Dean&Francis
ISSN 2959-6157

Human Pose Prediction Based on LSTM and MediaPipe

Qiang Zhou

Computer Science,Dongseo Univsersity
Email:zhouqiang19950222@gmail.com

Abstract:

Motion capture technology is a crucial digital animation production method in the fields of film and television and games. Unmarked motion capture technology has received increasing attention and wide application in recent years due to its ability to significantly reduce the time for data collection and processing. However, this technology is prone to generating incorrect human posture information when dealing with occlusion issues. Therefore, this paper attempts to combine Long Short-Term Memory networks (LSTM) for human posture prediction to correct or eliminate incorrect posture data, thereby enhancing the accuracy and reliability of motion capture.

Keywords: Unity3D, Mediapipe, LSTM, UCF101, YoYo, Deep learning, Motion Capture

I. INTRODUCTION

With the rapid development of the film and television and gaming industries, motion capture technology has become widely known for its ability to efficiently and conveniently create digital animations. Traditional motion capture technology is limited by equipment requirements and thus is difficult to be widely applied in all projects. Therefore, in order to better apply motion capture technology to practical projects, various motion capture technologies based on different principles have emerged[2,3].

MediaPipe, as an efficient tool in markerless motion capture technology[4,5], can significantly reduce the time for data collection and processing, making its application in monocular camera motion capture possible and thus replacing the expensive traditional motion capture solutions in the past. However, motion capture technology based on video analysis is prone to generating incorrect human pose information when dealing with issues such as occlusion. Although occasional errors in a single frame usually do not cause serious consequences, when applying human poses to skeletal animation, these errors can affect the smoothness of the animation.

In recent years, to enhance the accuracy of human pose recognition, researchers have proposed various methods, including physical-based pose correction [6] and human pose reconstruction [7]. Among them, the spatio-temporal two-stream pose recognition is a pioneering approach [8]. Compared with methods relying solely on the spatial dimension, this method significantly improves recognition accuracy by introducing temporal dimension assessment. Combined with MediaPipe, the spatio-temporal two-stream method can achieve more precise human model construction.

ISSN 2959-6157

II. RELATED WORK

A. MediaPipe

MediaPipe is a multimedia machine learning model application framework developed and open-sourced by Google, aiming to provide cross-platform efficient machine learning solutions. Fig.1 shows its logo. To achieve efficient computing on mobile platforms, MediaPipe has undergone specialized optimization, ensuring outstanding performance and efficiency.

Fig. 1. MediaPipe

MediaPipe has built a series of tool libraries, covering functions such as face recognition, pose recognition [9], and hand tracking. Therefore, developers do not need to write complex machine learning code from scratch, thus providing users with rich and intelligent experiences.

B. TensorFlow

TensorFlow is a symbolic numerical system based on dataflow programming and is widely used in the coding implementation of machine learning. Figure 3 shows its logo. Due to its ease of use and learnability, TensorFlow enables users to effectively solve practical human pose problems [10].

Fig. 2. TensorFlow

C. Unity3D

Unity3D (Fig.4) is a real-time 3D creation and operation platform, covering game development, film and television, and more. The Unity platform offers a complete set of software solutions that can transform digital information into any 2D or 3D content that can be interacted with in real time.

Fig. 3. Unity3D Engine

Unity3D employs multiple animation systems, among which skeletal animation is a widely used animation technique. The concept of skeletal animation originates from the human skeletal structure, driving the model's changes by controlling the movement and rotation of the bones.

III. PROCEDURE

A. Data analysis

When MediaPipe's Pose Solution successfully recognizes human postures, it returns the corresponding number of Landmarks based on the number of people identified in the image. As shown in Figure 4, each detection result contains 33 Landmarks.

