A Human Pose Estimation Method from Pseudo-Captured Two-Viewpoints Video

Ayaka Shimazu[1], Chun Xie[2][a], Satoru Tanigawa[3] and Itaru Kitahara[2][b]
[1]*Master's and Doctoral Program in Intelligent Mechanical Interaction Systems, University of Tsukuba, Ibaraki, Japan*
[2]*Center for Computational Sciences, University of Tsukuba, Ibaraki, Japan*
[3]*Faculty of Health and Sport Sciences, University of Tsukuba, Ibaraki, Japan*

Keywords: 3D Human Pose Estimation, Performance Analysis, Pseudo-Binocular Filming, Time-Warping.

Abstract: Motion analysis utilizing human pose estimation has garnered increasing attention within sports science, serving both preventive medicine and skill enhancement purposes. While techniques using 3D trackers and RGB-D cameras to estimate human poses are gaining attention, the widespread adoption is hindered by the requirement for extensive space and specialized equipment. This paper introduces a novel method to estimate the 3D human pose using RGB video data captured from 'pseudo' two-viewpoints. This approach involves performing the same motion in different directions and recording with a single camera. We confirm that the accuracy of 3D human pose estimation from video taken by a single camera is improved by the pseudo-two-viewpoints recording compared to existing methods using a single monocular RGB camera.

1 INTRODUCTION

This paper proposes a shooting video method called "pseudo-two-viewpoints recording" to achieve human pose estimation with sufficient accuracy for movement analysis. As illustrated in Figure 1, Pseudo-two-viewpoints recording involves using a single RGB camera to record the same movement performed twice but in different orientations by the same individual.

In sports science, performance analysis is a crucial component of preventive medicine and technical improvement. Movement analysis is typically based on human pose estimation, which utilizes videos to accurately capture and analyze athlete movements. This paper introduces a novel video capturing and processing method for human pose estimation and demonstrate our method using jumping as a specific example, which is closely tied to lower body performance and correlates with the risk of lower body injuries (Hewett et al. 2005).

Existing methods for 3D human pose estimation include marker-based methods (Bodenheimer et al. 1997), RGB-D camera-based methods (Zimmermann et al. 2018), monocular camera image-based methods (Liu et al. 2020), and multi-viewpoint image-based methods (Chen et al. 2020). While marker-based and RGB-D camera-based methods offer high accuracy in pose estimation, they require extensive space and specialized equipment, posing challenges for use beyond professional athletes. Conversely, the field of exercise physiology is increasingly recognizing the importance of preventive medicine for a diverse range of athletes, including amateur and junior athletes in addition to professionals.

Using a monocular camera to estimate human pose information is advantageous for acquiring measurement data easily. However, the accuracy of depth information derived from single-viewpoint observations is insufficient, posing challenges for movement analysis applications. By capturing motion from multiple viewpoints, we can estimate the 2D coordinates of the joints at each viewpoint and triangulate the 3D coordinates from matched joints in different views. In this way, the ambiguity in depth estimation can be resolved, leading to more accurate human pose estimation compared to the monocular video. However, this approach requires the use of multiple synchronized

[a] https://orcid.org/0000-0003-4936-7404
[b] https://orcid.org/0000-0002-5186-789X

Shimazu, A., Xie, C., Tanigawa, S. and Kitahara, I.
A Human Pose Estimation Method from Pseudo-Captured Two-Viewpoints Video.
DOI: 10.5220/0012921400003828
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 12th International Conference on Sport Sciences Research and Technology Support (icSPORTS 2024)*, pages 148-154
ISBN: 978-989-758-719-1; ISSN: 2184-3201

