Utilization of image (video) deep learning technology, and advanced body recognition technology research/development

Name	Nguyen Thi Phuong Hang	Course	Master
Student ID	208258		

1. Introduction

The project title: Utilization of Image (video) using deep learning technology, Advanced research/development of body recognition technology

Company name: Genie Soft Co., Ltd. (주식회사 지니소프트)

Project schedule

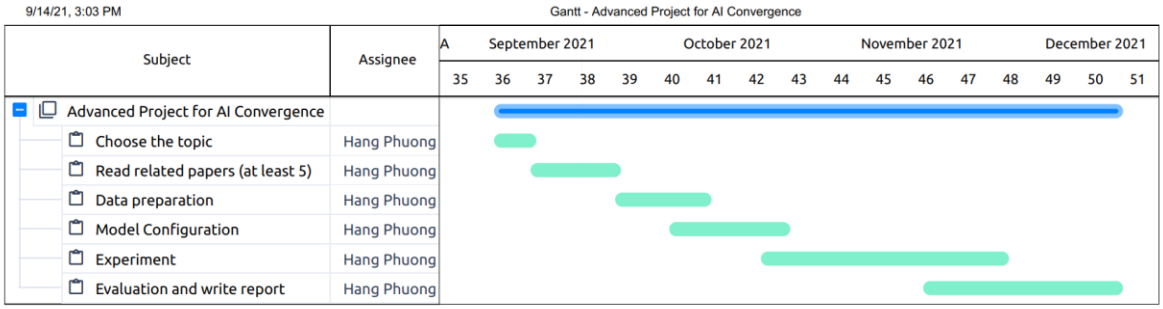


Figure 1: The time plan for project

In this project, I will apply the Mediapipe to extract the keypoints and transfer to 3D visualization and apply the Cosine Similarity to compare 2 poses.

2. Methods

2.1 Medipipe

MediaPipe Pose is a ML solution for high-fidelity body pose tracking, inferring 33 3D landmarks and background segmentation mask on the whole body from RGB video frames utilizing our BlazePose research that also powers the ML Kit Pose Detection API[1]. Here are the 33 landmarks that this model detects:

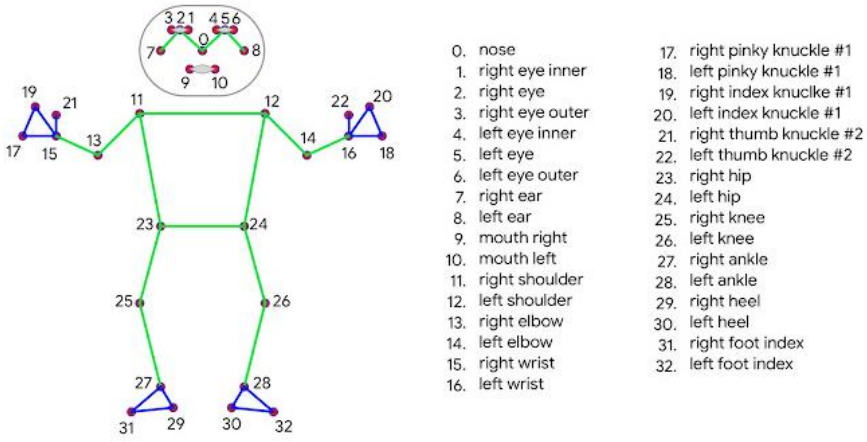


Figure 2: 33 joints landmarks [3]

The solution utilizes a two-step detector-tracker ML pipeline, proven to be effective in our MediaPipe Hands and MediaPipe Face Mesh solutions. Using a detector, the pipeline first locates the person/pose region-of-interest (ROI) within the frame[1]. The tracker subsequently predicts the pose landmarks and segmentation mask within the ROI using the ROI-cropped frame as input. Note that for video use cases the detector is invoked only as needed, i.e., for the very first frame and when the tracker could no longer identify body pose presence in the previous frame. For other frames the pipeline simply derives the ROI from the previous frame's pose landmarks.

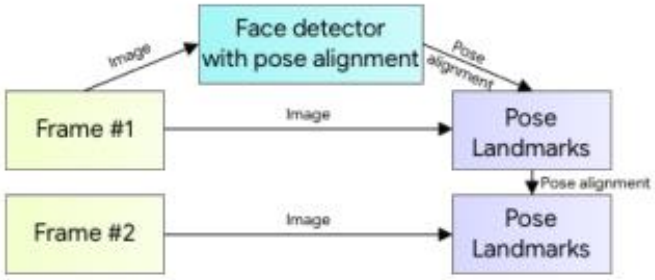


Figure 3: The human pose estimation pipeline [3]

The pose estimation component of the pipeline predicts the location of all 33 person keypoints with three degrees of freedom each (x, y location and visibility) plus the two virtual alignment keypoints described above. Unlike current approaches that employ compute-intensive heatmap prediction, our model uses a regression approach that is supervised by a combined heat map/offset prediction of all keypoints, as shown below. Specifically, during training we first employ a heatmap and offset loss to train the center and left tower of the network. We then remove the

heatmap output and train the regression encoder (right tower), thus, effectively using the heatmap to supervise a lightweight embedding.