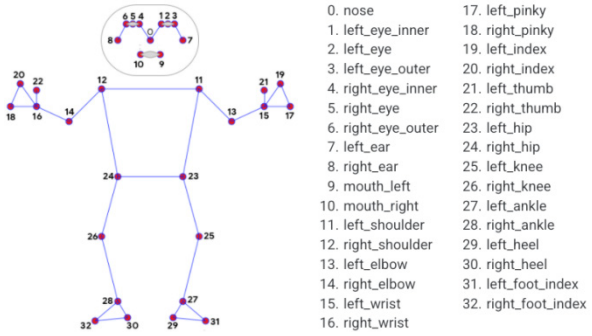

Fig. 4 pose landmarks

The obtained Landmarks cannot be directly used to generate animations and must undergo data reconstruction by simulating human joints [11]. Each Landmark consists of three components: x, y, and z, representing the coordinates of the point in three-dimensional space. Two adjacent Landmarks can form a vector, as shown in Figure 5, which indicates the direction of the bone.

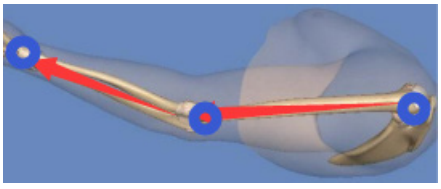

Fig. 5. Correspondence between Bones and Landmarks

The rotation angle and direction can be calculated through the dot product formula and the cross product formula.:

$$a \cdot b = \sum_{i=1}^{n} a_i b_i \tag{1}$$

$$A \times B = \|A\| \|B\| sin\theta n \tag{2}$$

It should also be noted that some of the landmarks provided by MediaPipe do not participate in the calculation of skeletal animation and should be excluded. Specifically, point 0 determines the orientation of the head, and points 7 and 8 determine the position of the head in space. Points

21 and 22 are the spatial coordinates of the thumbs. Since three points can determine a plane, points 21 and 22 do not participate in the training. Points 29 and 30 of the feet are the spatial coordinates of the heels. These points do not affect the rotation direction of the foot bones and therefore are not trained either.

Ultimately, a total of 21 Landmarks were selected as parameters for training the model, specifically including: 0, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 23, 24, 25, 26, 27, 28, 31 and 32.

B. Data acquisition

This experiment utilized the UCF101 video dataset, as shown in Figure 6. This dataset is quite challenging due to its inclusion of camera motion, variations in object appearance and posture, differences in object scale, changes in viewpoint, cluttered backgrounds, and diverse lighting conditions [12].

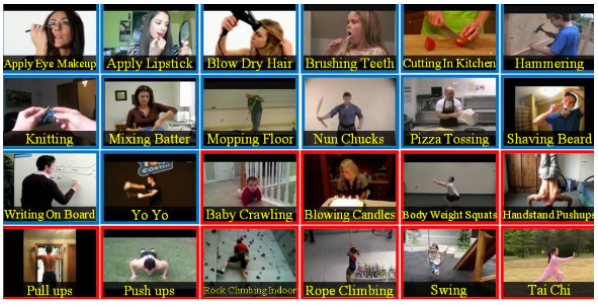

Fig. 6. UCF101

The image information provided by the video does not contain actual pose data, so manual adjustment processing is required. Figure 7 shows an example of the correct pose.

Fig. 7. Getting the correct pose and motion in Unity3D

Single-frame images cannot meet the demand for the quantity of the training set. Therefore, each frame of the video needs to be processed and adjusted to build a stable and sufficiently large dataset.

C. Model Design

Before designing the model, it is necessary to analyze the cases where human pose recognition fails. As shown in Figure 8, MediaPipe correctly identified the human pose information in the previous frame, but in the next frame, the legs crossed, resulting in detection failure. Such errors undoubtedly have a significant impact on the accuracy of animation information.

Fig. 8. MediaPipe misidentified the coordinates of both legs in subsequent frames, causing a deviation in the performance effect in Unity3D

However, we can also clearly observe that the upper body movement information identified in the consecutive frames is correct. Therefore, if all the landmarks are included in the training, it might cause the model to overlook minor recognition errors. To address this issue, I have divided the human body into five regions: left arm, right arm, left leg, right leg, and head, and trained the model separately for each region.

The model architecture adopts the Long Short-Term Memory (LSTM) network model. This model is highly suitable for analyzing time series data and can accurately capture the temporal dependencies between frames. Additionally, its gating mechanism effectively addresses the issues of vanishing and exploding gradients [13].

It can be foreseen that during high-speed movement, the changes in human body movements are extremely rapid, and the action trajectory is usually changed multiple times within one second. Therefore, the number of predicted steps should not be too large. Ultimately, I chose the Mean Squared Error (MSE) as the loss function and used the Adam optimizer to design the LSTM model as shown in Fig.9.