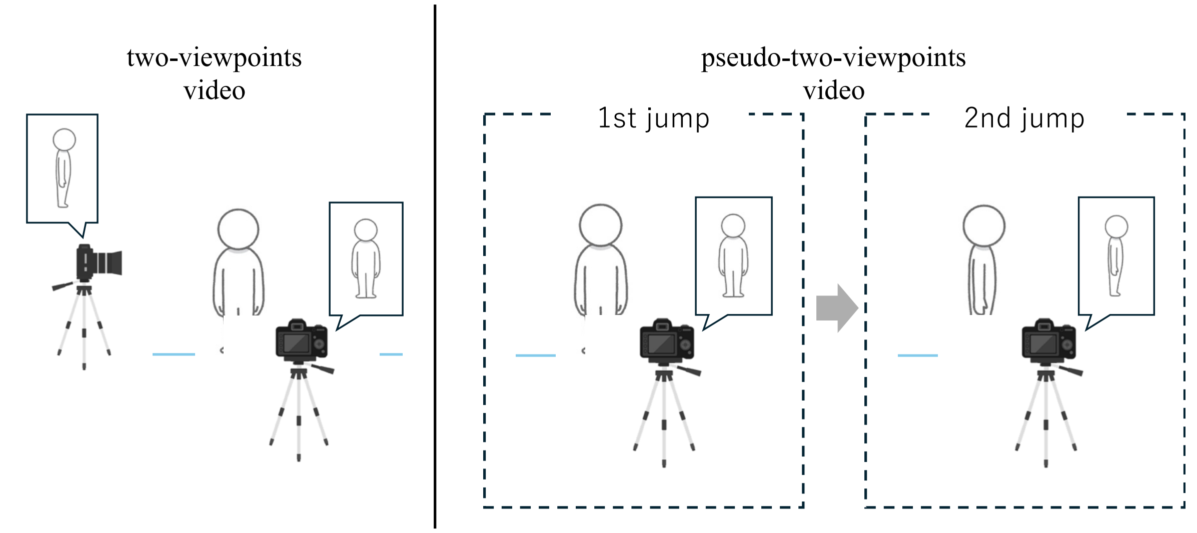

Figure 1: Two-viewpoints recording (left) and pseudo-two-viewpoints recording (right). In two-viewpoints recording, a frontal image and a side image are captured using two RGB cameras. In pseudo-two-viewpoints recording, the frontal and side images are obtained by changing the direction of the body and recording twice with a single RGB camera.

RGB cameras, which lead to extra cost and the setup process can be cumbersome. We therefore propose a new approach to solve this problem by estimating 3D human pose information from multi-view (pseudo-two-viewpoints) video data acquired using only a single stationary RGB camera.

The primary challenge in achieving human pose estimation through pseudo-two-viewpoints recording lies in temporal and spatial alignment. Temporal deviations occur due to varying movement timings, while spatial deviations arise from changes in camera and subject orientation. To address these issues, we introduce a time-warping method to correct temporal misalignment, along with camera calibration and joint triangulation, to consistently produce a 3D pose from two RGB videos. Our method achieves sufficient accuracy for movement analysis without requiring extensive space or specialized equipment, thus simplifying the process for various athletes.

2 RELATED WORKS

3D human pose estimation can be broadly classified into marker-based and marker-less methods. Timothy et al. utilized 25 reflective markers on the lower body to measure the posture and load on the knee during a jump landing, recording the 3D coordinates of the joints (Hewett et al. 2005).

Marker-less posture estimation methods can be divided into two types: those that estimate joint positions as 2D coordinates (2D posture estimation) and those that estimate 3D coordinates (3D posture estimation). An example of a 2D pose estimation method is HR-Net (Sun et al. 2019). This top-down method maintains high resolution throughout the process by adding subnetworks from high resolution to low resolution and connecting multi-resolution networks in parallel, achieving better results than bottom-up methods for single-person pose estimation. However, 2D pose information alone is insufficient for analyzing sports movements performed in 3D space, necessitating 3D pose estimation.

3D pose information is generally estimated by applying triangulation based on camera position and pose information to 2D pose data obtained from images taken from different viewpoints. This method, however, requires considerable time and expertise to set up and synchronize cameras.

GAST-Net (Liu et al. 2020) exemplifies a solution to this issue. By applying Graph Convolution Networks (GCN) to the time series information of the skeleton, it addresses the self-occlusion problem where joints are obscured by the subject's own body. Nevertheless, the accuracy of 3D human pose estimation using a monocular camera remains insufficient for sports motion analysis when compared to estimation from multiple viewpoints.

