Automatic Recognition of Human Workout / Exercises Using MediaPipe and Long-Short Term Memory Networks

[1,*]*You-Lin Yang* (楊侑霖), *and* [1,]*Yuan-Hsiang Chang* (張元翔)

[1] Department of Information and Computer Engineering,
Chung Yuan Christian University, Taoyuan, Taiwan.
[*]E-mail: qazzy5566@gmail.com

ABSTRACT

Daily workout or exercise is important to ensure good health in human's life. Because of advances of artificial intelligence (AI) techniques, the objective of this study is to develop a system for automatic recognition of human work/exercises using MediaPipe and Long-Short Term Memory (LSTM) networks. A video dataset was previously collected and used to train and test our system model. Our experimental study demonstrated that our system could achieve the accuracy, precision, and recall of over 90%, which indicates relatively good performance in the dataset. In summary, our system model could yield potential solutions to detect and classify human's workout/exercise behavior, thus leading to AI coach, multimedia, human-computer interaction, or other applications.

Keywords: *Action Recognition, Deep learning, Human Pose Recognition, LSTM Network.*

1. INTRODUCTION

Daily workout or exercise is very important for good health in human life. During human body exercise, it is often important to pay attention to correct posture frequently. Once the posture is not correct, it is easy to generate abnormal muscle soreness, which may lead to serious injury.

Therefore, maintaining good posture cannot be ignored. One way is to hire a personal trainer or coach for one-on-one instruction, but the cost is usually expensive. In the era of artificial intelligence (AI), an AI trainer or coach may provide potential and cost-effective solution to the problem.

In recent years, human posture recognition has been an active and ongoing research topic in the field of computer vision. This technology can be applied in various applications, such as medical treatment, exercises, sports, traffic safety, or Human Computer Interaction (HCI).

In medicine, human posture recognition can be used for gait analysis [1], hand tremor analysis for Parkinson's patients [2], Neurological tests [3], etc. It can also be applied in traffic safety [4], sign language recognition [5] and human-computer interaction (HCI) [6].

Luvizon, *et al.* [7] proposed a multitask framework using deep learning method. The model estimates human pose estimation from images, and then recognizes action in video sequences.

Studies in [8,9,10,11] provide different approaches for physical exercise recognition.

Shuo, *et al.* [8] presented a deep squat detection by using MediaPipe and YOLOv5. The results of their system can achieve an accuracy rate of over 96%.

Weeriya, *et al.* [9] also used MediaPipe for human body extraction. They built a system to classify three exercise poses and compared the performance of four different machine learning models.

Utkarsh and Shikha [10] developed a system that identifies different yoga poses. The body pose extraction method is also MediaPipe. They then train and test the data by using classification-based machine learning algorithms. The system achieved a good accuracy score of 94%.

Islam [11] developed a yoga poses recognition system but in a different method. The method is to use Microsoft Kinect to detect joint point of human body, and calculate the joint angles to check the correctness of different yoga poses. The system successfully works in real time.

In this study, our goal is to build an automatic recognition system that can automatically detect and classify human's workout or exercise behavior, with the potentials to help human rectify incorrect postures, thereby improving the efficiency of exercise and reducing the risk of injury.

In the next section, we will go through every detail of how we implement the data preprocessing, model training and performance measuring to make sure that the model we trained is well performed.

2. METHODS

Fig. 1 shows the overall block diagram of our system. A video dataset with 22 classes was collected and organized for the training and testing of our system. The MediaPipe pose detection from Google research is used for the human body detection. After acquiring the human pose features from a set of frames, the LSTM network is used as the classification model. The process was divided into training and testing stage.

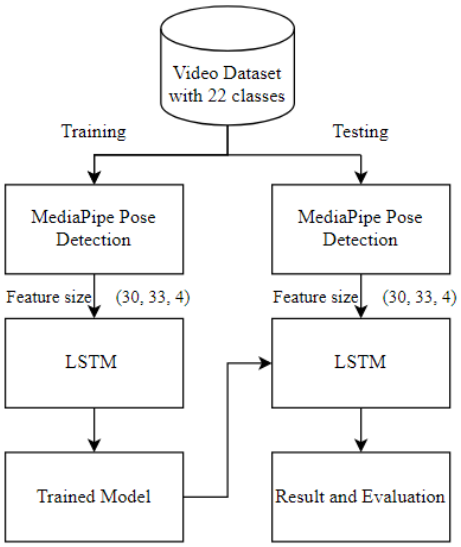

Fig. 1. Overall block diagram of our system for the automatic recognition of physical exercise behavior.