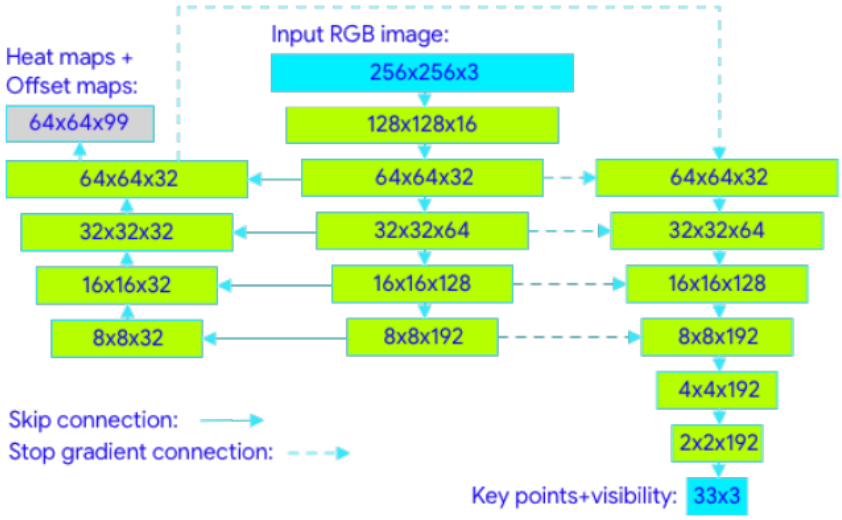


Figure 4: The tracking network architecture [1]

You run the detector in the first frame of the video to localize the person and provide a bounding box around it, after that the tracker takes over and it predicts the landmark points inside that bounding box ROI, the tracker continues to run on any subsequent frames in the video using the previous frame's ROI and only calls the detection model again when it fails to track the person with high confidence.

Keypoint Z-value estimate is provided using synthetic data, obtained via the GHUM[2] model (articulated 3D human shape model) fitted to 2D point projections[4].

Figure 5: Sample GHUM fitting for an input image. From left to right: original image, 3D GHUM reconstruction (different viewpoint) and blended result projected on top of the original image [2]

Their model works best if the person is standing 2-4 meters away from the camera and one major limitation of their model is that this approach only works for single-person pose detection, it's not applicable for multi-person detection.

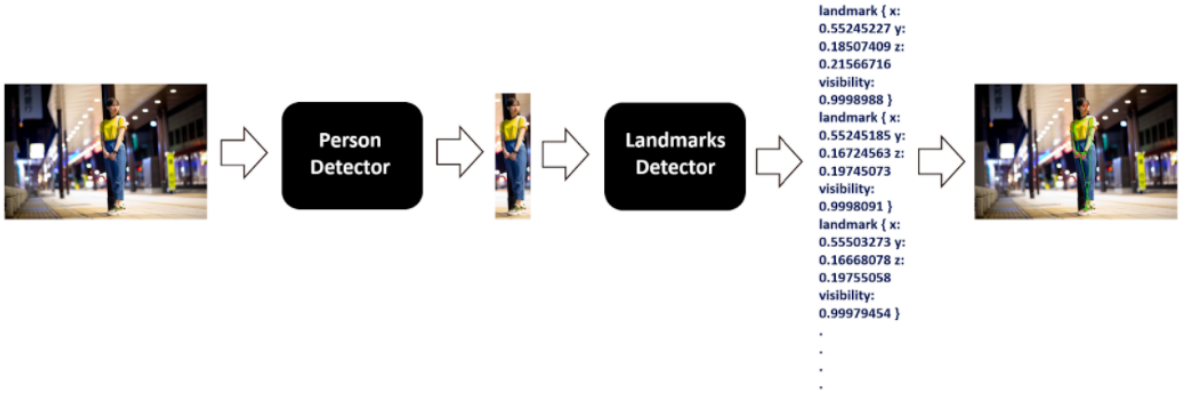


Figure 6: The model works

2.2 Results

To evaluate the quality of our models against other well-performing publicly available solutions, we use three different validation datasets, representing different verticals: Yoga, Dance and HIIT. Each image contains only a single person located 2-4 meters from the camera. To be consistent with other solutions, we perform evaluation only for 17 keypoints from COCO topology.

Method	Yoga mAP	Yoga PCK@0.2	Dance mAP	Dance PCK@0.2	HIIT mAP	HIIT PCK@0.2
BlazePose GHUM Heavy	68.1	**96.4**	73.0	**97.2**	74.0	**97.5**
BlazePose GHUM Full	62.6	**95.5**	67.4	**96.3**	68.0	**95.7**
BlazePose GHUM Lite	45.0	**90.2**	53.6	**92.5**	53.8	**93.5**
AlphaPose ResNet50	63.4	**96.0**	57.8	**95.5**	63.4	**96.0**
Apple Vision	32.8	**82.7**	36.4	**91.4**	44.5	**88.6**

Table 1: Result for methods on datasets on COCO topology (17 keypoints) [1]