ISSN 2959-6157

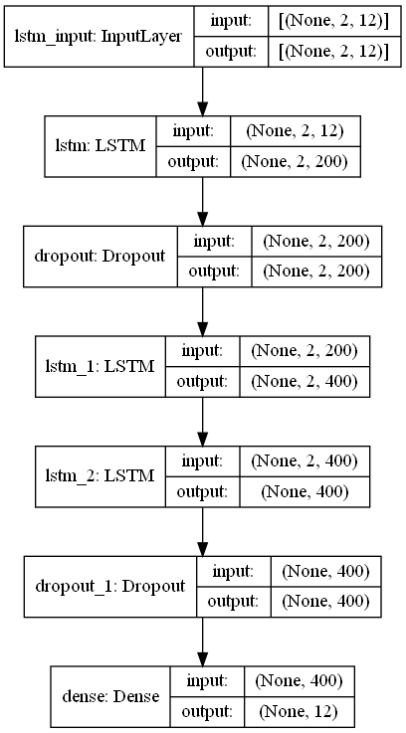

Fig. 9. Model

This model is used for the training and recognition of leg postures. There are 4 landmarks on the leg, so 12 input and output nodes are required. After 64 epochs of training, it can be seen that the error decreases rapidly.

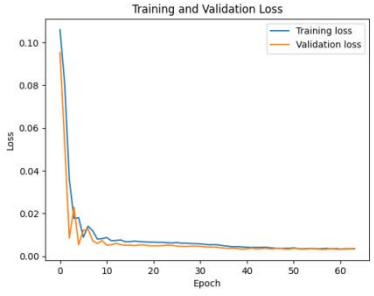

Fig. 10. Results

D. Predictive analysis

After the model training is completed, the video is imported into MediaPipe to extract the recognized pose information and input it into the model for analysis and prediction. As shown in Fig.11, MediaPipe recognizes the human pose data in the video and passes it to the model for analysis. The model predicts the movement trajectory of the third frame based on the action trends of the first two frames. In frames 97 and 110 of the video v_YoYo_g19_c04, there were problems with the recognition of the leg landmarks. After processing and prediction by the LSTM model, a new digital animation was generated, basically avoiding the original recognition errors.

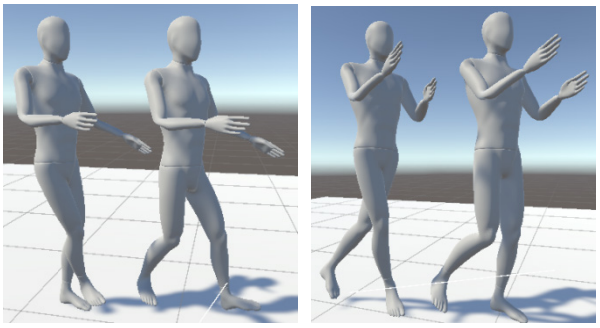

Fig. 11. The first image shows the comparison of the animation before and after processing the pose information at frame 97; the second image shows the corresponding situation at frame

IV. EXPERIMENTAL RESULTS

Firstly, in this experiment, LSTM was applied to predict and correct the video v_YoYo_g19_c04 in the UCF101 dataset. The corrected animation information fixed the recognition errors in most cases. However, as shown in Figure 12, the predicted animation still had the problem of crossed legs. This was because the movements in the video were very smooth, which led the model to fail to identify this error and thus failed to correct it successfully.

Fig. 12. In the 71st frame of the animation, the spatial positions of the legs front and back have been corrected, but the problem of crossing of predicted coordinates still exists.

V. CONCLUSION AND FUTURE WORK

In this experiment, the trained LSTM model was able to accurately achieve the expected goal, correcting the animation information misidentified by MediaPipe and converting it into the required animation. However, although the UCF101 video set was used, it did not cover all types of actions, and only the YoYo video animation was utilized. The reason is that creating the corresponding training set requires a significant amount of manual labor, making it impossible to precisely reproduce all actions in the training set. Additionally, due to the frame rate differences among various videos, there were deviations in action prediction, which was one of the problems discovered

during the prediction process.

