IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 July 2024, accepted 4 August 2024, date of publication 16 August 2024, date of current version 2 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3444790

RESEARCH ARTICLE

3D Baseball Pitcher Pose Reconstruction Using Joint-Wise Volumetric Triangulation and Baseball Customized Filter System

YUN-WEI CHIU[1], KUEI-TING HUANG[2], YUH-RENN WU[2], (Senior Member, IEEE), JYH-HOW HUANG[3], WEI-LI HSU[4], AND PEI-YUAN WU[5], (Member, IEEE)

[1]Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan
[2]Graduate Institute of Photonics and Optoelectronics, National Taiwan University, Taipei 10617, Taiwan
[3]Department of Sport Information and Communication, National Taiwan University of Sport, Taichung 40404, Taiwan
[4]Graduate Institute of Physical Therapy, National Taiwan University, Taipei 10055, Taiwan
[5]Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan

Corresponding author: Pei-Yuan Wu (peiyuanwu@ntu.edu.tw)

This work was supported in part by the National Science and Technology Council under Grant 113-2425-H-028-002, Grant 112-2425-H-028-002, Grant 111-2223-E-002-004-MY3, and Grant 112-2622-8-002-022; and in part by the Ministry of Education under Grant 112L900901.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of National Taiwan University Hospital (NTUH) under Application NTUH-REC No. 201807126RINA.

ABSTRACT 3D human pose estimation (HPE) has become increasingly important in baseball analytic, but there are several difficulties pertaining to pose estimation in real-world baseball pitching. First, in-the-wild baseball pitching lacks related 3D pose datasets and contains lots of joints occluded by other body parts. Second, baseball pitching contains dramatic velocity changes during arm acceleration phases. Due to these properties of pitching, it is difficult to use common filters to remove random noises while preserving high-frequency critical joint movements in pitching. To solve these problems, we propose joint-wise volumetric triangulation to reconstruct 3D human poses by utilizing the information of multi-view 2D joint heatmaps generated by 2D HPE methods. We also designed a baseball-customized filter system to remove noisy signal from pose movement while preserving the high-frequency pitching motion. Our proposed pose reconstruction scheme yields a 33.1 mm average position error and 0.35m/s (1.28 km/h) average velocity error on baseball pitching motion. Our work can be directly applied to estimate human poses either in indoor environment or real-world baseball field.

INDEX TERMS 3D human pose estimation, triangulation, heatmap, filtering, baseball.

I. INTRODUCTION

Human pose reconstruction has become increasingly important in domains like sports analytic [1], [2], [3], action recognition [4], human–computer interaction [5], [6], human rehabilitation [7], [8], [9], etc. 3D human pose estimation is especially important because it directly reflects the actual state of the human body. In the field of baseball, pitching is considered as one of the most significant actions in baseball games and baseball player training phases. By analyzing 3D pitching poses, couches and audiences can directly track the performance and physical state of the baseball pitcher. Therefore, 3D pose estimation has vast potential for tracking the baseball pitcher's behavior because it can directly obtain the joint positions of the baseball pitcher without extra sensors and equipment applied to human parts.

In several types of 3D pose estimation methods, multi-view 3D pose estimation is considered as the most feasible method to utilize in baseball analytic. Although recent works of pose estimation are focused on monocular 3D pose

The associate editor coordinating the review of this manuscript and approving it for publication was Lorenzo Mucchi.

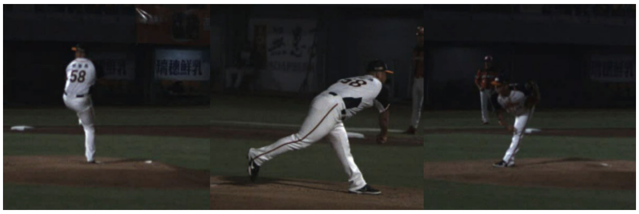


FIGURE 1. Self-occluded poses from baseball pitching.

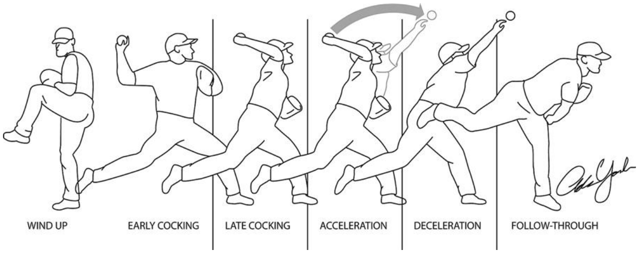


FIGURE 2. Phases of baseball pitching [14].

estimation, by comparing their performance on 3D pose datasets like Human3.6m [10], the accuracy of monocular 3D pose estimation is still low compared to multi-view 3D pose estimation [11]. Moreover, in epipolar geometry, because different 3D points in the world can correspond to the same 2D point in the image, it is impossible to get the actual 3D pose coordinates from a single camera without extra parameters like limb length. Due to such reason, most of the obtained 3D results from monocular 3D HPE models are relative coordinates with the hip centered at the origin, hindering us from directly computing the velocity of specific joints like shoulders and wrists from them. Since it is practical to set a calibrated camera system around the pitcher's circle for multi-view 3D pose estimation, in this work we focus on multi-view 3D HPE as a more suitable way to obtain the pose of the pitcher with higher accuracy.

However, there are several difficulties pertaining to pose estimation in baseball pitching. First, baseball pitching is an action with specific sports movements, which is different from normal daily motions in most of the 3D pose datasets [10], [12], [13], and there is no in-the-wild 3D baseball pose datasets yet to the best of our knowledge. Second, baseball pitching also contains lots of joints occluded by other body parts (Figure 1), which will further deteriorate the performance of the model. Third, baseball pitching contains dramatic velocity changes during arm acceleration phases. Due to these properties of pitching, its difficult to use common filters to remove random noises while preserving high-frequency critical joint movements in pitching.

In this paper, we develop a 3D baseball pose reconstruction system to solve the difficulties mentioned earlier. First, to fix detection error of pose estimation models, we present a novel joint-wise volumetric triangulation method to efficiently reconstruct pitcher's 3D pose. Unlike traditional triangulation methods which only take 2D joint predictions as input, our method can fully utilize the information of multi-view 2D joint heatmaps which indicates the probability of joint location without any extra training. More explicitly, we aggregate multi-view 2D heatmap information of joints to 3D heatmaps, which is used to calculate the final 3D pose. Second, to filter high-speed baseball pitching motion, we designed a baseball-customized filter system to remove noisy signal from pose movement while preserving the high-frequency pitching motion. More elaborately, we use filters with different parameter settings on different joints based on the characteristics of baseball pitching. Referring from Table 10, our proposed pose reconstruction scheme yields a 33.1 mm average position error and 0.35m/s (1.28 km/h) average velocity error on baseball pitching motion, which outperform other 3D reconstruction methods and filter systems (cf. Table 6,7,8,10). Our system can be operated at a reasonable computational speed of 25 fps. Our work can be directly applied to estimate human poses in various fields including indoor environment or real-world baseball field.

II. RELATED WORK

A. ML-BASED 3D HUMAN POSE ESTIMATION (HPE)

3D HPE is one of the most popular research topics in computer vision that involves estimating 3D human poses and bone orientations from 2D images or videos. Due to the complexity of human body, different HPE methods will give different human models. There are mainly three types of human models: skeleton-based model focused on human joints; skinned multi-person linear model (SMPL model) focused on human shape and body proportions, and surface-based model focused on dense correspondences of the human image.

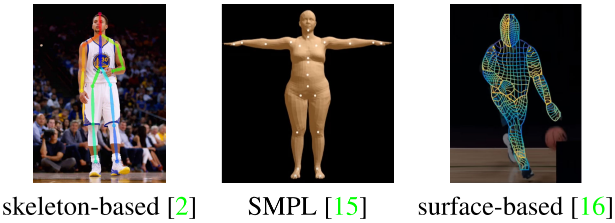

skeleton-based [2] SMPL [15] surface-based [16]

FIGURE 3. Three main kinds of human models.

With the growth of 3D human pose dataset like Human3.6m [10], MPI-INF-3DHP [13], CMU Panoptic [12], etc, machine learning (ML) based methods is now the mainstream method for 3D HPE. 3D pose estimation can be roughly divided into two categories based on the number of cameras: multi-view 3D pose estimation and monocular 3D pose estimation. With multiple cameras of different perspectives that can reduce the ambiguity of depth information, multi-view 3D HPE generally has better accuracy than monocular 3D HPE. Typical multi-view 3D HPE methods include multi-view 2D heatmaps fusing with models [17], [18], triangulation [19], [20], [21], and multiple view consistency [22]. Pavlakos et al. [17] uses 3D pictorial models to take 2D image features as input and return the final 3D human pose. Tome et al. [18] designed a multi-stage approach that refine the 3D estimation generated at each stage. Iskakov et al. [19] uses triangulation to aggregate the 2D image features for further training. Zhang et al. [20] presented adaptive fusion weight to reflect the 2D feature quality from each view and use it to reduce the responses at incorrect joint locations before 3D triangulation to get better results. Kocabas et al. [21] use the epipolar geometry to recover the 3D pose from predicted 2D poses and uses it as a supervision signal to train the 3D HPE model. Rhodin et al. [22] proposed a semi-supervised method by trying to make the 2D HPE models from each view to predict the same 3D pose. There are further some researches [23], [24] that combine human localization module and feature volume to estimate multiple people with multiple camera views. With labeled 3D pose datasets, these ML-based multi-view 3D HPE models can fully utilize the information of 2D features or heatmaps and perform well on the datasets they trained on, but they often require extra training on labeled datasets to maintain the accuracy of the pose estimation model. Most of the 3D human pose datasets [10], [25], [26] are collected in indoor scenarios. For most sports, despite some small datasets like KTH Multiview Football Dataset [27], there is often not enough in-the-wild multi-view 3D human pose datasets for ML-based 3D HPE to be properly trained. Figure 4 shows that, with 3D-based HPE models trained on indoor Human3.6m dataset, the reconstruction results for pitcher videos taken in-the-wild is far from satisfactory. Figure 4 also shows that our system yields more rational reconstruction results compared with other 3D-based HPE models (here the 2D pose heatmaps are all extracted from Alphapose for fair comparison).