In this research, we propose a human pose estimation method that ensure both ease of photography and accuracy of pose estimation using pseudo multiple viewpoints captured by a monocular RGB camera.

3 PROPOSED METHOD

Figure 2 illustrates the processing flow of the proposed method. A single fixed camera is used to capture images from two different angles (pseudo-two-viewpoints recording) by changing the subject's body direction and repeating the same action twice. The camera position and orientation (extrinsic parameters) are estimated based on landmarks set in the scene. 2D pose estimation is first performed on the two captured videos separately. Dynamic Time Warping (DTW) (Müller 2007) is then applied to the estimated 2D pose time series information to compensate for the temporal misalignment between the two series. Finally, the 3D positions of the joints are estimated by triangulating the corresponding 2D joints using the camera intrinsic and extrinsic parameters estimated in advance.

3.1 Definition of Coordinate Systems

Figure 3 shows the coordinate system defined this paper. The origin of the world coordinate system W is the center of the hula hoop placed in the scene. The X-axis is parallel to the horizontal axis of the image plane of camera 1, and the Y-axis is parallel to the horizontal component of the optical axis of camera 1. The Z-axis is obtained by the cross product of the X and Y axes and is orthogonal to the ground on which the hula hoop is placed. The camera coordinate system c_1 takes the optical center of camera 1 as the origin, with the horizontal axis of the image plane as the X_1-axis (points to the right), the vertical axis as the Y_1-axis (points downward), and the optical axis as the Z_1-axis (points forward). Let (t_x, t_y, t_z) be the world coordinate of the optical center of camera 1, and given that t_x is always zero, the rigid body transformation from world frame basis W to camera frame basis c_1 in homogenous coordinates is expressed by Equation (1).

$$\boldsymbol{c}_1 = \begin{bmatrix} 1 & 0 & 0 & t_x \\ 0 & 0 & -1 & t_z \\ 0 & 1 & 0 & -t_y \\ 0 & 0 & 0 & 1 \end{bmatrix} \boldsymbol{W} \tag{1}$$

In the pseudo-two-viewpoints recording, the position and orientation of the virtually positioned camera 2 need to be determined based on the viewport of camera 1, as it is virtually moved due to the subject's orientation change. Assuming that the camera 2 is a rotation of camera 1 by an angle θ around the Z-axis of the world coordinate system, the rotation matrix R_{12} between c_1 and c_2 are expressed by Equation (2).

