# 3D Human Motion Reconstruction in Unity with Monocular Camera

Tai-Wei Chen
National Chung Hsing University
Taichung City, Taiwan, ROC
ks800223@gmail.com

Wei-Liang Lin
National Chung Hsing University
Taichung City, Taiwan, ROC
wllin@nchu.edu.tw

*Abstract*— **This paper using a 3D pose estimator to predict human 3D poses. By combining the pose sequence information as a motion capture, we could reconstruct the human motion in Unity with any appearance. A potential application is collecting a compact human 3D activity dataset.**

## I. INTRODUCTION

With the rapid development of deep learning technology, more and more researchers are studying AI related fields. The construction of human motion done in this paper involves many sub-projects, such as the flow of people in the navigation system, people behavior prediction, people tracking, etc.

Many dangerous scenarios, such as the sudden encounter of pedestrians while training autonomous driving, are difficult to collect in the real world.

But in Unity, you can easily simulate a variety of situations. Observation from any angle is easy. Furthermore, scripts can usually help annotation.

Through a 2D pose estimation, we obtain a pedestrian's bounding box. Combined with tracking by detection, we intercept each pedestrian from every frame as the input to the 3D pose estimator as shown in Figure 1.

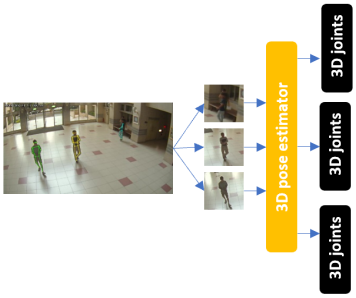


Figure 1. Get each frame from the video stream and do AlphaPose to calculate the position of the pedestrian through the skeleton as input.

We convert human 3D joints that predicted by 3D pose estimator into motion capture (BVH format), and then, the FBX format through Blender. Unity can access FBX format (Figure 2).

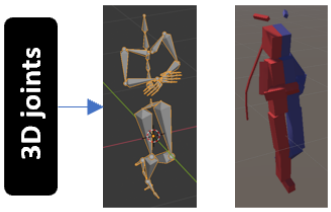


Figure 2. With this skeleton animation, you can apply it to any human model in Unity

Finally, regarding the paths of pedestrians, we use Zhang's camera calibration method [5] to convert pixel coordinates into 3D coordinates which is used in Unity.

## II. RELATED WORK

**AlphaPose [3]:** In obtaining the bounding box, we first adopted traditional object detection method. However, the results were not satisfactory because the bounding box obtained through traditional object detection method usually vary within a large range. But bounding boxes calculated through the estimated poses does not have this problem. We chose AlphaPose to find poses. With pose information, we can easily calculate a pedestrian center and separate each pedestrian.

**MOTDT [2]:** Before entering the 3D pose estimator, it is necessary to do pedestrian tracking in order to obtain continuous inputs. Since we use AlphaPose as preprocess, we also tried a companion software, PoseFlow, which is a tracking algorithm based on pose information previously. However, when the pedestrians on the screen overlap, this method loses the pedestrian ID. We combine the results of AlphaPose and MOTDT algorithm to solve this problem.

**3D Pose Estimator [1]:** There is a lot of 3D reconstruction research, but most of them focus on the restoration of mesh [6]. Different from them, we don't use SMPL model since we

just need the 3D joints. We use pre-trained model [1] to predict 3D joints and combine them into a motion capture.

**Zhan's camera calibration:** Figure 3 show the conversion formula, where $S$ is a scale factor, u, v represents pixel coordinates. With the intrinsic-parameter matrix and the extrinsic-parameter matrix, we can convert u, v to $X\ Y\ Z$. Since all we need is the paths of pedestrians, we can simplify the formula and ignore $Z$ [5].

$$S\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{dx} & 0 & u_0 \\ 0 & \frac{1}{\mathrm{d}y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} * [R \mid T] * \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

Figure 3. conversion formula of pixel coordinate to 3D coordinate

## III. APPLICATIONS

Currently, we have two possible applications: 4D Map and navigation environment.

4D Map: Unlike GOOGLE map, this method can record the history in a specific field because it contains continuous time information. And in the Unity virtual environment, a person under the monitor has become a virtual model, so there is no privacy issue.

Navigation environment: Reinforcement learning usually uses a virtual environment since it requires a lot of iterations. However, in the virtual environment, the movement of pedestrians is often subject to a fixed mathematical formula. Our method collects real world data, and the resulting data can be a supplement to the original virtual environment.

## IV. RESULTS AND CONCLUSIONS

With this motion capture, we not only record 3D information of pedestrians, but also time information. Compared with photo format storage, motion capture only records skeleton motion information, greatly reducing the storage space.

Table I. compares the size of the motion capture with 3 pedestrians and 150 frames in different data types. Our BVH shows more space-saving. Also, our BVH is not limited by the mesh, and the same motion capture can be applied to any appearance model. A future improvement is to just save BVH information in TEXT format.

**Acknowledgements** The study is supported by Ministry of Science and Technology (MOST) under MOST 109-2218-E-005-011.

TABLE I.

| | ***File size (Mb)*** | ***Compression ratio (%)*** | ***3D information*** |
|---|---|---|---|
| Photos from film (JPG) | 26.1 | 0 | No |
| BVH | 1.1 | 95.8 | Yes |
| FBX | 4.8 | 81.6 | Yes |

## REFERENCES

[1] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In CVPR, 2018J.

[2] C Long, A Haizhou, Z Zijie, and S Chong. 2018. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. ICME.

[3] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In ICCV, 2017.

[4] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," arXiv preprint arXiv:1802.00977, 2018.

[5] Z. Zhang, "A flexible new technique for camera calibration," IEEE Trans. Pattern Anal. Mach.

[6] T. Alldieck, M. Magnor, W. Xu, C. Theobalt and G. Pons-Moll Video Based Reconstruction of 3D People Models, Computer Vision and Pattern Recognition (CVPR), 2018.