Monocular 3D HPE recovers 3D human poses with single-view images. The advantage of monocular 3D HPE is the low hardware requirement and no need of the calibrated camera system, but the pose ambiguity of the 3D pose projection becomes a real problem, and because different 3D points in the world can correspond to the same 2D point in the image, most of the monocular methods can only obtain relative coordinates of human poses that do not give the actual location of humans in the environment. Some researches [31], [32], [33], [34] claim that they can recover the absolute coordinates of human poses, but most of them require extra parameters like limb length. Their performance is also bad compared with multiview 3D HPE [19], [20]. Methods of monocular 3D HPE can be categorized into several classes. Some methods [17], [35] directly predict a volumetric heatmap of joint location and take the maximum of the heatmap as final estimation. References [36], [37], [38], and [39] treat HPE as a regression problem that estimate the location of joints relative to the root joint. Many researches [40], [41], [42], [43], [44] use 2D human pose predictions as input and lift them to 3D spaces. References [45], and [46] take a sequence of images as input and try to solve the ambiguity problem of monocular 3D HPE by considering the temporal consistency of the pose series predicted by the model.

In several papers [17], [19], [21], [23], [24], [47], volumetric heatmap which indicates the probability of the 3D joint in the space is utilized to generate 3D human poses. They define a volumetric heatmap with a range that covers the entire body, then uses ML models or projection geometry to aggregate the 2D joint heatmaps generated by the 2D HPE or pose features to the volumetric heatmap. After heatmap aggregation, 3D CNN and softmax function [48] is utilized to generate the final human pose predictions. Compared to ML based 3D HPE, where a large venue of volumetric heatmap is needed to aggregate all the information around the whole body for further training, our method considers only the space around each joint, thereby achieving faster computation speed.

B. MULTI-VIEW TRIANGULATION

Multi-view triangulation is an important topic in computer vision that involves estimating a 3D point in the world with its 2D projection points from multi-view images given. There are roughly three traditional multi-view triangulation methods [49], [50]: Direct linear transform (DLT) method that expresses the relationship between 3D prediction and 2D projections as linear equations and solves them; Midpoint method that minimizes sum of the distances from the 3D prediction to all the epipolar lines generated by 2D projections; Reprojection optimization method including L_2 method and L_∞ method that tries to find the point which has the minimum reprojection error between 2D image points and 2D points reprojected from 3D prediction. These traditional methods are considered to be fast and accurate enough that recent related studies [51], [52], [53], [54], [55] still use them to reconstruct 3D points, but these methods are quite sensitive to outliers, so recently there are some new triangulation methods [56], [57] that consider the robustness of the triangulation algorithm to outliers.

All of these triangulation methods above takes 2D points as input to reconstruct 3D joints. In 3D human pose reconstruction, 2D HPE is often used to generate 2D human joints. These joint predictions are actually generated by joint heatmaps that indicates the probability of the joint position. If we use these triangulation methods which only takes 2D joints as input, the rich information provided by the joint heatmap will be wasted. In our work, we designed a triangulation method that can fully utilize the information of these 2D joint heatmaps efficiently.

C. HUMAN POSE FILTERING

After receiving 3D human poses from 3D HPE or motion capture system like Vicon, due to the inaccuracy of the observation, the estimated 3D joint trajectory often contains noisy signal. Low-pass filters are often used for human pose filtering. These methods refine human motion data

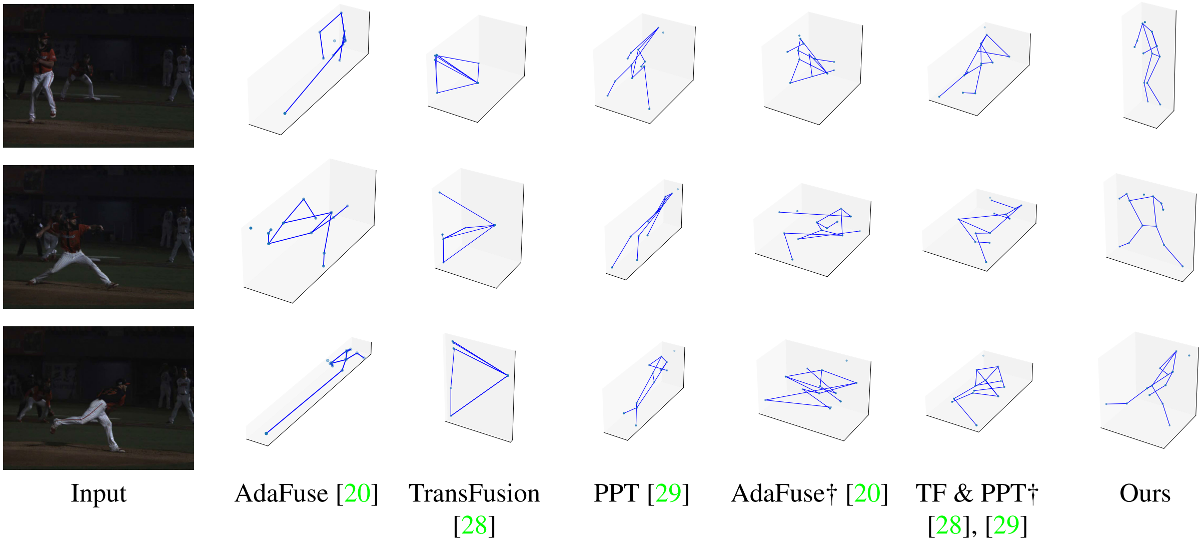


FIGURE 4. Reconstruction results of multi-view 3DHPE methods on Tainan baseball dataset (cf. Section IV-C). 3DHPE models are selected from the state-of-the-art multi-view 3DHPE on Human3.6m [10]. Reconstruction results with dagger (†) indicate that 2D pose heatmaps in 3D HPE are extracted from alphapose [30].

with parameters independent of input. There are several related researches [58], [59] utilizing classic low-pass filters. For example, Mathys et al. [60] filter human motions with Gaussian filter [61], which modifies signal with a Gaussian distribution as its kernel. As Butterworth filter [62] is a type of signal processing filter that has stable and flat frequency response in its passband, Crenna et al. [63] uses Butterworth filter to filter out white Gaussian noise in human motions. One Euro filter [64] is a first order low-pass filter with an adaptive cutoff frequency, which reduces jitter while preserving the high frequency motion signals. Zou et al. [65] thereby utilize one Euro filter to reduce high-frequency noises while reconstructing human dynamics. By using position, velocity, or angular velocity of human joints to represent the system state, Kalman filter [66] and its variants [67], [68] are often used to refine human motion sequences [69], [70], [71], [72]. These common filters work well on general human motions, but when filtering high speed motions like pitching, some parameter or structure adjustment must be done to maintain the performance of the filter.

Recently, due to the popularity of machine learning, some human pose filter based on temporal models [73], [74], [75] have been proposed. However, these ML based filters often require large annotated human pose datasets like Human3.6m [10] or 3DPW [76]. This is often impractical in sports related researches, where it is difficult to collect and annotate enough in-the-wild data to train these ML based filters.

D. SPORTS ANALYSIS USING POSE ESTIMATION

By tracking human poses without extra sensors and equipment, pose estimation facilitate performance monitoring in sports analytic area. For easier implementation and faster computation speed compared with 3D pose estimation, 2D pose estimation is utilized in many sports-related researches. Li et al. [1] performed a baseball evaluation system using 2D pose estimation model OpenPose [2] to estimate whether a baseball hitter performs a good swing. Jiang et al. [77] proposed a lightweight temporal-based 2D HPE designed for efficient and effective golf swing analysis. Kurose et al. [78] performed method for objective form analysis that can evaluate the quality of play performed by the tennis player using observable information such as estimated joint position of the player and the result of the game. Yan et al. [79] presented an automatical clipping system to summarize the sports video stream using 2D poses predicted by 2D pose estimation models as input. To solve the poor detection issue of certain sports movement, instead of using pretrained 2D HPE models, several researches [77], [80], [81] use their own dataset of specific sports to train their 2D HPE models. Due to the variety of sports and the fact that these sports-related researches are often trained on self-collected data, it is difficult to have a common standard to compare these methods.

In related researches using 3D pose estimation, 3D pose skeletons can provide more information about human than 2D pose skeletons such as joint angle measurement and speed analysis. However, 3D HPE is also more challenging because most of the 3D HPE models require calibrated camera system and more complicated models and algorithms than 2D HPE models. However, as illustrated in Figure 4, directly using multi-view 3D HPE on in-the-wild scenarios often leads to poor performance. Because of the data variety of the 2D pose datasets [82], [83] and the lack of the related multi-view in-the-wild 3D human pose datasets, 2D HPE models often perform better than multi-view 3D HPE models in outdoor scenarios. Due to this reason, instead of directly

TABLE 1. Comparison between 3D HPE methods.