$$\boldsymbol{R}_{12} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

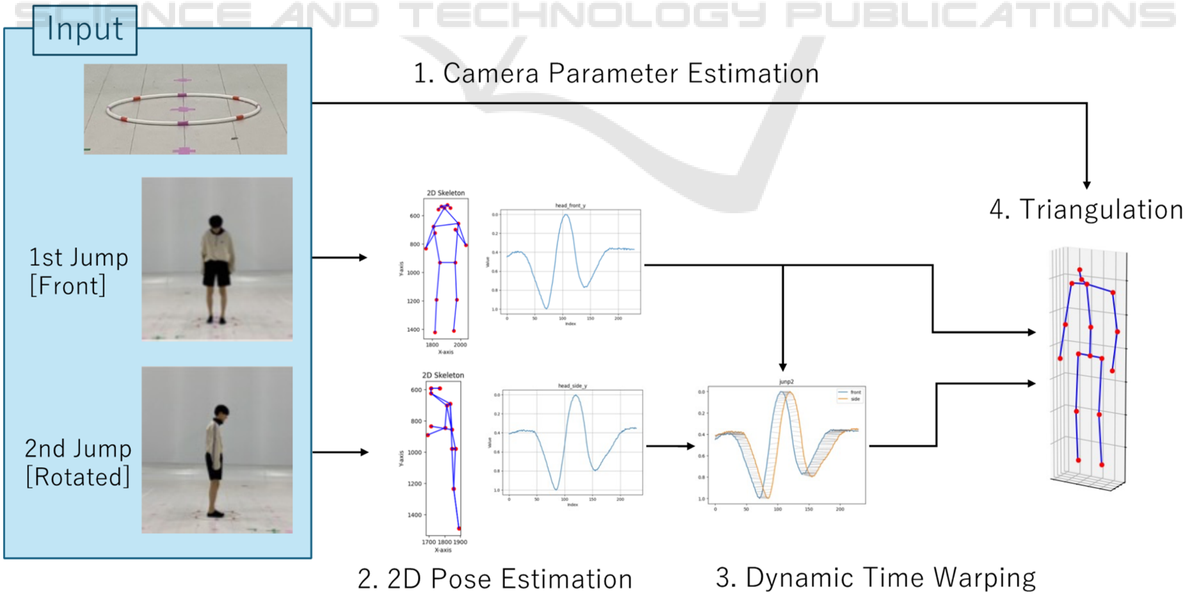

Figure 2: The 3D human pose is estimated using two videos of jumps in different directions as input. The 3D coordinates of the joints are estimated by applying joint-based time alignment and joint coordinate triangulation.

3.2 Spatial Alignment

3.2.1 Camera Position Estimation

The position and orientation (extrinsic parameters) of camera 1 are obtained using landmarks whose 3D coordinates are known. In this research, a hula hoop with markings is used as a landmark, as shown in Figure 4. Using the correspondence information between the 3D coordinates of the four landmark points and their 2D coordinates observed in the image, a rotation matrix $\boldsymbol{R}_1$ and a translation matrix $\boldsymbol{t}_1$ are obtained using the IPPE method (Collins and Bartoli 2014). The intrinsic parameters $\boldsymbol{K}$ are determined based on Zhang's method(Zhang 2000). The projective transformation matrix $^1\boldsymbol{P} = \boldsymbol{K}(\boldsymbol{R}_1|\boldsymbol{t}_1)$ is then obtained by combining the intrinsic parameters $\boldsymbol{K}$ with the extrinsic parameters.

3.2.2 Pseudo-Camera Position Estimation

The intrinsic parameters of the second viewpoint camera are the same as those for the first viewpoint. From rotation in Equation (2), the perspective projection matrix of the second viewpoint can be obtained by Equation (3).

$$^2\boldsymbol{P} = \boldsymbol{K}\left(\boldsymbol{R}_1\begin{bmatrix}\cos\theta & -\sin\theta & 0\\ \sin\theta & \cos\theta & 0\\ 0 & 0 & 1\end{bmatrix}\middle|\boldsymbol{t}_1\right) \qquad (3)$$

The rotation angle θ of the camera is considered to coincide with the rotation angle of the subject's head around the Z-axis. Therefore, we first estimate the head posture using 6DRepNet360 (Hempel, Abdelrahman, and Al-Hamadi 2023) when the subject is standing upright in both videos (Figure 5), and estimate θ as the relative rotation of head around Z-axis.

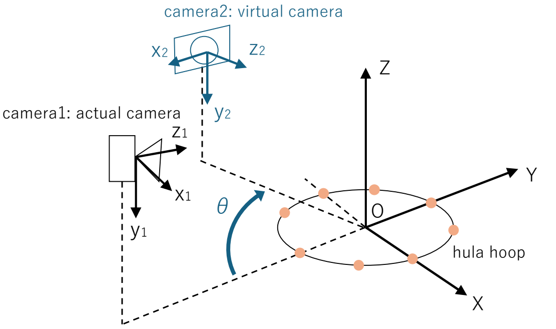

Figure 3: World coordinate system and camera coordinate system.

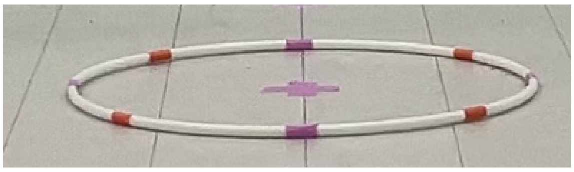

Figure 4: A hula hoop with target markers.