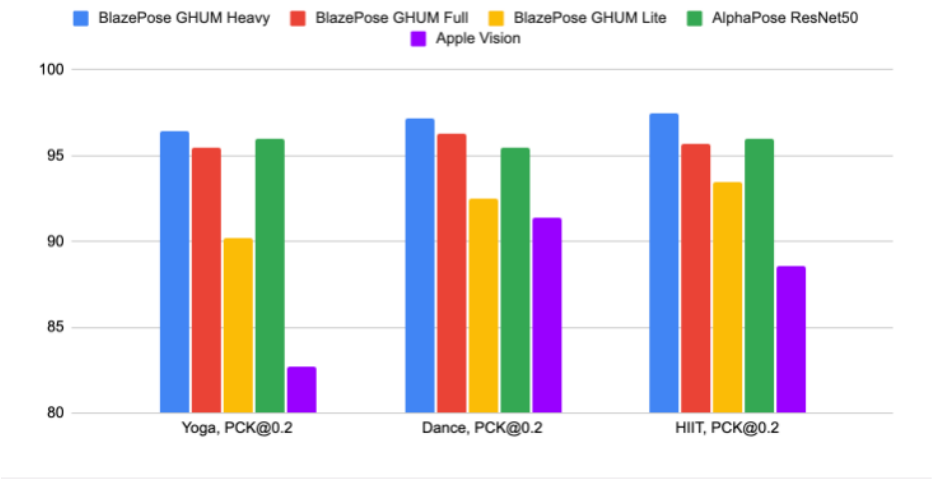


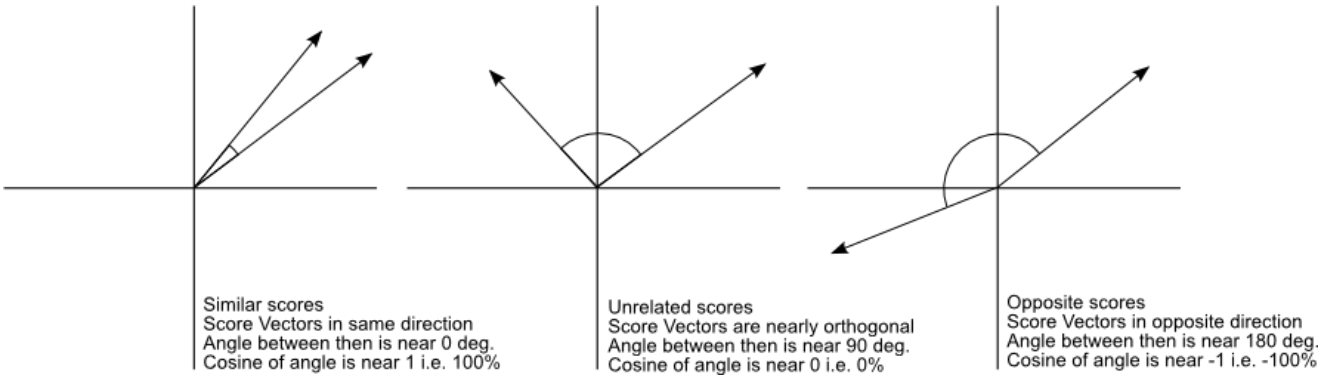
Table 2: Quality evaluation in [PCK@0.2 [1]]

2.3 Cosine Similarity

Cosine similarity of two vectors are defined as follows:

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}$$

The keypoints are converted into a vector and plotted in high dimensional space. This vector plotting is compared to another vector plot from our benchmark image.



To calculate the Cosine distance:

$$D(Fxyz, Gxyz) = \sqrt{2 * (1 - cosineSimilarity(Fxyz, Gxyz))}$$

If D(Fxyz, Gxyz) is larger, the similarity is smaller. It means 2 poses are different. And to calculate the distance of angle:

$$D(Fi,Gj) = \frac{Fi - Gj}{(Fi + Gj)/2}$$

3. System configuration

In this project, I listed the angle which need to calculate in whole body. Base on the computational geometry, I can find the values of these angles.

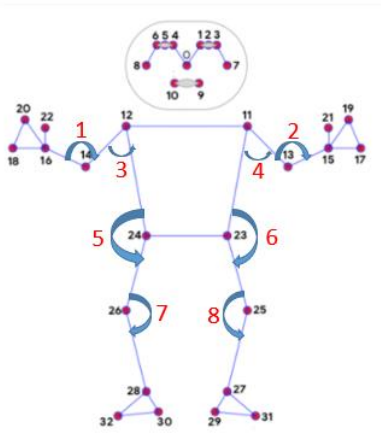


Figure 7: The angle list on human

The figure 7 presents the system configuration pipeline of my study. The first step, I use the Mediapipe pose to extract the 33 key points, then I calculate the angles list as figure 6. The next step, I transfer the 3D visualization and I use the dynamic time warping to evaluate the motion of human and compare the motion of non-professional people and experts in field for example, yoga, ... and so on.