3D HPE methods	Advantages	Disadvantages
ML-based Multiview 3D HPE	• High accuracy.	• Low Robustness. • Requires related dataset.
ML-based Monocular 3D HPE	• Low hardware requirement.	• Less accuracy • Pose ambiguity.
ML-based 2D HPE + Multi-view triangulation	• High accuracy. • High robustness.	• Needs calibrated camera system.

FIGURE 5. In the wild performance of 2D HPE [30].

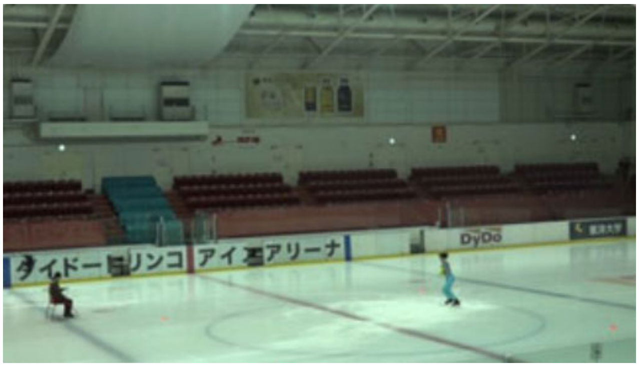


FIGURE 6. Human pose capturing in large skating venue [3].

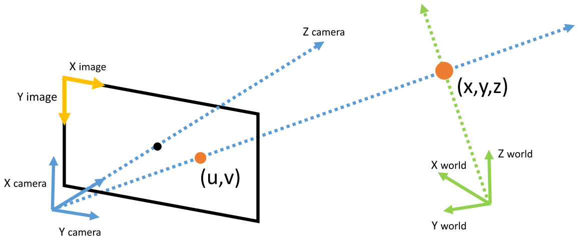


FIGURE 7. Pinhole camera model, where (x, y, z) represents a 3D point in the world and (u, v) is its 2D image projection.

using 3D HPE models [19], [20], [29] trained on 3D human pose datasets, most of the studies [3], [53], [54], [55], [84] use two-step methods where 2D HPE models [2], [30] are combined with 3D triangulation algorithms that aggregate multi-view 2D poses to a 3D pose. The comparison between 3D HPE methods are presented in Table 1. There are lots of practical difficulties while reconstructing 3D poses in real-world environment. For example, to solve the problem of unclear human pose images captured in large skating venue, Tian et al. [3] performed a transformation system and a time smoothness system attempting to effectively handle the performance of the multi-view 3D pose estimation. To solve the problem of noisy joint estimations on sports field and long processing time of multiple people 3D HPE, Bridgeman et al. [85] presents a new multi people 3D HPE method that applies a fast greedy algorithm to efficiently identify correspondences between 2D poses in multi-view videos. Their approach also uses a temporal skeleton association and filtering method to correct errors of the estimated poses. There are also some researches [86], [87] using monocular 3D pose estimation to predict relative 3D human poses for further analysis.

III. METHOD

In this section, we elaborate on our proposed methods of joint-wise volumetric triangulation and 3D baseball pose reconstruction system. Section III-A introduce basic algorithms that will be used in our triangulation method including camera geometry and two-view DLT triangulation. In Section III-B, our 3D pose reconstruction method using joint-wise volumetric triangulation would be described. In Section III-C, we introduce our baseball customized filter system that can filter out noises while the high frequency signal of the pitching motion is preserved.

A. CAMERA GEOMETRY AND TWO-VIEW DLT TRIANGULATION

Refer to [49], by assuming the relation between the camera and the world as the ideal pinhole camera model in Figure 7, the projection from a 3D point (x, y, z) to the 2D image point (u, v) can be formulated with:

$$Z_i \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} [R|T] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = M \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \quad (1)$$

Here $R \in \mathbb{R}^{3*3}$ represents the rotation matrix and $T \in \mathbb{R}^{3*1}$ represents the translation matrix from world coordinate to camera coordinate, f_x, f_y indicate the focal length, and u_0, v_0 indicate the image coordinate of the lens center from the camera. All parameters above can be obtained by camera calibration, which sets several reference points in the world to calculate the relations between world coordinate and image coordinate. In other words, with calibrated camera, we can represent the relation of 3D point in the world and its 2D projection on the image though the camera matrix M. Here Z_i represents the distance of X to the focal plane in camera coordinate system.

Direct linear transform (DLT) triangulation method aims to find the position of 3D point X with its 2D projections from multi-view cameras. More elaborately, consider two cameras with their camera matrices M_1, M_2 and their corresponding 2D image projections (u_1, v_1), (u_2, v_2). In each view c, we can represent the relation between $x_c = [u_c, v_c, 1]^T$ and X as $k_c x_c = M_c X$, where k_c is the unknown scale factor. Since x_c and $M_c X$ are two parallel vectors, their cross product $x_c \times M_c X$ vanishes. We can list three equations from $x_c \times M_c X = 0$, where m_c^i means the ith row of M_c:

$$u_c(m_c^3 X) - (m_c^1 X) = 0, \tag{2}$$

$$v_c(m_c^3 X) - (m_c^2 X) = 0, \tag{3}$$

$$u_c(m_c^2 X) - v_c(m_c^1 X) = 0, \tag{4}$$

where (4) is redundant as it is a linear combination of (2) and (3). Therefore, with two views $c = 1, 2$, there will be essentially four equations to solve X as follows:

$$\begin{bmatrix} u_1(m_1^3) - (m_1^1) \\ v_1(m_1^3) - (m_1^2) \\ u_2(m_2^3) - (m_2^1) \\ v_2(m_2^3) - (m_2^2) \end{bmatrix} X = 0. \tag{5}$$

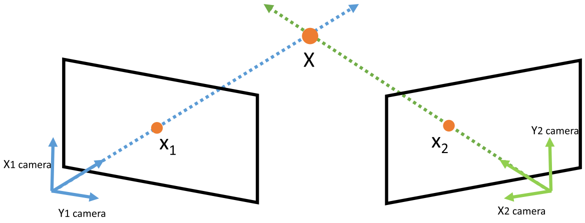


FIGURE 8. Two-view triangulation.

B. JOINT-WISE VOLUMETRIC TRIANGULATION

The main disadvantage of the traditional triangulation methods is that they only use 2D joint predictions to reconstruct 3D human poses, which ignores the information of 2D joint heatmaps from 2D pose estimation model. Our joint-wise volumetric triangulation is designed to efficiently utilize the information of these 2D joint heatmaps.

The pseudo code of our joint-wise volumetric triangulation are presented in Algorithm 1. Our method assumes to have videos from C synchronized cameras capturing human motions. Each camera is calibrated to get its projection matrix M_c. For each frame c, we have 2D joint heatmap information H_c^{2D} from the prediction of 2D pose estimation model, and the main goal is to reconstruct the 3D human pose $J_i^{3D} \in \mathbb{R}^3$ from a 3D volumetric heatmap $H_i^{3D} \in \mathbb{R}^{n*n*n}$ that aggregates the information from these 2D heatmaps for each joint i.

TABLE 2. Notation.

Symbols	Meaning
M_c (Alg. 1)	Camera Matrix of camera c
$H^{2D}_{c,i}$ (Alg. 1)	2D joint heatmap of camera c (joint i)
J^{2D}_i (Alg. 2)	2D human pose generated by $H^{2D}_{c,i}$ (joint i)
$V^{3D}_{i,p,q,r}$ (Alg. 1)	3D coordinate of voxels in volumetric heatmaps (joint i , index=(p,q,r))
H^{3D}_i (Alg. 1)	volumetric heatmap value (joint i)
J^{3D}_i (Alg. 1)	3D human pose (joint i)
$M_{i,t}$ (Alg. 3)	Total mean (joint i, time t centered in the interval)
$SD_{i,t}$ (Alg. 3)	Standard deviation (joint i, time t centered in the interval)
$O_{i,t}$ (Alg. 3)	Outlier indication (joint i, time t)

Algorithm 1 Joint-Wise Volumetric Triangulation

```
Require: 2D joint heatmaps $H^{2D}_{c,i}$ of $I$ joints in $C$ camera views, Camera matrices $M_c$ in $C$ camera views
Ensure: 3D human pose $J^{3D}_i$
$J^{3D}_{raw,i} \leftarrow$ Simple triangulation($H^{2D}_{c,i}, M_c$)
for $i \leftarrow 1$ to $I$ do
    $V^{3D}_i \leftarrow$ Mesh grids around $J^{3D}_{raw,i}$
    $H^{3D}_i \leftarrow \sum_{c=1}^{C} H^{2D}_{c,i}(M_c V^{3D}_{c,i})$
    $\hat{J}^{3D}_i \leftarrow$ 3D point that has the greatest heatmap value from $H^{3D}_i$.
    $J^{3D}_i = \sum_{k \in N_i} \frac{e^{\alpha * H^{3D}_{i,k}}}{\sum_{l \in N_i} e^{\alpha * H^{3D}_{i,l}}} V^{3D}_{i,k}$
    where $N_i = \{(p,q,r) : \|V^{3D}_{i,p,q,r} - V^{3D}_{i,p^*_i,q^*_i,r^*_i}\|_2 < d\}$
end for
```