2.1. MediaPipe Pose Detection

MediaPipe is an open source machine learning framework developed by Google research. It provides many different solutions such as detection, classification for image, audio, text and language. MediaPipe is also compatible with C/C++, Android, IOS, and Python programming environment, which allows the developers to build their applications fast and easily on different systems.

In this research, we used MediaPipe's pose landmark detector as the first step. This detector identifies 33 different body landmarks from an image or video frame, with each landmark having four values: (x, y, z) coordinates and visibility. All of the coordinates were normalized by the image width and height.

Fig. 2. shows the 33 body landmarks on the human body detected by MediaPipe pose, all 33 landmarks and 4 coordinates will be used to train the model in this study.

2.2. LSTM Network

Long Short-Term Memory (LSTM) networks were introduced by Hochreiter and Schmidhuber in 1997 [12]. The LSTM network is essentially a gradient based recurrent neural network (RNN), and allows us to capture long-term information by controlling the gate signals. Through these gates, we are able to decide which information should be forgotten, recorded, updated and output for further prediction.

Due to the design structure, LSTM is suitable for processing and predicting time series and sequential data, such as video classification or text analysis, and locally in space and time, which allow us to train model.

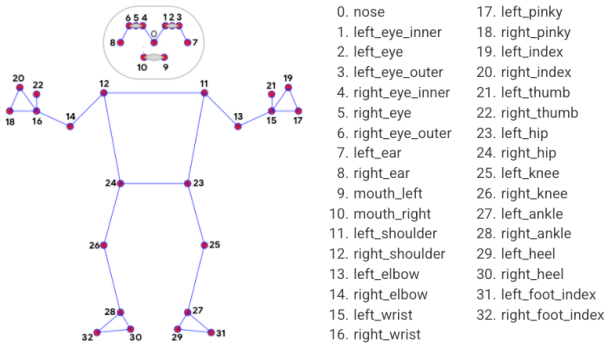

Fig. 2. The 33 landmarks extracted by MediaPipe pose detection.

2.3. System Architecture

Fig. 3. Shows the system architecture for the automatic recognition of physical exercise behavior using a workout/exercise dataset.

The first step is video preprocessing, all the videos from the dataset were processed by MediaPipe pose detection on a frame by frame basis. If MediaPipe does not detect human body, we then ignore that frame and continue on the next frame. After pose detection, it will generate 33 landmarks on human body for each frame that is successfully detected. Each landmark contains 4 values with (x, y, z) coordinates and visibility. Finally, each video will generate an array of landmarks, the length of which varies according to the detection conditions and the length of the video. To create the dataset instances, these arrays will be divided into groups of 30 frames. At the same time, in order to increase the amount of data, the landmarks are cut into landmark 0 to landmark 29, landmark 15 to landmark 44, and so on, until there is no way to cut out the complete 30 landmarks. Overall, the total data set for training is composed of 8,514 landmark arrays of size (30, 33, 4) with 22 different classes, the dataset was split into 80% for training and 20% for testing.

After the data preprocessing is completed, the second step is the training stage. The network architecture used in this study is composed of LSTM

network and linear fully connected layer. The input size of the LSTM block is 132 which corresponds to the size of the landmark features, and the output size of 128, and the number of recurrent layers is 1. The output of LSTM block is fed into linear layer, and the output of linear layer with 22 units will be the classification result of our system.