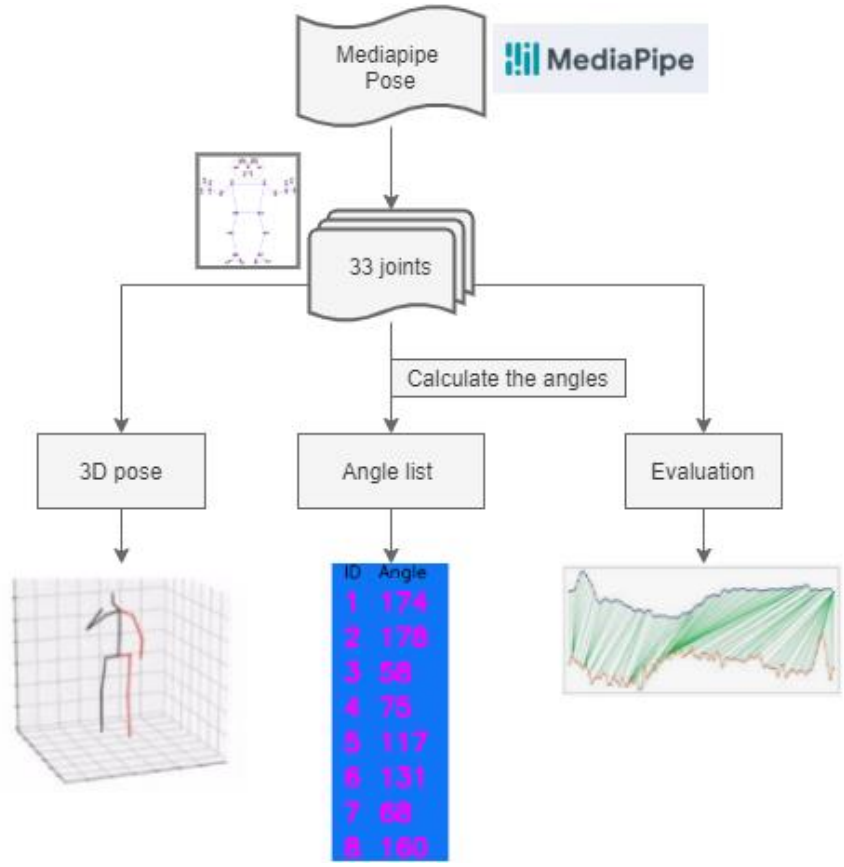
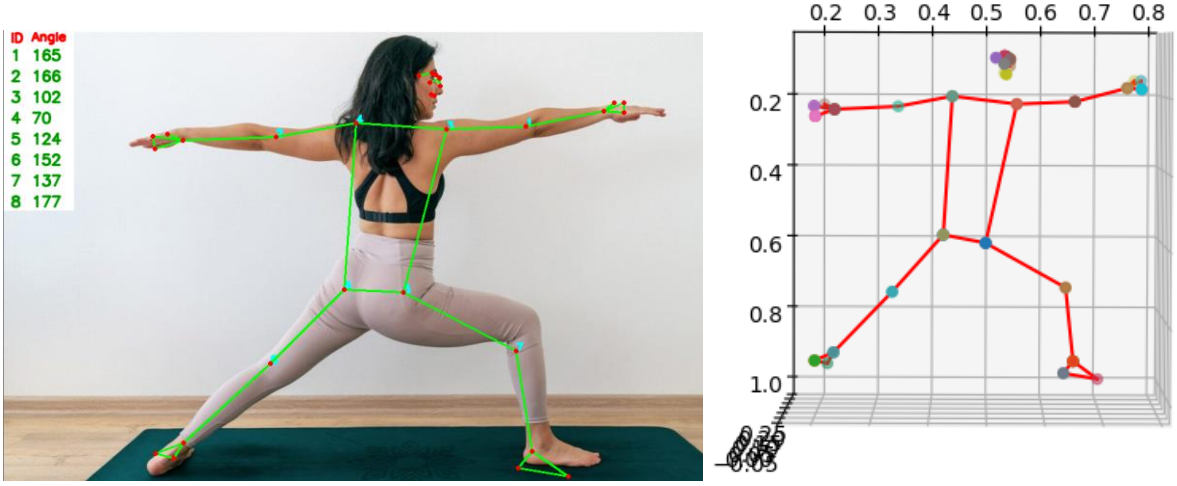
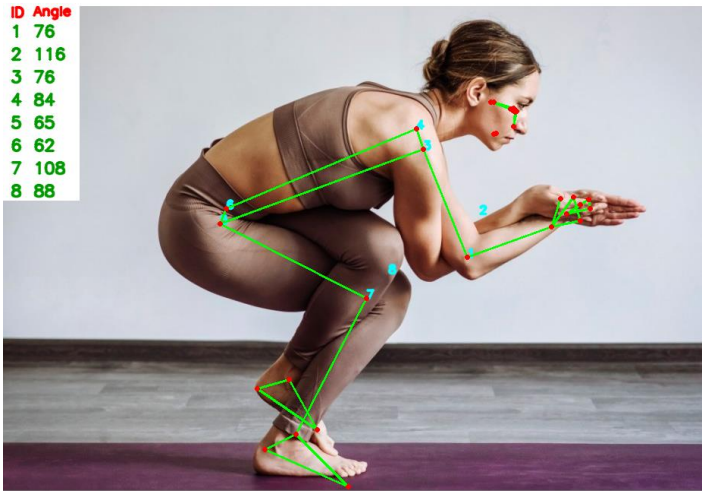