Most human joint heatmap aggregation methods [17], [19], [21], [23], [24], [47] take human pelvis as the center of the heatmap and consider the large space all around the body. However, a clever placement of the domain for the volumetric heatmap near the real position of the joint with a smaller side length D can increase accuracy while using less mesh grids to achieve better computing speed. Towards this end, as illustrated in Figure 10 and algorithm 2, we use a simple triangulation method to reconstruct a preliminary 3D human pose. We first find 2D poses by finding arguments of the maxima (argmax) of the 2D joint heatmaps. After that, two-view DLT triangulation $Tri(J^{2D}_{j,i}, J^{2D}_{k,i}, M_j, M_k)$ mentioned in Section III-A is utilized to get the 3D joint reconstruction results for joint i from every two views j, k. We compare their reprojection errors and choose the 3D joint $J^{3D}_{raw,i} \in \mathbb{R}^3$ with the lowest error as the preliminary 3D pose. After finding J^{3D}_{raw}, we place the volumetric heatmap H^{3D}_i of joint i around the position $J^{3D}_{raw,i}$, then perform discretization

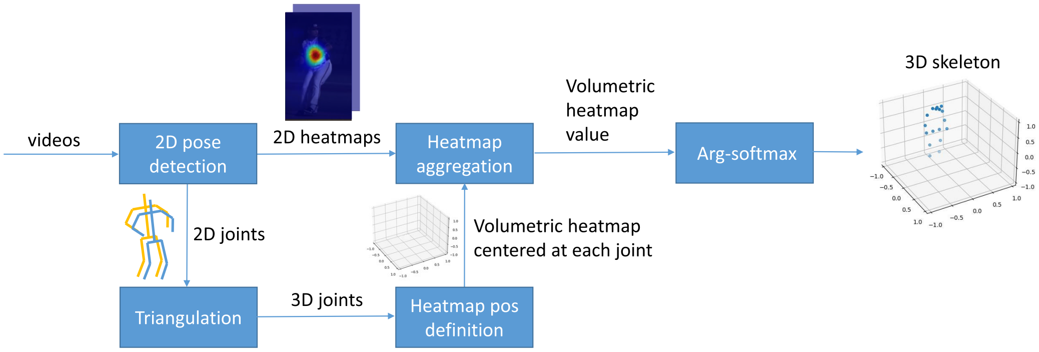


FIGURE 9. Flow chart of joint-wise volumetric triangulation.

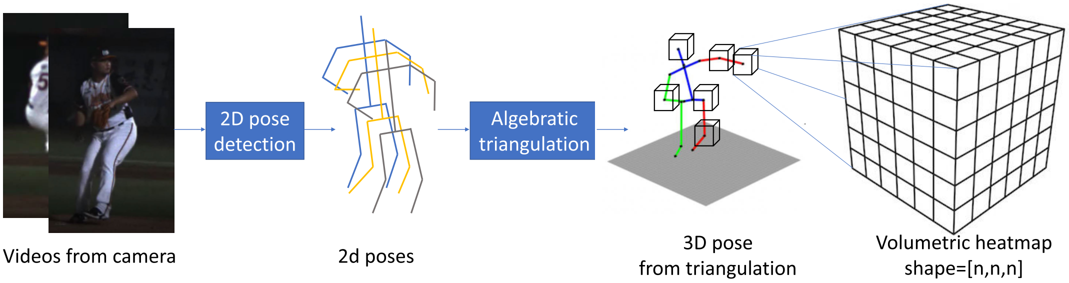


FIGURE 10. Position definition of joint-wise volumetric heatmap.

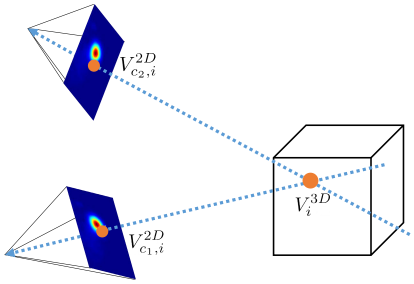


FIGURE 11. Aggregate information of 2D heatmaps into volumetric heatmap.

from every axis to get $V^{3D} \in \mathbb{R}^{I*n*n*n*3}$, which indicates the 3D coordinate that corresponds to each voxel in the volumetric heatmaps, namely $(V^{3D}_{i,p,q,r,1}, V^{3D}_{i,p,q,r,2}, V^{3D}_{i,p,q,r,3})$ indicates the 3D coordinate of the (p, q, r)-th voxel in H_i^{3D}.

After determining the coordinates in the volumetric heatmaps, we then aggregate 2D heatmap values from every camera to the volumetric heatmap. Projective geometry is utilized to project the 3D coordinates in V^{3D} to each camera view c. Refer to section III-A, the projection from a 3D point to the image point can be formulated as matrix multiplication $k_c \left[V^{2D}_{c,i,p,q,r,\cdot}, 1\right]^T = M_c V^{3D}_{i,p,q,r,\cdot}$, where k_c can be eliminated through normalization. The volumetric heatmap is then calculated as the summation of the projected 2D heatmap values from every camera view:

$$H^{3D}_{i,p,q,r} = \sum_{c=1}^{C} H^{2D}_{c,i}(V^{2D}_{c,i,p,q,r,\cdot}). \tag{6}$$

Here the right hand side in (6) is computed with pixels in $H^{2D}_{c,i}$ through nearest neighbor interpolation. We pick the point that has the greatest heatmap value $\hat{J}_i^{3D} = V^{3D}_{i,p_i^*,q_i^*,r_i^*}$, where

$$(p_i^*, q_i^*, r_i^*) = \arg\max_{p,q,r} H^{3D}_{i,p,q,r}, \tag{7}$$

as the preliminary estimation of J_i^{3D}. We then estimate J_i^{3D} as the centroid of mesh grids in H_i^{3D} that are at most distance d away from $\hat{J}_i^{3D}$, where each mesh grid is weighted by taking softmax [48] over H_i^{3D}, namely:

$$J_i^{3D} = \sum_{k \in N_i} \frac{e^{\alpha * H^{3D}_{i,k}}}{\sum_{l \in N_i} e^{\alpha * H^{3D}_{i,l}}} V^{3D}_{i,k}, \tag{8}$$

where $N_i = \{(p, q, r) : \|V^{3D}_{i,p,q,r} - V^{3D}_{i,p_i^*,q_i^*,r_i^*}\|_2 < d\}$.

C. FILTER SYSTEM CUSTOMIZED FOR BASEBALL PITCHING

Baseball pitching is a motion that contains dramatic velocity changes during arm acceleration phases. With off-the-shelf low-pass filter that is often used to filter human poses, it is likely to cause delay or loss in the signals from the pitching arm because of the high frequency movement. As there is

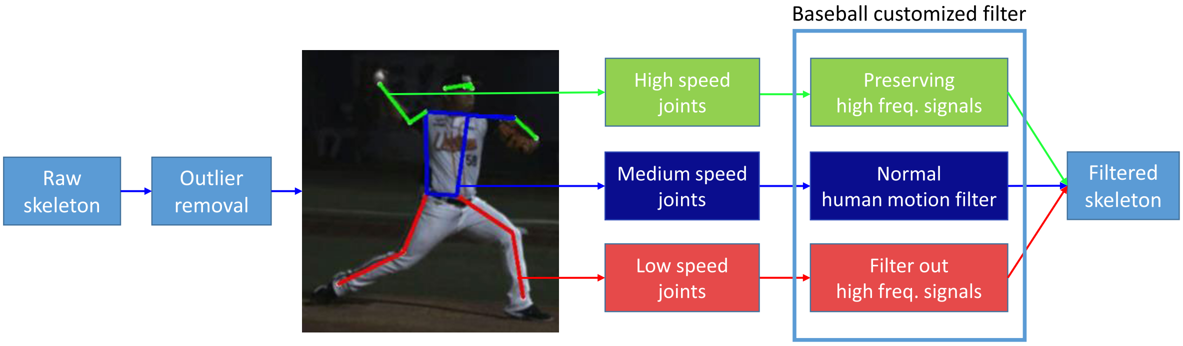

Baseball customized filter
Raw skeleton
Outlier removal
High speed joints
Medium speed joints
Low speed joints
Preserving high freq. signals
Normal human motion filter
Filter out high freq. signals
Filtered skeleton

FIGURE 12. Flowchart of baseball customized filter system.

Algorithm 2 Simple Triangulation

```
Require: 2D joint heatmaps $H^{2D}_{c,i}$ of $I$ joints in $C$ camera views, Camera matrices $M_c$ in $C$ camera views
Ensure: a raw 3D human pose $J^{3D}_{raw,i}$
for $i \leftarrow 1$ to $I$ do
    for $j \leftarrow 1$ to $C$ do
        $J^{2D}_{c,i} \leftarrow argmax(H^{2D}_{c,i})$
    end for
end for
for $i \leftarrow 1$ to $I$ do
    $L \leftarrow \infty$
    for $j \leftarrow 1$ to $C - 1$ do
        for $k \leftarrow j + 1$ to $C$ do
            $\hat{J}^{3D}_i \leftarrow Tri(J^{2D}_{j,i}, J^{2D}_{k,i}, M_j, M_k)$
            $\hat{L} \leftarrow 0$
            for $l \leftarrow 1$ to $C$ do
                $\hat{L} \leftarrow \hat{L} + \|J^{2D}_{l,i} - M_l\hat{J}^{3D}_i\|_2$
            end for
            if $\hat{L} < L$ then
                $L \leftarrow \hat{L}$
                $J^{3D}_{raw,i} \leftarrow \hat{J}^{3D}_i$
            end if
        end for
    end for
end for
```

not enough 3D labeled pitching pose data for us to train pose smoothing spatio-temporal models, a customized filter system designed for pitching motion is a must. Figure 12 shows the flowchart of our baseball customized filter system. It contains two modules: outlier removal and a filter system based on traditional low-pass filters that will filter noises at different levels depending on the action characteristics of the joints during pitching, which will be elaborated in Section III-E.