3.3 Temporal Alignment

The difference in the timing of the start of movement and the length of the dwell time causes a temporal posture shift between the videos. Figure 6 illustrates the processing flow of the temporal alignment. To correct the temporal misalignment, we focus on the head movement, which usually has minimum occlusion and allows us to obtain relatively stable coordinates of the joint points. In detail, HR-net (Sun et al. 2019) is applied to the video frames and extract the vertical coordinates of the subject's head over time. Using these coordinates, three key frames in both videos are detected: the maximum bending point before the jump, the maximum reaching point, and the maximum bending point after the landing. After that, DTW is applied to match these three key frames and dynamically adjust the speed of the virtual viewport video to align it with the actual viewport video. By doing this, we obtain a pair of time-aligned 2D pose data for the subject's jump.

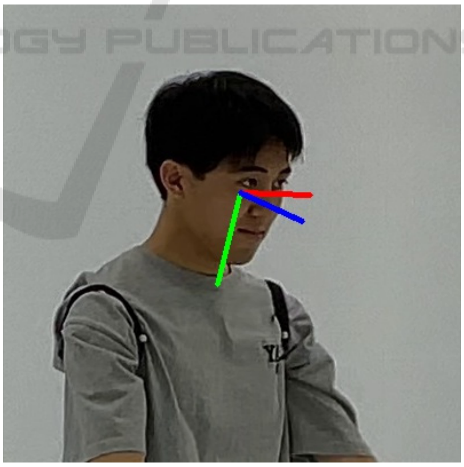

Figure 5: Visualization of head pose estimation results using 6DRepNet360.

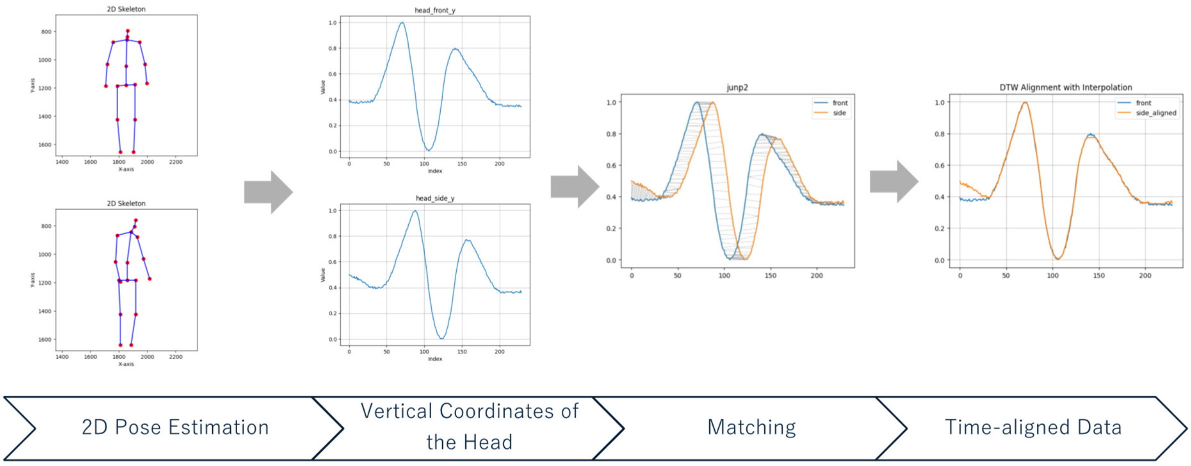

Figure 6: Time alignment using head vertical coordinates: the head vertical coordinates are taken from the 2D pose estimation results and matched by DTW for temporal alignment.

3.4 Estimation of 3D Coordinates of Joint Points

The 3D coordinates of the joints are calculated from the time-corrected 2D skeleton using the method described in section 3.2 and the camera parameters obtained in section 3.1. Let $\boldsymbol{X}$ be the world coordinate of the joint point and (u_1, v_1) and (u_2, v_2) be the coordinates on the images from cameras 1 and 2, respectively. The following equation holds:

$$\begin{cases} u_1 {}^1\boldsymbol{P}_3^T\boldsymbol{X} = {}^1\boldsymbol{P}_1^T\boldsymbol{X} \\ v_1 {}^1\boldsymbol{P}_3^T\boldsymbol{X} = {}^1\boldsymbol{P}_2^T\boldsymbol{X} \\ u_2 {}^2\boldsymbol{P}_3^T\boldsymbol{X} = {}^2\boldsymbol{P}_1^T\boldsymbol{X} \\ v_2 {}^2\boldsymbol{P}_3^T\boldsymbol{X} = {}^2\boldsymbol{P}_2^T\boldsymbol{X} \end{cases} \tag{4}$$

By solving this system of equations, we can obtain the 3D coordinates $\boldsymbol{X}$ of the joint points. This process takes as input the 2D coordinates corresponding to each of the seven joint points in each frame of the two scenes and the camera parameters obtained in section 3.1. The output includes 17 joints, and the skeletal connections between joints are drawn for each frame to generate the 3D skeletal time series data of the jumping motion.