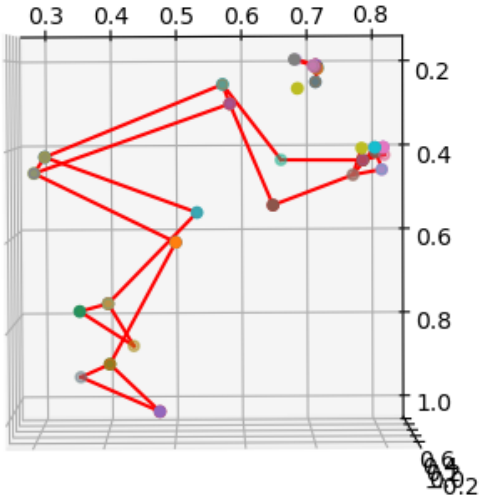
Figure 8: The system configuration

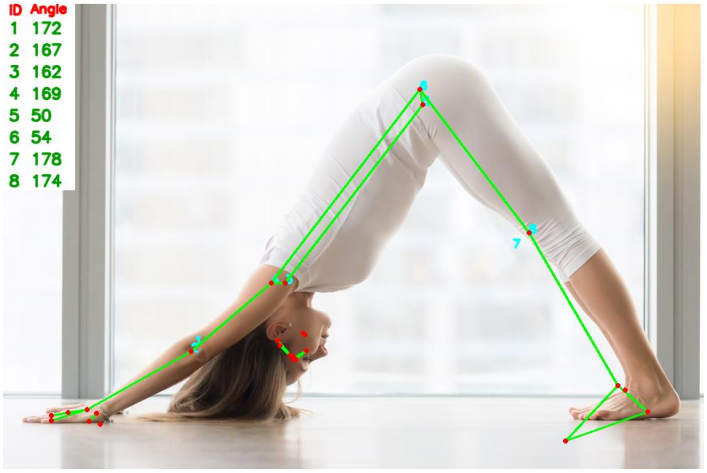
4. Results

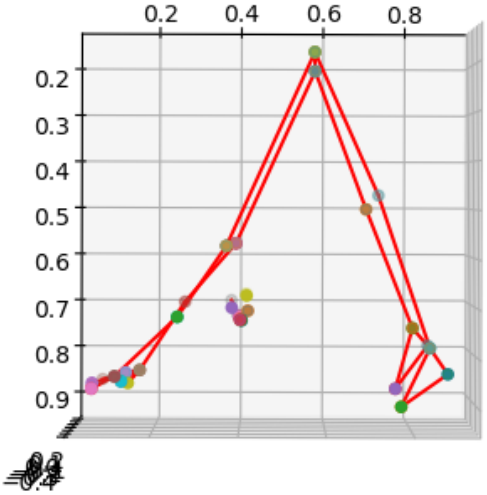
4.1. Result in images to extract the keypoints and calculate the angles