Therefore, using the LSTM model to correct the predicted human pose information is an effective method, but under the limitations of the training set and video sources, there are still certain restrictions. In the future, if the diversity and scale of the training set can be increased, the accuracy of the prediction results will be further improved.

References

[1] Liu, S., Zhang, J., Zhang, Y. et al. A wearable motion capture device able to detect dynamic motion of human limbs. Nat Commun 11, 5615 (2020). https://doi.org/10.1038/s41467-020-19424-2

[2] [2]Robert M. Kanko, Elise K. Laende, Elysia M. Davis, W. Scott Selbie, Kevin J. Deluzio,Concurrent assessment of gait kinematics using marker-based and markerless motion capture,Journal of Biomechanics,Volume 127,(2021),110665,ISSN 0021-9290,https://doi.org/10.1016/j.jbiomech.2021.110665.

[3] Takeda, I., Yamada, A., & Onodera, H. (2020). Artificial Intelligence-Assisted motion capture for medical applications: a comparative study between markerless and passive marker motion capture. Computer Methods in Biomechanics and Biomedical Engineering, 24(8), 864–873. https://doi.org/10.1080/10255842.2020.1856372

[4] Robert M. Kanko, Elise Laende, W. Scott Selbie, Kevin J. Deluzio,Inter-session repeatability of markerless motion capture gait kinematics,Journal of Biomechanics,Volume 121,2021,110422,ISSN 0021-9290,https://doi.org/10.1016/j.jbiomech.2021.110422.

[5] Wade L, Needham L, McGuigan P, Bilzon J. 2022. Applications and limitations of current markerless motion capture methods for clinical gait biomechanics. PeerJ 10:e12995 https://doi.org/10.7717/peerj.12995

[6] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. 2020. PhysCap: physically plausible monocular 3D motion capture in real time. ACM Trans. Graph. 39, 6, Article 235 (December 2020), 16 pages. https://doi.org/10.1145/3414685.3417877

[7] [10]L. Malekian and R. Lapeer, "Self-Occluded Human Pose Recovery in Monocular Video Motion Capture," 2024 14th International Conference on Pattern Recognition Systems (ICPRS), London, United Kingdom, 2024, pp. 1-6, doi: 10.1109/ICPRS62101.2024.10677815. keywords: {Optical filters;Measurement;Three-dimensional displays;Motion estimation;Pose estimation;Predictive models;Cameras;human pose estimation;self-occlusion;single view video;SMPL model;machine learning;deep learning},

[8] Chenchu Xu, Zhifan Gao, Heye Zhang, Shuo Li, Victor Hugo C. de Albuquerque,Video salient object detection using dual-stream spatiotemporal attention,Applied Soft Computing,Volume 108,2021,107433,ISSN 1568-4946,https://doi.org/10.1016/j.asoc.2021.107433.

[9] Kim, J.-W.; Choi, J.-Y.; Ha, E.-J.; Choi, J.-H. Human Pose Estimation Using MediaPipe Pose and Optimization Method Based on a Humanoid Model. Appl. Sci. 2023, 13, 2700. https://doi.org/10.3390/app13042700

[10] L. Xie and X. Guo, "Object Detection and Analysis of Human Body Postures Based on TensorFlow," 2019 IEEE International Conference on Smart Internet of Things (SmartIoT), Tianjin, China, 2019, pp. 397-401, doi: 10.1109/SmartIoT.2019.00070. keywords: {object detection;body postures;TensorFlow;Faster R-CNN},

[11] Kim, J.-W.; Choi, J.-Y.; Ha, E.-J.; Choi, J.-H. Human Pose Estimation Using MediaPipe Pose and Optimization Method Based on a Humanoid Model. Appl. Sci. 2023, 13, 2700. https://doi.org/10.3390/app13042700

[12] Doan T N. An efficient patient activity recognition using LSTM network and high-fidelity body pose tracking[J]. International Journal of Advanced Computer Science and Applications, 2022, 13(8).

[13] Shrestha M, Pandey S P. Human action recognition using deep learning methods[C]//Proceedings of the International Conference on Machine Learning and Data Engineering. Berlin/Heidelberg, Germany: Springer Nature, 2023: 345-356.