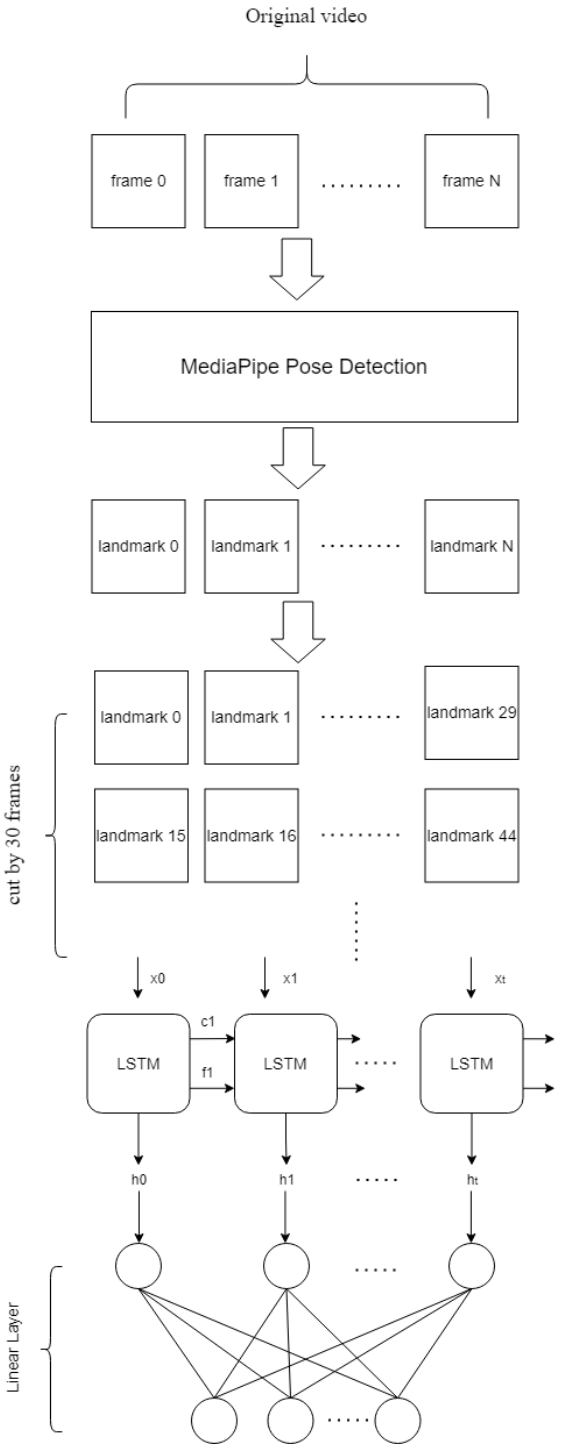

Fig. 3. System architecture to detect human body from video and recognize the exercise posture.

2.4. Model Training

After the model was built, the model was trained for 30 epochs and the batch size is 64. The loss function used cross entropy loss and the optimizer used Adam optimizer, and the learning rate start with 0.002, and was divided by 2 on every 10 epochs. All the training was performed on a laptop and the total training time was approximately 30 minutes.

3. RESULTS

In this section, the experimental results are presented.

3.1. Video Dataset

A workout / exercises video dataset was previously acquired from the Internet [13]. The video dataset consists of 22 different classes of workout/exercises behaviors. Fig. 4. shows the 22 classes as defined in the dataset. Fig. 5. shows some of the video frames in the dataset.

Fig. 4. The 22 classes defined in the workout/exercises video dataset.

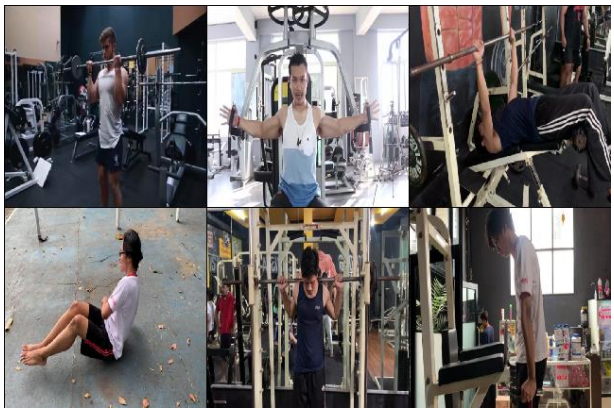

Fig. 5. Sample video frames in the workout/exercises videos dataset.

3.2. Model Training

In this study, the video dataset was randomly divided into 80% training and 20% testing. Figs. 6 and 7. show the accuracy and loss after training for 30 epochs. As shown in the figures, both the training and testing sets achieved over 90% accuracy. Although there were no dropout layers in our system model, overfitting didn't occur during training and testing, thus achieving good classification results.

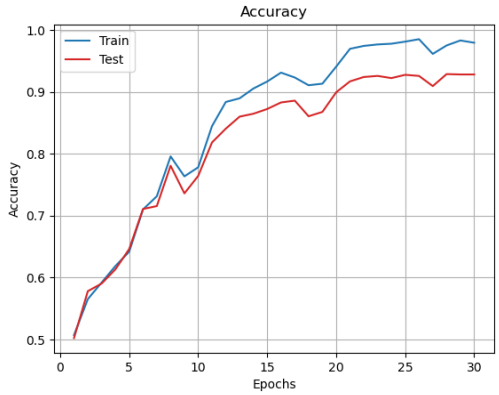

Fig. 6. The accuracy per epoch of the LSTM model.