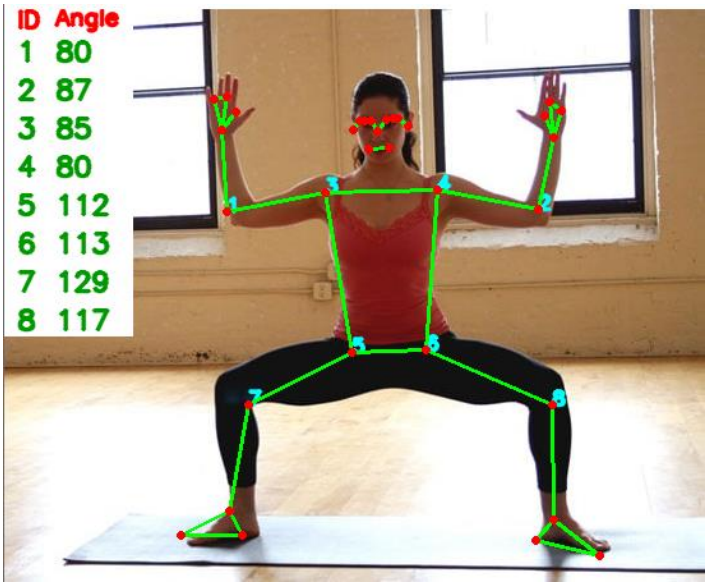


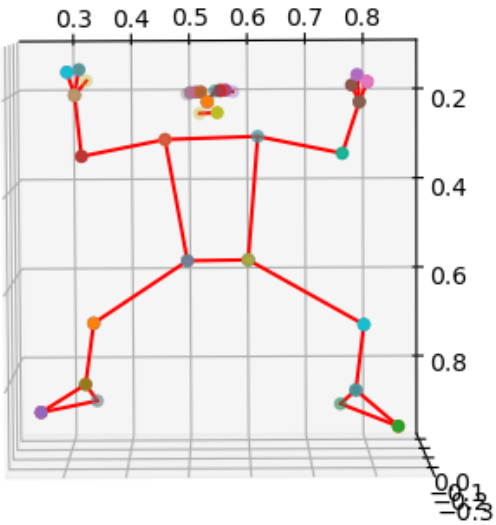

ID Angle
1 76
2 116
3 76
4 84
5 65
6 62
7 108
8 88


0.3
0.4
0.5
0.6
0.7
0.8
0.2
0.4
0.6
0.8
1.0


ID Angle
1 172
2 167
3 162
4 169
5 50
6 54
7 178
8 174


0.2
0.4
0.6
0.8
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9


ID Angle
1 80
2 87
3 85
4 80
5 112
6 113
7 129
8 117


0.3
0.4
0.5
0.6
0.7
0.8
0.2
0.4
0.6
0.8

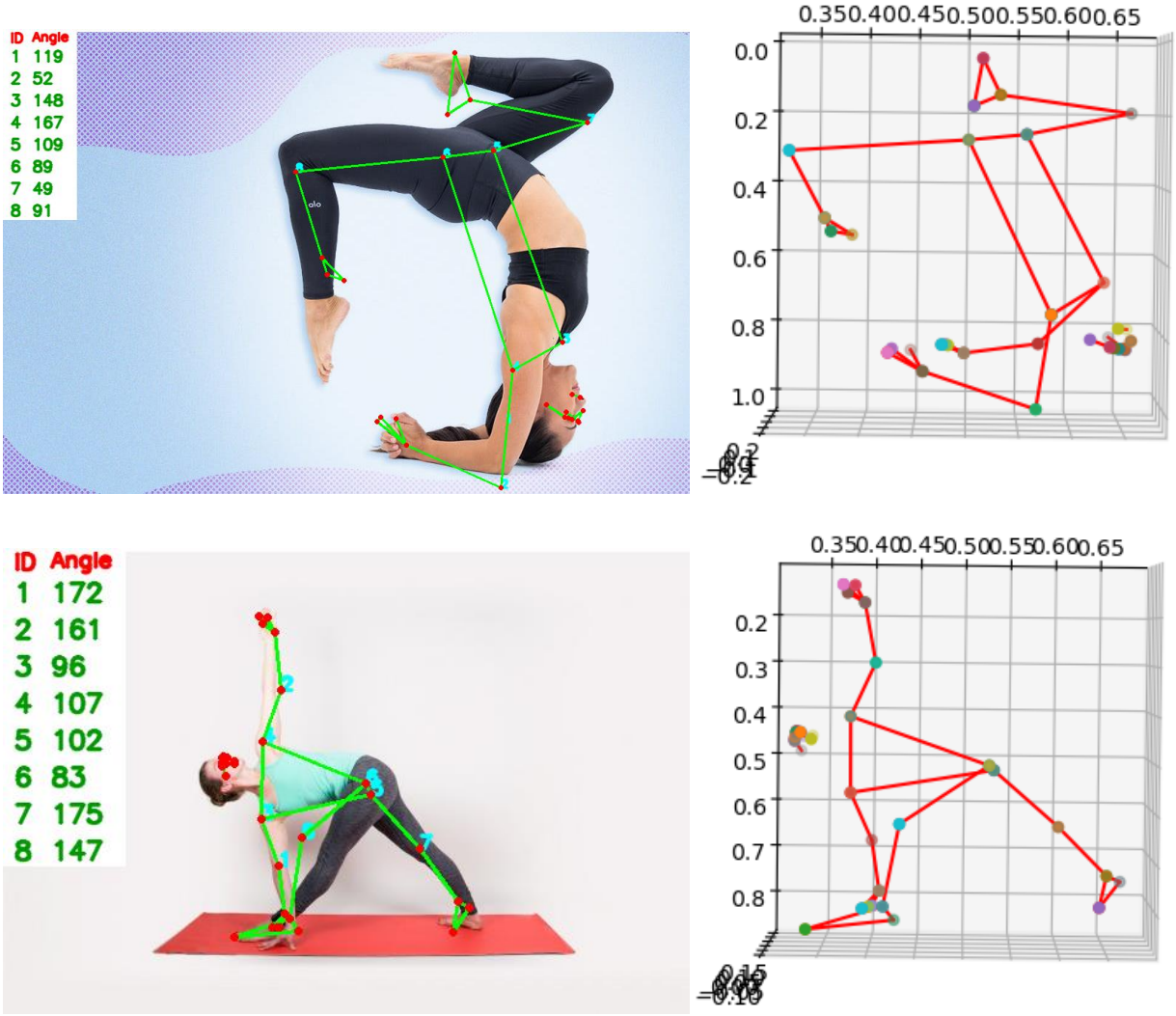
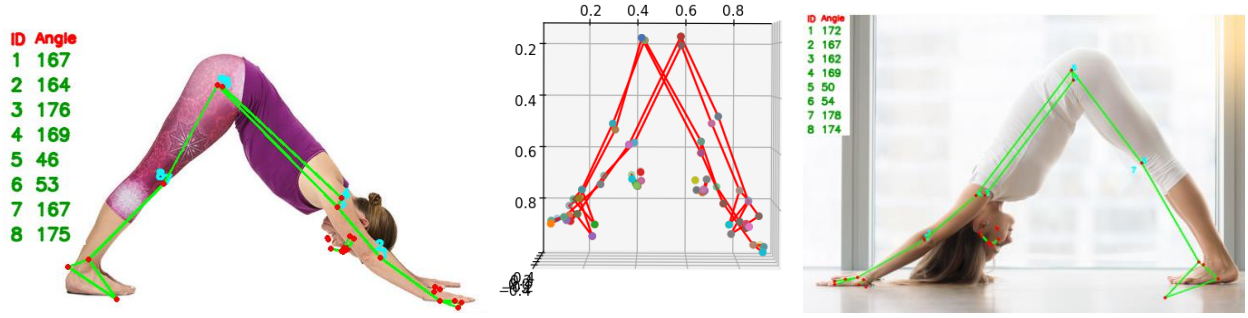