4 EXPERIMENTS

4.1 Experimental Setup

In the experiment, a RGB camera mounted on Apple iPadPro 12.9 (6th generation) is used to capture the videos. We note that the iPad is frequently used to record video in sports scenes. The camera is fixed on a tripod so that the optical center is 1.14 meters above the ground. The roll and tilt angles of the camera are set to zero, and the pan angle is adjusted so that the subject appears at the center of the screen. A white hula hoop with a diameter of 0.8 meters is used as a landmark for calibration and as a marker for the jump position. The motion video is captured with a resolution of 3840 pixels × 2160 pixels and a frame rate of 60 fps with no zooming (maximum wide angle).

As shown in Figure 7, the participant stands upright in the center of the hula hoop and performs a vertical jump with the camera facing forward. Before and after each jump, the participants were instructed to stand still in the upright position for about 3 seconds.

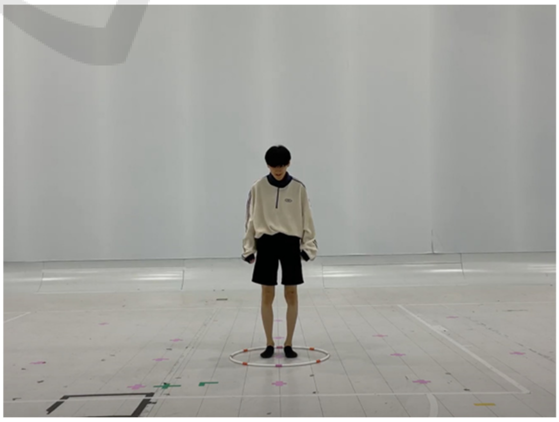

Figure 7: Scene of the shooting experiment. Participants stand upright in the center of the hula hoop and perform a vertical jump.

4.2 Evaluation Metrics

As an index of 3D human pose estimation accuracy, we use Mean Per Joint Position Error (MPJPE), which is the average distance between the predicted and reference positions of a joint point, and P-MPJPE, which is calculated after a rigid body transformation of the ground truth (GT) for translation, rotation, and scale. P-MPJPE is calculated with respect to a coordinate system transformed with the coordinates of the waist as the origin. The Percentage of Correct 3D Keypoints (3DPCK) is an index that indicates the percentage of successfully detected joints, where the distance between the predicted position of a joint and the reference position is within a predefined threshold. In this experiment, the threshold for 3DPCK is 150 mm, which is commonly used.

4.3 Results

4.3.1 Quantitative Evaluation

Table 1 shows the estimation accuracy of the 3D skeleton from the pseudo-two-viewpoints video. The MPJPE and 3DPCK of the proposed method are 106.5 mm and 86.0%, respectively. The P-MPJPE of the proposed method is less than half that of the baseline method. This confirms that the proposed method improves the accuracy of 3D skeleton estimation. The standard deviation of P-MPJPE has also decreased, indicating that the proposed method is stable with less variation in the estimated result. Figure 8 shows the P-MPJPE for each joint. The MPJPE of the proposed method is smaller for all joints. The standard deviations are also smaller for all joints, indicating that the proposed method is stable. In particular, the MPJPE of the lower body (hips, knees, and ankles), which is an important index in jump motion analysis, is kept low, suggesting that the 3D skeletal posture estimation from the pseudo-two-viewpoints video is effective for performance analysis purposes.

Table 1: Accuracy of 3D skeleton estimation. The proposed method demonstrates greater accuracy compared to the existing method.