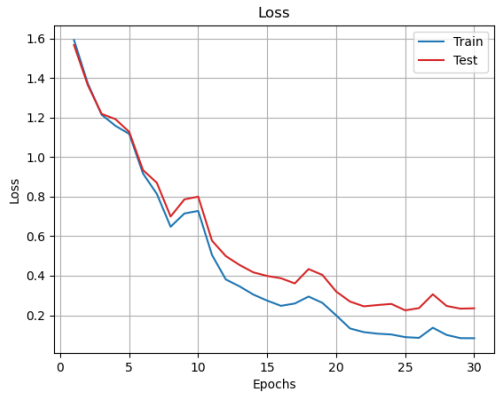

Fig. 7. The loss per epoch of the LSTM model.

Fig. 8. shows the confusion matrix that allows us to observe the model performance thoroughly. From the confusion matrix, it can be observed that some classes easily affect each other's prediction accuracy due to their high similarity in actions, but in the professional world of fitness, they are different movements. There are also some classes that are easier to achieve with higher accuracy, because of their unique movements.

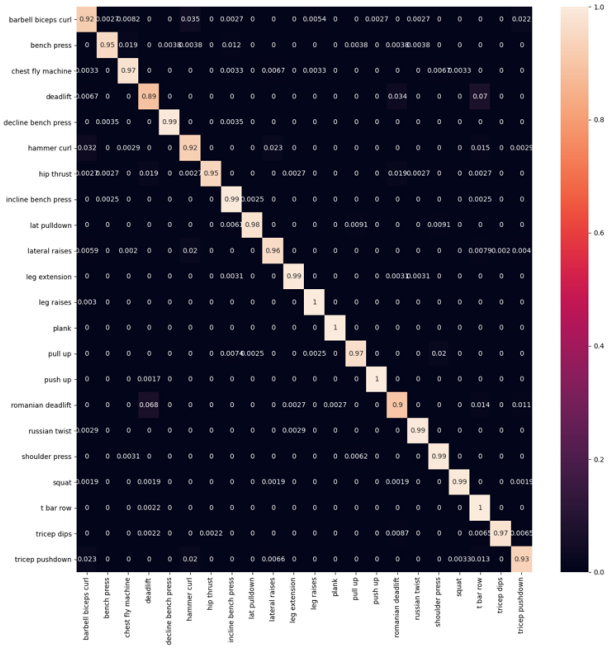

Fig. 8. The confusion matrix in the testing video dataset.

3.3. Performance Metrics

Table 1. shows the Performance Metrics in this study, that allow us to evaluate the performance of our system model more clearly. A model with high precision will be more rigorous in prediction, and a model with high recall can easily find all the correct target, F1-score is the harmonic mean of the two, which is a relatively balance indicator to see the overall performance of this model.

Table 1. Performance Metrics.

	Train Set	Test Set
Accuracy	97.95%	92.83%
Precision	97.81%	92.29%
Recall	97.73%	92.18%
F1-Score	97.77%	92.24%

As shown in the table, the measurement, including accuracy, precision and recall, all exceeded 90%, indicating good performance.

3.4. Video Testing

Fig. 9. shows the implementation of the classification model in the video dataset. The input is a video, and our system model will recognize the actions of the people in the video at regular intervals. Finally, the classification results (i.e., the 22 classes) were shown on upper-left corner of the video frames. The results show that MediaPipe pose detection has high reliability detecting human body key point for classification model to predict and classify. In addition, the LSTM model could also successfully identify the temporal sequences of motions and classify different exercise movements.

In the video test, the computation time for the first action recognition was 0.3 seconds, and the average computation time after that was 0.004 seconds.

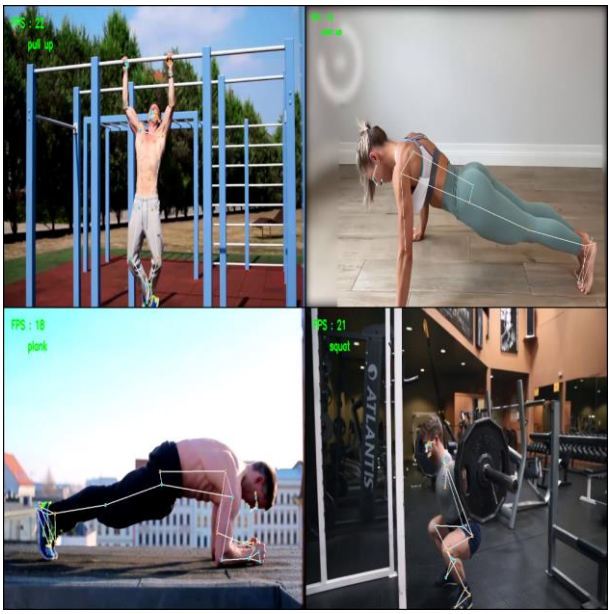

Fig. 9. Model testing using the workout / exercises video dataset.