D. OUTLIER REMOVAL

After 3D human pose reconstruction with triangulation, there are still many outliers in the reconstructed 3D pose signal. We designed a simple but effective algorithm to detect the outliers. For every time frame t, we compute the mean and standard deviation of the pose data close to it. The outliers can be detected if the distance between the joints at frame t and the mean are larger than a margin of error. Algorithm 3 shows the detail of the outlier detection algorithm, where $M_{i,t}$, $SD_{i,t}$ indicate the mean and standard deviation of the i^{th} joint around frame t:

Algorithm 3 Outlier Detection

```
Require: 3D poses containing $I$ joints from a time clip of total $T$ frames $\left[J^{3D}_{i,t}\right]$, window size $2W + 1$, outlier detection constant $\beta$, number of iterations $N$.
Ensure: An $I * T$ array $\left[O_{i,t}\right]$ indicating which joints in which frames are outliers.
for $n \leftarrow 1$ to $N$ do
    for $t \leftarrow 1$ to $T$ do
        for $i \leftarrow 1$ to $I$ do
            $M_{i,t} \leftarrow \Sigma^{t+W}_{s=t-W} J^{3D}_{i,s}/(2W + 1)$
            $SD_{i,t} \leftarrow sqrt(\Sigma^{t+W}_{s=t-W}(J^{3D}_{i,s} - M_{i,t})^2/(2W + 1))$
            if $\|J^{3D}_{i,t} - M_{i,t}\|_2 > \beta * SD_{i,t}$ then
                $O_{i,t} \leftarrow$ True
            else
                $O_{i,t} \leftarrow$ False
            end if
        end for
    end for
end for
```

After outlier detection, considering speed and signal smoothing, we simply replace the outliers with the linear interpolation of adjacent joints. The process of outlier detection and replacement will be iterated for $N = 4$ times.

E. FILTER SYSTEM DEPENDING ON CHARACTERISTICS OF PITCHING

We designed a filter system to remove the noise from the pitching movement without eliminate the pitching signals. The backbone of the filter system includes a Median filter [88] to wipe out outliers, a Gaussian filter [61] to filter out Gaussian noises from the pose movement, and a fourth Butterworth filter [62] to further decrease high-frequency

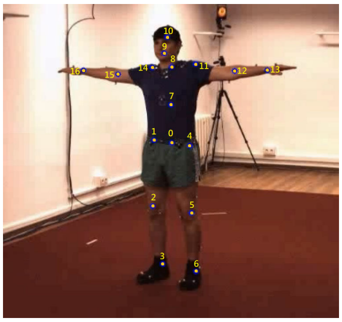


FIGURE 13. Human3.6M annotation format.

noises. Depending on the characteristics of pitching action, we separated the movement of body parts in the human pose into three types: fast movement, medium movement, and slow movement. For a right handed pitcher, Table 3 shows how we categorize the body parts:

TABLE 3. Types of human body movements.

Body parts	Type of movement	Body parts	Type of movement
Left wrist	Fast	Right wrist	Fast
Left elbow	Medium	Right elbow	Fast
Left shoulder	Medium	Right shoulder	Medium
Left hip	Slow	Right hip	Slow
Left knee	Slow	Right knee	Slow
Left ankle	Slow	Right ankle	Slow

For joints that have fast movement like the pitching arm, the movement contains lots of high frequency signals, so we use Median filter with smaller window length l^{fast}, Gaussian filter kernel with smaller standard deviation σ^{fast}, and Butterworth low-pass filter with higher critical frequency w^{fast}, to preserve the high frequency signals of the joint movement. For joints that have slow movement like the lower body, we use Median filter with larger window length l^{slow}, Gaussian filter kernel with bigger standard deviation σ^{slow}, and Butterworth low-pass filter with lower critical frequency w^{slow}, to filter out high frequency noises as much as we can. Detailed parameter settings are presented in Table 9.

IV. DATASETS AND EVALUATION METRICS

A. HUMAN3.6M DATASET

Human3.6m [10] is a widely used dataset in 3D pose estimation. It contains 3.6 million images from 11 subjects performing actions in 17 daily scenarios: discussion, taking on the phone, walking dogs, etc. Each action video is captured by four synchronized cameras at 50 Hz, and a high-speed motion capture VICON system is utilized to capture the 3D joint positions from subjects. Figure 13 shows the annotation format of Human3.6m.

B. MSL BASEBALL DATASET

[1]MSL Baseball Dataset is a baseball player pose dataset recorded in Movement Science Lab (MSL) by National Taiwan University. The dataset contains 15 pose sequences composed of 79200 images of a subject doing three different baseball actions: hitting, pitching, punching. Each action video is captured by four synchronized cameras at 300 fps. A fixed VICON system with 10 cameras operating at 120 fps is utilized to capture human 3D joint position from markers attached to the subjects. The camera setting as well as annotation format of the MSL Baseball Dataset are shown in Figure 14 and Figure 15.

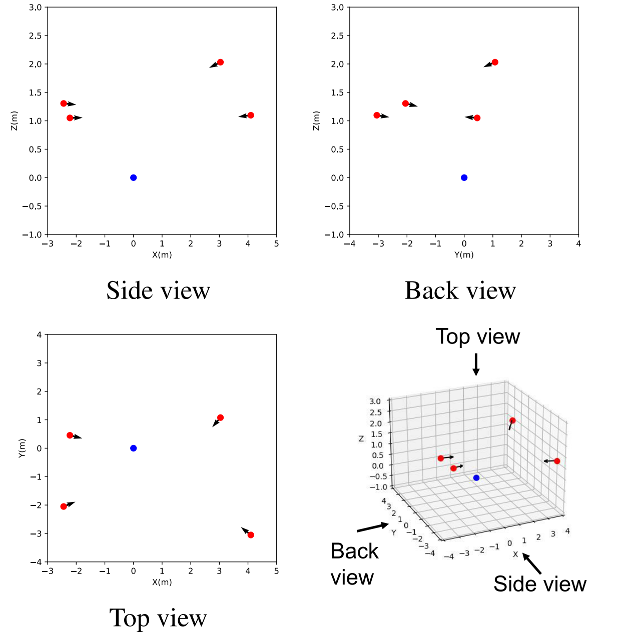


FIGURE 14. Camera setting of MSL baseball dataset. The red dots indicate camera position and the black arrows indicate camera orientation.

C. TAINAN BASEBALL DATASET

Tainan Baseball Dataset is a baseball pitching pose dataset recorded in Tainan Municipal Baseball Stadium by National Taiwan University of Sports. The dataset contains 27 pitching pose sequences composed of 32805 images from different pitchers. Due to the difficulty of real-world marking and capturing, there is no joint label in the dataset. Our multi-view system consists of three synchronized cameras at 300 fps to capture videos from pitchers. Figure 16 shows the angles of the pitchers photographed by each camera.

D. EVALUATION METRICS

We use two of the most widely used evaluation metrics in pose estimation: mean per joint position error (MPJPE) and Procrustes aligned mean per joint position error (P-MPJPE) [11]. MPJPE is calculated from the mean of Euclidean distance between joint predictions J and the ground truth joints J_{gt}:

$$MPJPE(J) = \frac{1}{T * I} \sum_{t=1}^{T} \sum_{i=1}^{I} \|J_{i,t} - J_{gt,i,t}\|_2 \tag{9}$$

[1] This project is reviewed and approved by Jen-Ai Hospital Institutional Review Board. Both MSL Baseball Dataset and Tainan Baseball Dataset have been obtained with the informed consent of involved subjects.

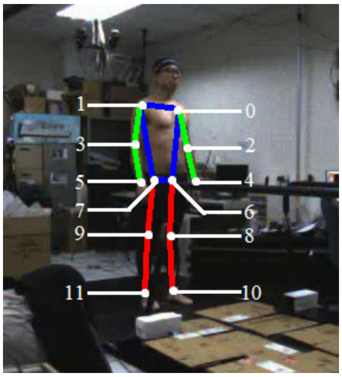


FIGURE 15. MSL baseball dataset annotation format.

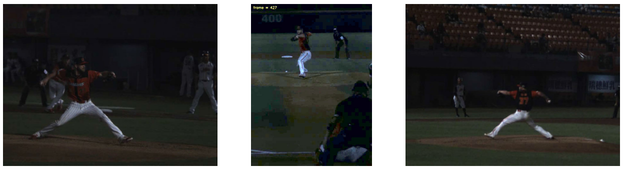


FIGURE 16. Camera setting of Tainan baseball dataset.

where $J_{i,t}$ denotes the i^{th} joint position at frame t. T is the total frame number. I is the joint number in a single pose. P-MPJPE is a variant of MPJPE. It is similar to MPJPE since it measures the Euclidean distance between joint predictions and the ground truth joints, but before calculating, an additional alignment step is applied to the prediction. The prediction was shifted, scaled, and rotated trying to minimize the error:

$$P-MPJPE(J)=\frac{1}{T*I}\sum_{t=1}^{T}\sum_{i=1}^{I}\|P(J_{i,t},J_{gt,i,t})-J_{gt,i,t}\|_2 \tag{10}$$

where $P(J_{i,t},J_{gt,i,t})$ denotes Procrustes superimposition [89] to make $J_{i,t}$ similar to $J_{gt,i,t}$ by minimizing the Procrustes distances between them. These two evaluation matrices are used in most of the 3D HPE studies [11].