	MPJPE [mm] ↓	P-MPJPE [mm] ↓	P-MPJPE std ↓	3DPCK [%] ↑
GAST-Net	-	163.8	144.2	-
Ours	97.99	53.41	72.52	89.6

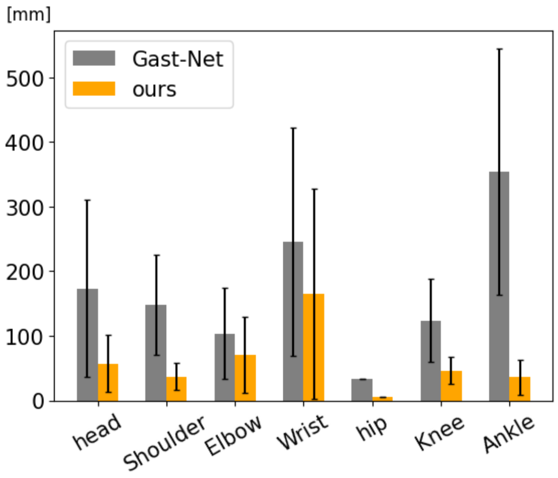

Figure 8: Comparison of P-MPJPE for each joint between the estimation results of the existing method (GAST-Net) and the proposed method. Error bars indicate standard deviations. The proposed method demonstrates greater accuracy for all joints.

4.3.2 Qualitative Evaluation

The estimated 3D poses are evaluated qualitatively. The three estimation results compared are the reference image estimated by the two-viewpoints recording (ground truth), the image estimated by the pre-trained GAST-Net, and the image estimated by the proposed method. The estimation results at two representative time points are shown in Figure 9. The results show that both GAST-Net and our method produce sufficiently accurate results at the reaching point, while our method significantly outperformed GAST-Net for forward-leaning and knee-bending motions, such as the maximum bending before jumping and bending after landing.

5 LIMITATIONS

The accuracy of human pose estimation in this system depends on the similarity between the two repeated motions. If the similarity between the two motions is not enough, some errors may occur during time alignment and triangulation. Acceptable thresholds for motion repeatability errors are currently under investigation and require verification using a larger dataset. Repeating the motion multiple times may also reduce reproducibility due to fatigue or other factors. Moreover, since this system replaces traditional two-viewpoint recordings with two separate recordings of the same motion, it requires twice the number of recordings compared to conventional two-viewpoint motion analysis.