4. DISCUSSION AND CONCLUSION

In this study, the MediaPipe pose detection could accurately detect key points of the human body, and the detected landmarks could be used to train the model and has achieved good performance.

However, MediaPipe does not support multi-person detection. Therefore, if MediaPipe is used for personal exercise monitoring at home, it may be sufficient, but may not be suitable for use in public places or group activities. To implement multi-person detection with MediaPipe for interactive use, additional tools are required.

Although MediaPipe has a good performance on detection, some errors are still inevitable. That is, the detection is susceptible to noise, if there are obstacles blocking the human body, or other people walking next to the user, these will affect the model predicting by getting the wrong landmark. In practical applications it is necessary to pay attention to the human exerciser and the surrounding environment.

By using the extracted human landmark data through MediaPipe, the LSTM model could effectively recognize different posture movements, and there is no need for complex model. As a result, the training time is also reduced.

Finally, when using video testing, we found that the FPS is unstable or even lagging at some time. If it is to be developed into a real-time system in the future, more optimization processing will be required.

This study combines MediaPipe and LSTM Model for automatic recognition of physical exercise, and has a very good accuracy, also it does not cost too much time in the process of training model.

Moreover, this method could also be implemented in different applications, such as multimedia, transportation, and real-time interactive games, etc., which will be a great help to developers for future implementation.

As for the future direction of this research, more datasets can be added to improve the functionality of this model, since there are still many different kinds of exercises that are not included in this research.

In order to have more complete practicality and stability, we can reduce the complexity of the model so that the model can be applied to real-time systems, or use the faster model such as GRU for implementation.
For the system that will be used in in public places or group activities, we can also use the YOLO to achieve multi-person detection.

5. REFERENCES

[1] A. Gupta, *et al.*, "Knee Flexion/Extension Angle Measurement for Gait Analysis Using Machine Learning Solution "MediaPipe Pose" and Its Comparison with Kinovea®," *IOP Conference Series*: *Materials Science and Engineering***,** Vol. 1279, No. 1, 2023.

[2] G. Güney, *et al.*, "Video-based hand movement analysis of parkinson patients before and after medication using high-frame-rate videos and MediaPipe," *Sensors*, Vol. 22, No. 7992, 2022.

[3] T. Rumambi and M. Hermita, "Motion Detection Application to Measure Straight Leg Raise ROM Using MediaPipe Pose," *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS).* IEEE, pp. 1-5, 2022.

[4] A. J. S. Ong, *et al.*, "LSTM-based Traffic Gesture Recognition using MediaPipe Pose," *TENCON 2022 - 2022 IEEE Region 10 Conference (TENCON), Hong Kong*, pp. 1-5, 2022.

[5] A. Halder and A. Tayade, "Real-time vernacular sign language recognition using MediaPipe and machine learning," *International Journal of Research Publication and Reviews*, Vol. 2, No. 5, pp. 9-17, 2021.

[6] S. Agrawal, A. Chakraborty and M. Rajalakshmi, “Real-Time Hand Gesture Recognition System Using MediaPipe and LSTM,” *International Journal of Research Publication and Reviews*, Vol 3, No. 4, pp. 2509-2515, 2022.

[7] D. C. Luvizon, D. Picard and H. Tabia. “2d/3d pose estimation and action recognition using multitask deep learning,” *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, pp. 5137-5146, 2018.

[8] Z. Shuo, *et al*., “Human deep squat detection method based on MediaPipe combined with Yolov5 network,” *Chinese Control Conference (CCC), Hefei, China*, pp. 6404-6409, 2022.

[9] S. Weeriya, *et al.*, “Machine Learning-Based Exercise Posture Recognition System Using MediaPipe Pose Estimation Framework,” *International Conference on Advanced Computing and Communication Systems (ICACCS)*, Vol. 1, pp. 2003-2007, 2023.

[10] B. Utkarsh, and G. Shikha, “Yoga pose detection and classification using machine learning techniques,” *Int Res J Mod Eng Technol Sci*, Vol. 3, No. 12, pp. 13-15, 2021.

[11] M. U. Islam, “Yoga posture recognition by detecting human joint points in real time using Microsoft Kinect,” *IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pp. 668-673, 2017.

[12] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, Vol. 9, No. 8, pp. 1735-1780, 1997.

[13] Workout / Exercises Video, Kaggle, https://www.kaggle.com/datasets/hasyimabdillah/workout fitness-video