Similar to MPJPE, mean per joint velocity error (MPJVE) and mean per joint acceleration error (MPJAE) can be calculated to evaluate the performance of joint velocity and acceleration:

$$MPJVE(J)=\frac{1}{T*I}\sum_{t=1}^{T}\sum_{i=1}^{I}\|V_{i,t}-V_{gt,i,t}\|_2 \tag{11}$$

$$MPJAE(J)=\frac{1}{T*I}\sum_{t=1}^{T}\sum_{i=1}^{I}\|A_{i,t}-A_{gt,i,t}\|_2 \tag{12}$$

where we calculate the velocity of from joint positions and acceleration from velocity:

$$V_{i,t}=J_{i,t}-J_{i,t-1}, A_{i,t}=V_{i,t}-V_{i,t-1} \tag{13}$$

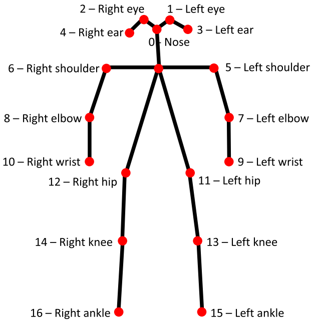


FIGURE 17. Alphapose annotation format.

V. EXPERIMENT

A. JOINT-WISE VOLUMETRIC TRIANGULATION

1) EXPERIMENT SETTINGS

Figure 18 shows the framework of the experiment on triangulation methods. To test the robustness of the triangulation methods from 2D pose predictions, we use the simplest model in Alphapose (Simple Baseline) [30] as our 2D detection model to generate 2D pose estimations. After triangulation, outlier removal algorithm mentioned in Section III-D are applied after every triangulation method to reduce the error from the outliers. Parameter settings of the joint-wise volumetric triangulation and outlier removal are shown in Table 5. We tested parameter settings on other subjects in Human3.6m and picked the setting with a reasonable computational speed (25 fps). MPJPE and P-MPJPE are used to evaluate the error of the pose. All of the algorithms are implemented on Intel Core i9-13900K CPU@5.80 GHz. Alphapose model are executed on RTX 4090.

We evaluate the performance of triangulation methods on the test set of Human3.6m (S11) and pitching set in MSL Baseball Dataset. Because of the difference in joint annotation between Alphapose (Figure 17), Human3.6m (Figure 13) and MSL Baseball Dataset (Figure 15), we only consider joints in Table 4 that are labeled in all three datasets.

TABLE 4. Joint matching between alphapose and Human3.6m.

Bodyparts	Alphapose	Human3.6m	MSL Baseball
shoulder	5,6	11,14	0,1
elbow	7,8	12,15	2,3
wrist	9,10	13,16	4,5
hip	11,12	4,1	6,7
knee	13,14	5,2	8,9
ankle	15,16	6,3	10,11

2) EXPERIMENT RESULT

We compare our joint-wise volumetric triangulation method with several triangulation methods including L2 triangulation and midpoint triangulation [49], two robust triangulation methods for 3D object reconstruction [56], [57] that can tolerate erroneous 2D joint predictions, and a 2D heatmap

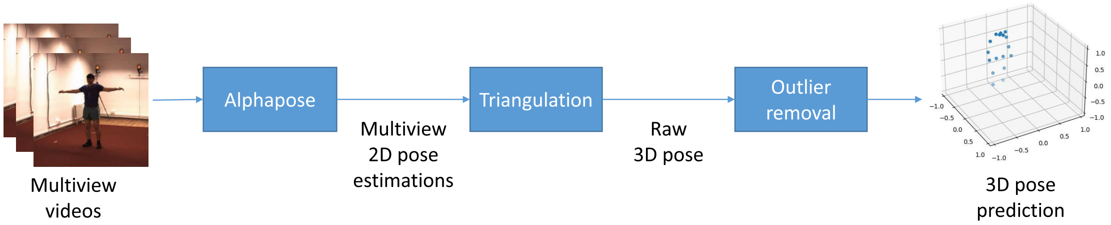


FIGURE 18. Flow chart for the experiment on joint-wise volumetric triangulation.

TABLE 5. Parameter setting.

Parameters	Value	Meaning
Joint-wise volumetric triangulation		
D (Sec. III-B)	1 (m)	Side length of volumetric heatmap
n (Sec. III-B)	32	Size of the volumetric heatmap ($n * n * n$)
d (Eq. 8)	0.2 (m)	Boundary of area definition for softmax function
α (Eq. 8)	2	Parameter in softmax function
Outlier removal		
N (Alg. 3)	4	Iteration of outlier detection
W (Alg. 3)	30	Window size of outlier detection
β (Alg. 3)	1.5	Boundary of outlier detection

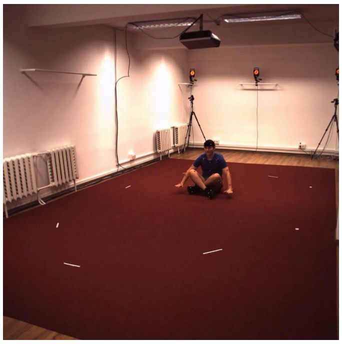
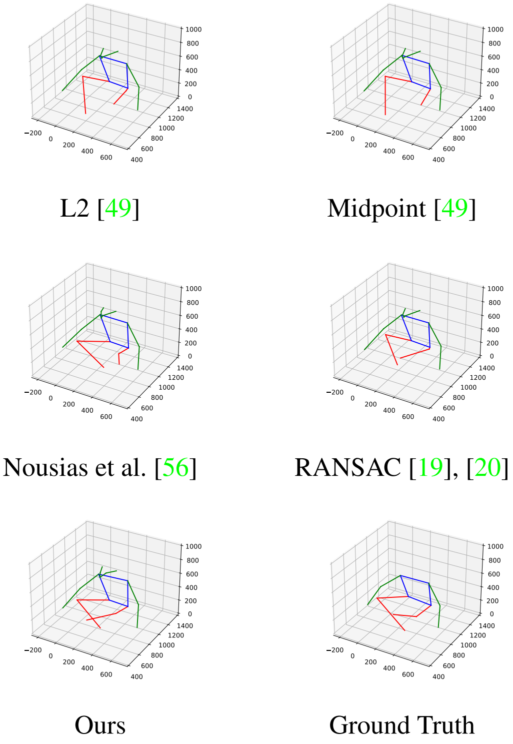


FIGURE 19. Example of sitting in Human3.6m.



L2 [49]

Midpoint [49]

Nousias et al. [56]

RANSAC [19], [20]

Ours

Ground Truth

FIGURE 20. Visualization of predictions of Figure 19.

triangulation method from the state-of-the-arts [19], [20] multi-view 3D human pose estimation on human3.6m that uses softmax function to generate 2D joints from 2D heatmaps followed with RANSAC [90] algorithm to reconstruct final 3D joint prediction.

Table 6,7,8 presents the experiment results on Human3.6m and MSL Baseball Dataset. Our joint-wise volumetric triangulation outperforms all of the other triangulation methods on every kind of motion in terms of P-MPJPE by a large margin. Although [56], [57] attain lower MPJPE on few kinds of motion, our method has lower average MPJPE and is more robust to heavily occluded poses like "Sit" in Human3.6m. As an example illustrated in Figure 20, Figure 21 and Figure 22, the accuracy of traditional triangulation methods are influenced by wrong 2D pose predictions in camera 1 and camera 3 as they value wrong predictions as much as right predictions. Our triangulation method can consider extra information on 2D pose heatmaps and reduce the impact of wrong 2D predictions which has less response on volumetric heatmaps, so it can still correctly reconstruct 3D human poses in the case of 2D joint prediction error. Our triangulation algorithm operates at 25 fps, though slower than other triangulation methods (>40 fps), but still meets the need of most real time applications.

B. BASEBALL CUSTOMIZED FILTER SYSTEM

1) EXPERIMENT SETTINGS

To evaluate the performance of the filter system on pitching motions, we use the pitching set of MSL Baseball Dataset and Tainan Baseball Dataset for testing. We take one of the pitching motion to adjust the parameters of our filter system and use the rest to evaluate the performance. Parameter settings of our baseball customized filter system are shown in Table 9. Raw 3D pitching poses are pre-generated by the joint-wise volumetric triangulation method mentioned in Section III-B, while we use the more accurate Alphapose 2D HPE model (Fast Pose with ResNet152 backbone). Filter systems are then applied to remove the noises of the predicted

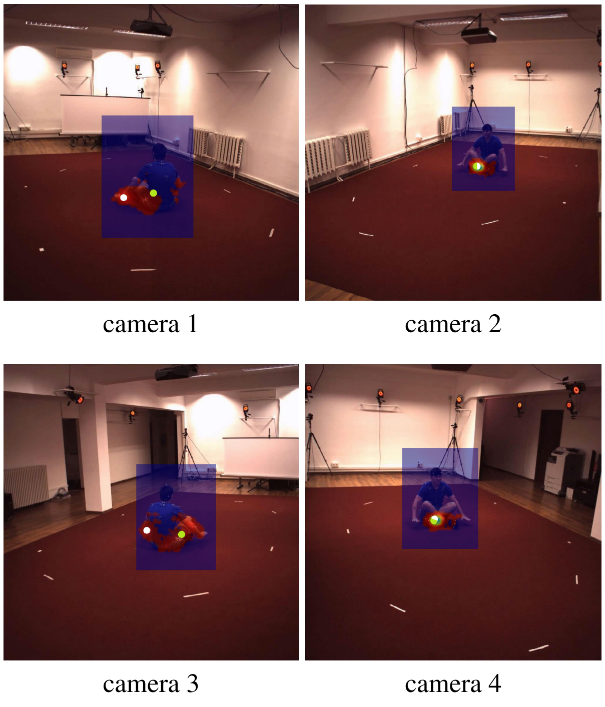


FIGURE 21. 2D joint (left foot) heatmaps generated by alphapose [30]. Here white dots indicate 2D joint predictions, while green dots indicate ground truth labels.