Figure 9: Results in yoga pose images and their 3D visualization

4.2 Compare the similarity of 2 poses in images.

4.2.1 The comparison of 2 opposite direction poses



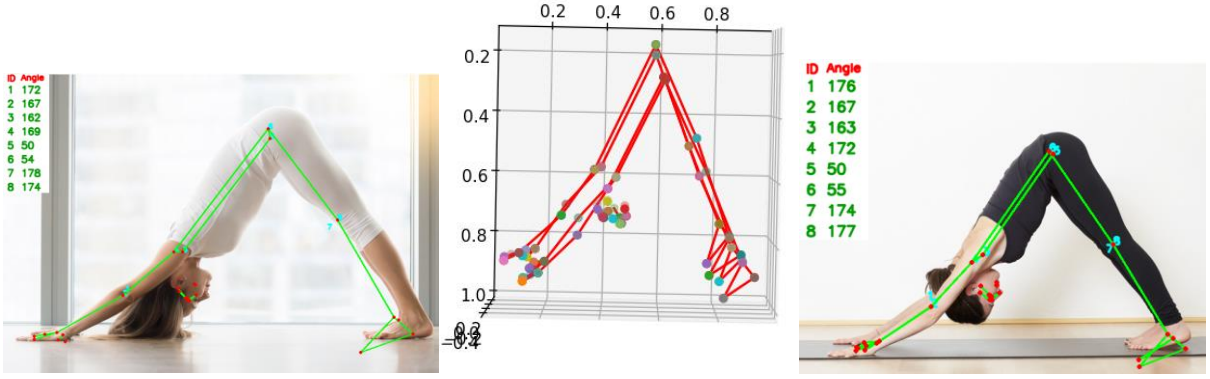
Distance coordinates	Score (%)	Distance angle	Score (%)
0.58	42	0.03	97

Figure 10: Results in 2 opposite direction poses

The score of distance angle is high and the score of distance coordinates is low.

4.2.2 The comparison of 2 completely similar poses

The scores of both distance coordinates and distance angle are high.

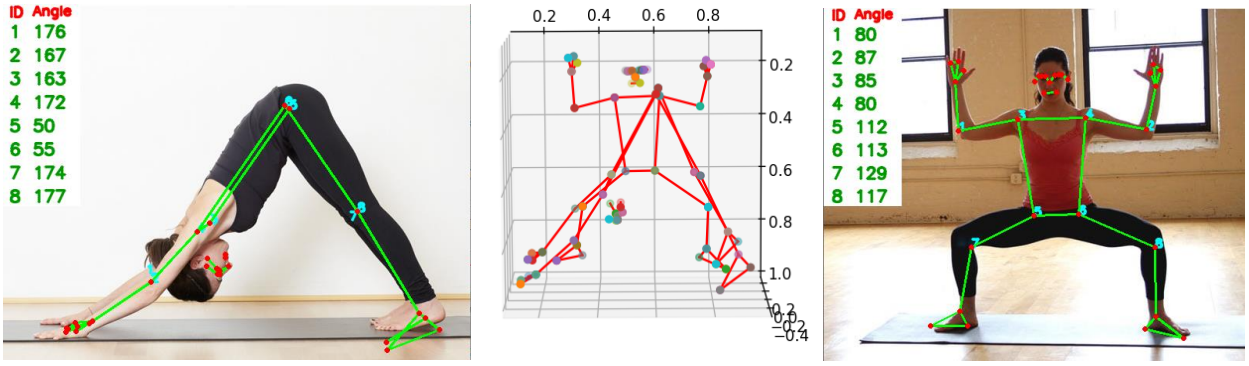


Distance coordinates	Score (%)	Distance angle	Score (%)
0.0	100	0.01	99

Figure 11: Results in 2 completely similar poses

4.2.3 The comparison of 2 completely different poses

The scores of both distance coordinates and distance angle are low.