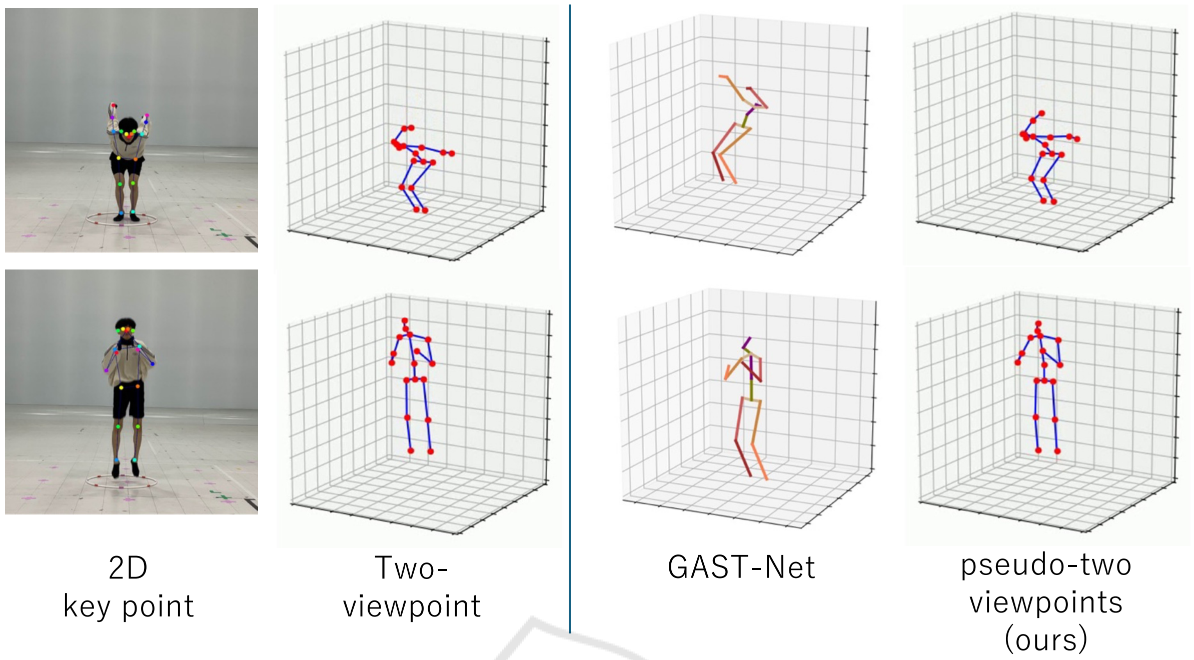

Figure 9: Visualization of 3D human pose estimation results for bending (top) and reaching point (bottom). The accuracy of GAST-Net is good at the reaching point but significantly decreases during bending motions, while the proposed method accurately estimates the 3D coordinates of joints in both situations.

6 CONCLUSIONS

This paper proposed a method for estimating the 3D human pose from pseudo-two-viewpoints video using only a single monocular RGB camera, to construct a 3D human pose estimation system that could easily capture images. Spatiotemporal deviations caused by the use of pseudo-two-viewpoints images were compensated by camera calibration and Dynamic Time Warping (DTW). Experimental results showed that the proposed method improves estimation accuracy compared to existing methods.

REFERENCES

Bodenheimer, Bobby, Chuck Rose, Seth Rosenthal, and John Pella. 1997. “The Process of Motion Capture: Dealing with the Data.” In *Eurographics*, 3–18. Vienna: Springer Vienna.

Chen, Liangjian, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, and Xiaohui Xie. 2020. “MVHM: A Large-Scale Multi-View Hand Mesh Benchmark for Accurate 3D Hand Pose Estimation.” *ArXiv [Cs.CV]*. arXiv. http://arxiv.org/abs/2012.03206.

Collins, Toby, and Adrien Bartoli. 2014. “Infinitesimal Plane-Based Pose Estimation.” *International Journal of Computer Vision* 109 (3): 252–86.

Hempel, Thorsten, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 2023. “Towards Robust and Unconstrained Full Range of Rotation Head Pose Estimation.” *ArXiv [Cs.CV]*. arXiv. http://arxiv.org/abs/2309.07654.

Hewett, Timothy E., Gregory D. Myer, Kevin R. Ford, Robert S. Heidt Jr, Angelo J. Colosimo, Scott G. McLean, Antonie J. van den Bogert, Mark V. Paterno, and Paul Succop. 2005. “Biomechanical Measures of Neuromuscular Control and Valgus Loading of the Knee Predict Anterior Cruciate Ligament Injury Risk in Female Athletes: A Prospective Study.” *The American Journal of Sports Medicine* 33 (4): 492–501.

Liu, Junfa, Juan Rojas, Zhijun Liang, Yihui Li, and Yisheng Guan. 2020. “A Graph Attention Spatio-Temporal Convolutional Network for 3D Human Pose Estimation in Video.” *ArXiv [Cs.CV]*. arXiv. http://arxiv.org/abs/2003.14179.

Müller, Meinard. 2007. “Dynamic Time Warping.” In *Information Retrieval for Music and Motion*, 69–84. Berlin, Heidelberg: Springer Berlin Heidelberg.

Sun, Ke, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. “Deep High-Resolution Representation Learning for Human Pose Estimation.” *ArXiv [Cs.CV]*. arXiv. http://arxiv.org/abs/1902.09212.

Zhang, Z. 2000. “A Flexible New Technique for Camera Calibration.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (11): 1330–34.

Zimmermann, Christian, Tim Welschehold, Christian Dornhege, Wolfram Burgard, and Thomas Brox. 2018. “3D Human Pose Estimation in RGBD Images for Robotic Task Learning.” *ArXiv [Cs.CV]*. arXiv. http://arxiv.org/abs/1803.02622.