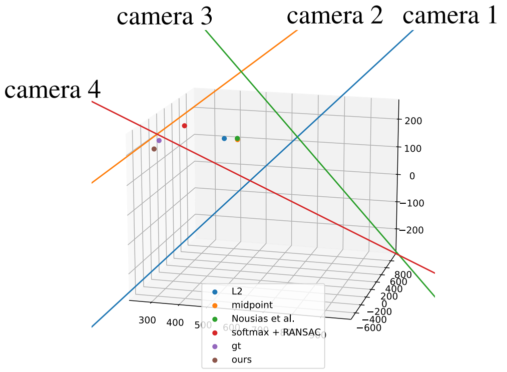


FIGURE 22. Visualization of 3D reconstruction results of Figure 21. Lines indicate epipolar lines generated by 2D joint predictions in each camera.

TABLE 6. Evaluation result on Human3.6M (MPJPE(mm)).

Methods	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.
L2 (2004) [49]	44.6	45.7	58.4	50.2	56.9	44.9	48.6	62.3
Midpoint (2004) [49]	42.1	44.0	53.5	48.1	52.5	44.9	46.6	53.1
Nousias et al. (2019) [56]	**42.0**	42.8	50.1	**46.5**	51.5	**44.5**	44.0	53.0
Lee et al. (2020) [57]	43.6	**42.1**	50.9	48.6	52.0	45.2	44.2	50.7
Softmax + RANSAC [19], [20]	44.5	45.7	58.6	49.9	55.1	44.7	48.0	53.6
Ours	44.2	44.0	**48.9**	48.1	**44.8**	46.9	**41.5**	**39.5**
	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	**Avg.**
L2(2004) [49]	75.8	56.1	55.4	48.3	46.7	50.8	45.0	52.6
Midpoint(2004) [49]	77.3	51.2	53.1	45.9	45.6	50.4	45.1	50.2
Nousias et al. (2019) [56]	57.6	50.0	52.1	45.7	**43.8**	47.6	**44.0**	47.6
Lee et al. (2020) [57]	57.1	50.6	52.5	46.2	45.6	48.2	45.5	48.2
Softmax + RANSAC [19], [20]	75.5	61.0	55.4	47.1	46.9	51.1	45.6	52.2
Ours	**45.6**	**43.5**	**50.0**	**45.1**	45.0	**46.9**	47.2	**45.4**

pitching pose, as evaluated by MPJPE, MPJVE, and MPJAE. The environment settings are the same as that mentioned in Section V-A1.

TABLE 7. Evaluation result on Human3.6M (P-MPJPE(mm)).

Methods	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.
L2 (2004) [49]	38.0	39.0	51.3	42.6	50.4	38.6	40.9	55.7
Midpoint (2004) [49]	35.4	37.2	46.4	40.6	46.1	38.6	39.5	46.1
Nousias et al. (2019) [56]	35.9	36.4	44.1	39.6	45.2	37.9	38.4	48.9
Lee et al. (2020) [57]	37.7	36.0	44.2	42.2	46.3	38.4	38.8	47.1
Softmax + RANSAC [19], [20]	38.2	39.5	52.1	42.6	49.2	38.7	40.4	47.9
Ours	**31.7**	**31.3**	**38.2**	**35.0**	**35.9**	**34.0**	**34.1**	**29.2**
	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	**Avg.**
L2 (2004) [49]	68.1	51.0	45.9	40.7	39.5	43.9	36.3	45.5
Midpoint (2004) [49]	70.3	45.4	43.5	38.5	39.0	44.1	36.7	43.2
Nousias et al. (2019) [56]	52.0	43.7	43.9	38.5	36.7	41.5	35.9	41.2
Lee et al. (2020) [57]	51.9	45.2	45.4	39.6	38.8	42.6	37.1	42.1
Softmax + RANSAC [19], [20]	67.7	58.6	45.9	39.9	40.2	44.4	37.5	45.5
Ours	**35.6**	**34.3**	**36.6**	**32.8**	**31.8**	**38.0**	**31.6**	**34.0**

TABLE 8. Evaluation result on MSL baseball dataset.

Methods	MPJPE(mm)	P-MPJPE(mm)
L2 (2004) [49]	43.0	34.8
Midpoint (2004) [49]	42.4	33.5
Nousias et al. (2019) [56]	36.3	28.6
Softmax + RANSAC [19], [20]	43.7	35.3
Ours	**33.2**	**24.9**

TABLE 9. Parameter setting.

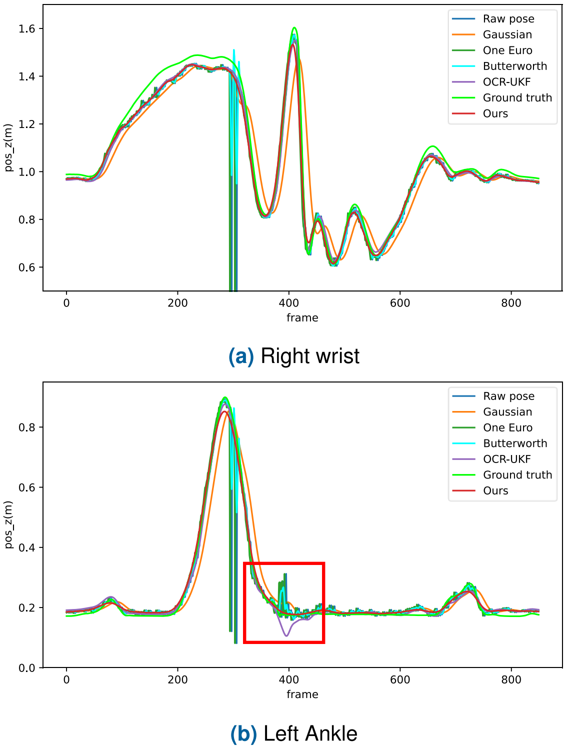
Parameters	Value	Meaning
Outlier removal		
N (Alg. 3)	4	Iteration of outlier detection
W (Alg. 3)	30	Window size of outlier detection
β (Alg. 3)	1.5	Boundary of outlier detection
Baseball customized filter		
l^{slow} (Sec. III-E)	21 (frame)	Window length of Median filter (slow movement)
l^{med} (Sec. III-E)	11 (frame)	Window length of Median filter (medium movement)
l^{fast} (Sec. III-E)	7 (frame)	Window length of Median filter (fast movement)
σ^{slow} (Sec. III-E)	10	Standard deviation of Gaussian filter kernel (slow movement)
σ^{med} (Sec. III-E)	8	Standard deviation of Gaussian filter kernel (medium movement)
σ^{fast} (Sec. III-E)	5	Standard deviation of Gaussian filter kernel (fast movement)
w^{slow} (Sec. III-E)	90 (Hz)	Cutoff frequency of Butterworth filter (slow movement)
w^{med} (Sec. III-E)	120 (Hz)	Cutoff frequency of Butterworth filter (medium movement)
w^{fast} (Sec. III-E)	120 (Hz)	Cutoff frequency of Butterworth filter (fast movement)

TABLE 10. Filter experiment results on MSL baseball dataset.

Methods	MPJPE(mm)	MPJVE(mm/frame)	MPJAE(mm/frame^2)
Raw poses	37.6	17.1	30.0
Gaussian filter [61]	37.5	1.58	0.66
Butterworth [62]	37.8	6.46	4.72
One Euro [64]	36.4	8.46	13.0
OCR-UKF [72]	34.3	1.30	0.46
Ours	**33.1**	**1.19**	**0.43**

2) EXPERIMENT RESULT

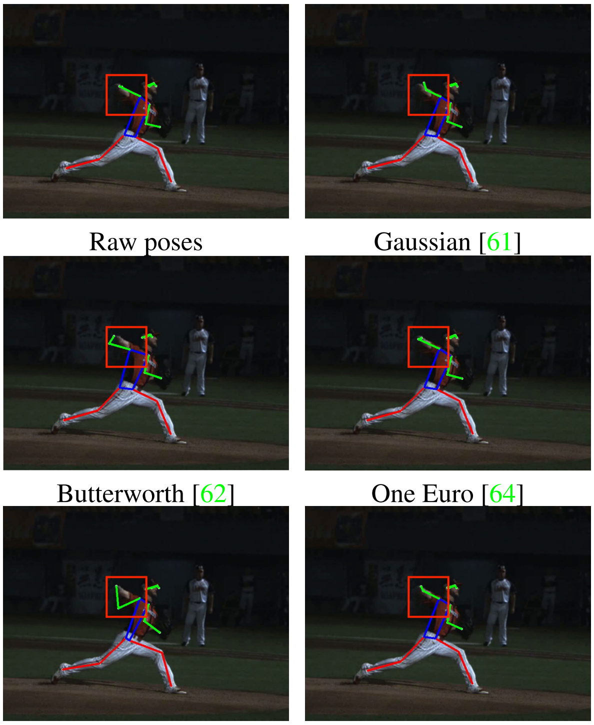
We compare the performance of our filter system in Section III-C on the baseball pitching motion against other filter systems including three classic low-pass filters: Gaussian filter [61], 4^{th} order Butterworth filter [62], and one Euro filter [64]. For Kalman filter prototype, we pick a human motion filter system [72] that also takes MSL Baseball Dataset as test dataset for comparison. Since there are only 15 pose sequences, it is infeasible to train ML based filter systems like SmoothNet [73]. Table 10 shows the result of the filter experiment on MSL Baseball Dataset. As observed, our baseball customized filter system reduces MPJPE by 12%, MPJVE by 93% and MPJAE by 99% compared to raw poses and outperforms all the other filter systems. As MSL Baseball Dataset was captured at 300 fps, the MPJVE and MPJAE of our system correspond to 0.35m/s (1.28 km/h)



(a) Right wrist

(b) Left Ankle

FIGURE 23. Trajectory of human joints in MSL baseball dataset.