Distance coordinates	Score (%)	Distance angle	Score (%)
0.72	28	0.61	39

Figure 12: Results in 2 completely different poses

4.2 Result in video

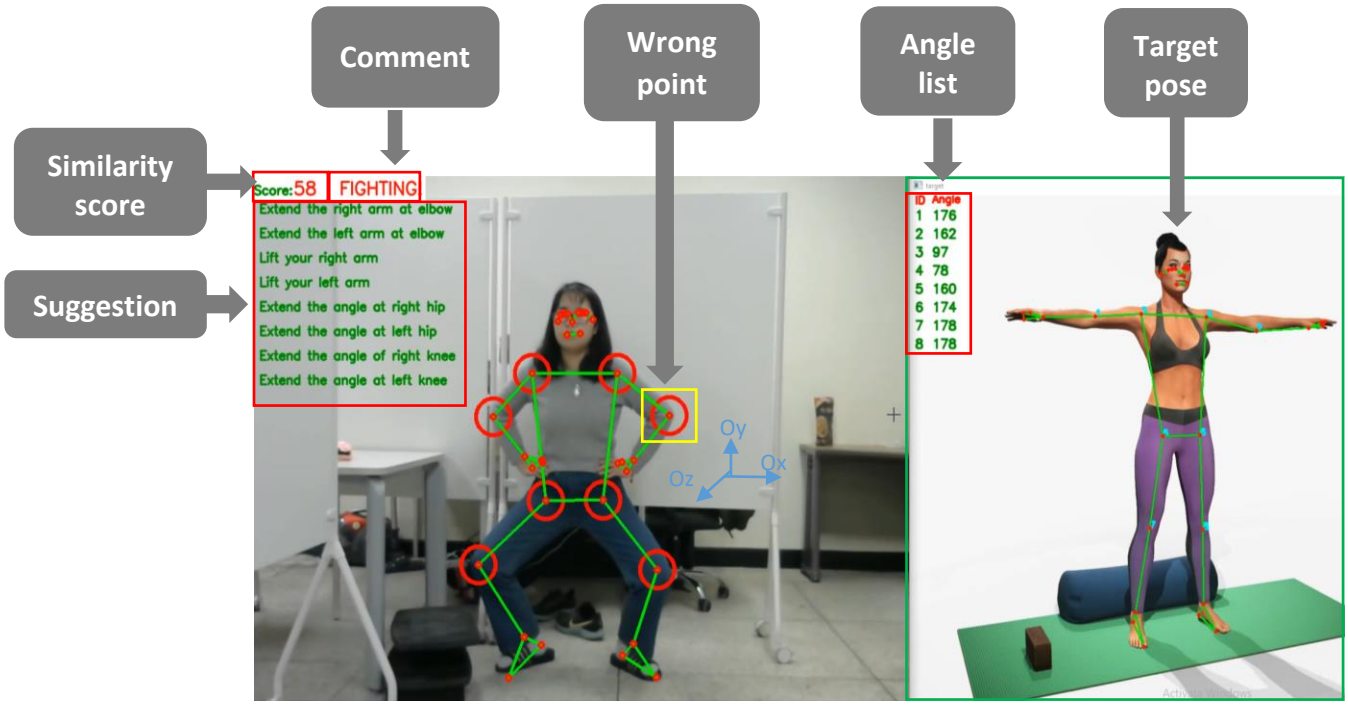


Figure 13: Screen of real-time video in wrong pose

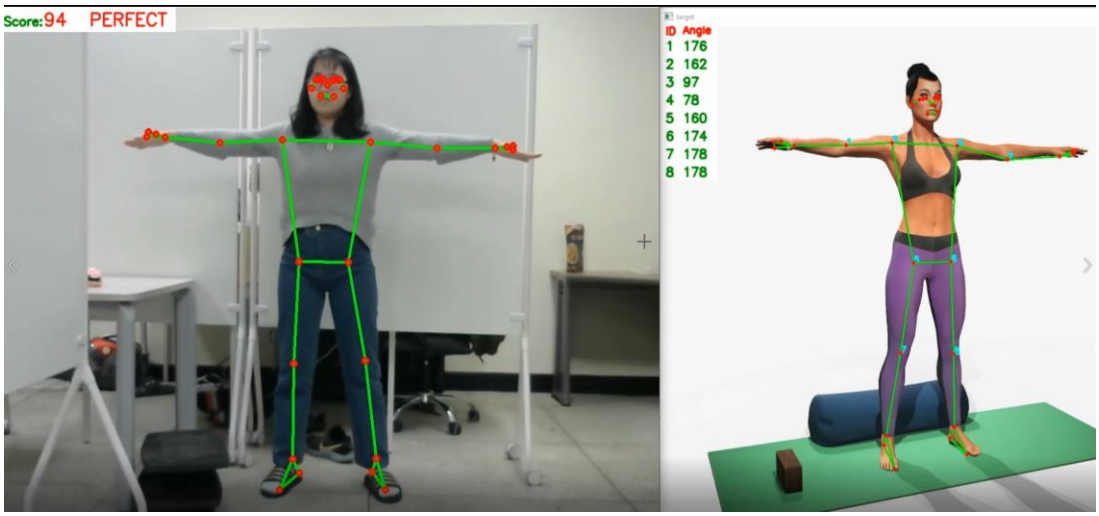


Figure 14: Screen of real-time video in correct pose

5. Discussion

On figure 9, we can see the results of pose image of human objects, the keypoints and the eight main angles were calculated and displayed on these images. And the corresponding 3D visualization were also present on the results. Beside the correct keypoints, we can see the incorrect keypoints from some poses and this is our limitation. When I applied in video, the video with the fast speed is better than the slow video and if any segments of body was hidden, the joints recognition is limited. On figure 10, figure 11 and figure 12, I performed to compare 2 poses images and I realized that the coordinates comparison is not enough to compare 2 poses so I used both the angle comparison and coordinates comparison.

+ If 2 poses are completely different, the scores of both coordinates comparison and angle comparison are low

+ If 2 poses are completely similar, the scores of both coordinates comparison and angle comparison are high

+ If 2 poses are completely similar and has opposite direction, the score of coordinates comparison is low and the score of angle comparison is high

In real-time video, the one side is the target pose image and one side is the real-time video. User should follow the target pose and complete the pose. If the pose is correct,

the “PERFECT” will display with high score. And if the pose is not correct, the “FIGHTING” will display with lower score and the suggestions to do a correct pose.

6. Conclusion

This project develops a fitness trainer to promote daily home training based on the real-time human pose similarity evaluation while comparing with the pose of experts or target pose. 3-dimesional joints were extracted by Mediapipe and main angle list was calculated to make an effective decision about the pose of trainee. As the results, the effectiveness of the proposed home training systen was confirmed by yoga exercises. The score of similarity on this system was calculated on both coordinates comparison and angle comparison. The limitation of this system is the difficult pose in yoga exercise and all points should be visible in image and video. Because of the short time, the front-end of this system is not perfect and hope it will be improved in the future. This application is suitable for home training with only low-cost single camera like yoga, fitness,…

7. References

[1] Mediapipe pose: https://google.github.io/mediapipe/solutions/pose

[2] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, Cristian Sminchisescu , GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models

[3] Valentin Bazarevsky and Ivan Grishchenko, On-device, Real-time Body Pose Tracking with MediaPipe BlazePose, 2020

[4] Model card: Mediapipe BlazePose GHUM 3D, 2019: https://drive.google.com/file/d/10WlcTvrQnR_R2TdTmKw0nkyRLqrwNkWU/preview