Raw poses

Gaussian [61]

Butterworth [62]

One Euro [64]

OCR-UKF [72]

Ours

FIGURE 24. Filtering results on tainan baseball dataset.

average velocity error and 38.7 m/s^2 average acceleration error, respectively.

Figure 23 compares the filtered trajectory, among various filter systems, of the left leg and right wrist joints from the

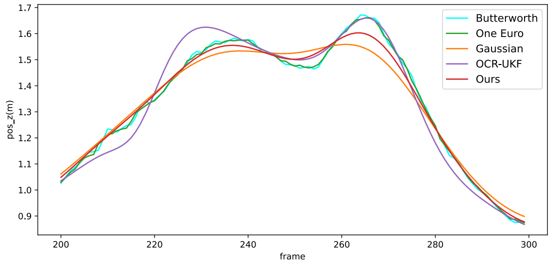


FIGURE 25. Trajectory of right wrist in tainan baseball dataset.

pitching pose. We observe large phase shift error in Gaussian filtering as well as signal jittering in one Euro filtering and Butterworth filtering, while our filter system does not have such issues. Comparing with OCR-UKF, our filter system shows better robustness when dealing with outliers in the raw poses.

To evaluate our filter system on real-world scenario, we also tested our filter system on in-the-wild Tainan Baseball Dataset. Video demonstration of the filtering poses can be found at [91]. Figure 24 and Figure 25 show the filtering results of a pitcher (subject 574) in Tainan Baseball Dataset during the arm acceleration phase. We can observe that our baseball customized filter, Butterworth filter and one Euro filter preserve most of the high frequency signal in the pitching action without signal lag, while our filter system has less jittering than the other two filters.

VI. CONCLUSION AND FUTURE WORK

This work focuses on multi-view 3D baseball pitcher pose reconstruction using 2D synchronized videos as input. We present a novel joint-wise volumetric triangulation method to aggregate the information from 2D joint heatmaps efficiently with volumetric heatmaps focused on preliminary joint predictions. Our approach outperforms other triangulation methods in terms of MPJPE and P-MPJPE. To filter out noises while preserving the high-speed baseball pitching motion, we designed a baseball customized filter system to categorize joints based on characteristics of pitching action and filter them separately. Compared with common filters for human pose filtering, our filter system achieves better velocity and acceleration performance than other common filter system for human pose filtering.

At present, our joint-wise volumetric triangulation operates without considering the relations between pose sequences, and our reconstruction system is customized for pitching motion of a single player. As future work, we expect to make our joint-wise volumetric triangulation time-aware and extend our system to cover other baseball movements such as batting or catching, that can be applied to multiple players on the field. By that, we expect to find more applications in live baseball game tracking.

VII. ACKNOWLEDGMENT

The authors thank Chiao Lai for providing MSL Baseball Dataset captured by Movement Science Laboratory.

REFERENCES

[1] Y.-C. Li, C.-T. Chang, C.-C. Cheng, and Y.-L. Huang, "Baseball swing pose estimation using openpose," in *Proc. IEEE Int. Conf. Robot., Autom. Artif. Intell. (RAAI)*, Apr. 2021, pp. 6–9.

[2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[3] L. Tian, X. Cheng, M. Honda, and T. Ikenaga, "Multi-view 3D human pose reconstruction based on spatial confidence point group for jump analysis in figure skating," *Complex Intell. Syst.*, vol. 9, no. 1, pp. 865–879, Feb. 2023.

[4] M. N. Saroja, K. R. Baskaran, and P. Priyanka, "Human pose estimation approaches for human activity recognition," in *Proc. Int. Conf. Advancements Electr., Electron., Commun., Comput. Autom. (ICAECA)*, Oct. 2021, pp. 1–4.

[5] Y. Cheng, P. Yi, R. Liu, J. Dong, D. Zhou, and Q. Zhang, "Human-robot interaction method combining human pose estimation and motion intention recognition," in *Proc. IEEE 24th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2021, pp. 958–963.

[6] S. D. Al-Sheekh and M. D. Younus, "Real-time pose estimation for human–robot interaction," in *Proc. 2nd Annu. Int. Conf. Inf. Sci. (AiCIS)*, 2020, pp. 86–90.

[7] H. Yan, B. Hu, G. Chen, and E. Zhengyuan, "Real-time continuous human rehabilitation action recognition using OpenPose and FCN," in *Proc. 3rd Int. Conf. Adv. Electron. Mater., Comput. Softw. Eng. (AEMCSE)*, Apr. 2020, pp. 239–242.

[8] Y.-P. Huang, Y.-J. Chou, and S.-H. Lee, "An OpenPose-based system for evaluating rehabilitation actions in Parkinson's disease," in *Proc. Int. Autom. Control Conf. (CACS)*, Nov. 2022, pp. 1–6.

[9] K. Otsuka, N. Yagi, Y. Yamanaka, Y. Hata, and Y. Sakai, "Joint position registration between OpenPose and motion analysis for rehabilitation," in *Proc. IEEE 50th Int. Symp. Multiple-Valued Log. (ISMVL)*, Nov. 2020, pp. 100–104.

[10] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[11] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and Ling Shao, "Deep 3D human pose estimation: A review," *Comput. Vis. Image Understand.*, vol. 210, Sep. 2021, Art. no. 103225.

[12] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social interaction capture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 190–204, Jan. 2019.

[13] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 506–516.

[14] D. Abbasi. (2021). *Phases of Throwing*. [Online]. Available: https://www.orthobullets.com/shoulder-and-elbow/3039/phases-of-throwing

[15] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[16] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.

[17] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1263–1272.

[18] D. Tome, M. Toso, L. Agapito, and C. Russell, "Rethinking pose in 3D: Multi-stage refinement and recovery for markerless motion capture," in *Proc. Int. Conf. 3D Vis. (3DV)*, 2018, pp. 474–483.

[19] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, "Learnable triangulation of human pose," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7717–7726.

[20] Z. Zhang, C. Wang, W. Qiu, W. Qin, and W. Zeng, "AdaFuse: Adaptive multiview fusion for accurate human pose estimation in the wild," *Int. J. Comput. Vis.*, vol. 129, no. 3, pp. 703–718, Mar. 2021.

[21] M. Kocabas, S. Karagoz, and E. Akbas, "Self-supervised learning of 3D human pose using multi-view geometry," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1077–1086.

[22] H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation learning for 3D human pose estimation," in *Proc. ECCV*, Sep. 2018, pp. 750–766.

[23] H. Tu, C. Wang, and W. Zeng, "VoxelPose: Towards multi-camera 3D human pose estimation in wild environment," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 197–212.

[24] Y. Chen, R. Gu, O. Huang, and G. Jia, "VTP: Volumetric transformer for multi-view multi-person 3D pose estimation," *Appl. Intell.*, vol. 53, pp. 26568–26579, Aug. 2022.

[25] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, pp. 4–27, Mar. 2010.

[26] C. K. Ingwersen, C. Mikkelstrup, J. N. Jensen, M. R. Hannemose, and A. B. Dahl, "Sportspose—A dynamic 3D sports pose dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5219–5228.

[27] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan, "Multi-view body part recognition with random forests," in *Proc. 24th Brit. Mach. Vis. Conf.*, 2013, pp. 1–11.

[28] H. Ma, L. Chen, D. Kong, Z. Wang, X. Liu, H. Tang, X. Yan, Y. Xie, S.-Y. Lin, and X. Xie, "Transfusion: Cross-view fusion with transformer for 3D human pose estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2021, pp. 1–15.

[29] H. Ma, Z. Wang, Y. Chen, D. Kong, L. Chen, X. Liu, X. Yan, H. Tang, and X. Xie, "PPT: Token-pruned pose transformer for monocular and multi-view human pose estimation," in *Proc. ECCV*, 2022, pp. 424–442.

[30] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7157–7173, Jun. 2023.

[31] D. C. Luvizon, D. Picard, and H. Tabia, "Consensus-based optimization for 3D human pose estimation in camera coordinates," *Int. J. Comput. Vis.*, vol. 130, no. 3, pp. 869–882, Mar. 2022.

[32] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt, "XNect: Real-time multi-person 3D motion capture with a single RGB camera," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 88:1–88:17, Aug. 2020.

[33] G. Moon, J. Y. Chang, and K. M. Lee, "Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10132–10141.

[34] B. X. Nie, P. Wei, and S.-C. Zhu, "Monocular 3D human pose estimation by predicting depth on joints," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3467–3475.

[35] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D pose estimation and action recognition using multitask deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 1802, pp. 5137–5146.

[36] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Computer Vision—ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham, Switzerland: Springer, 1711, pp. 536–553.

[37] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei, "Deep kinematic pose regression," in *Computer Vision—ECCV 2016*, G. Hua and H. Jégou, Eds., Cham, Switzerland: Springer, 2016, pp. 186–201.

[38] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3D human pose estimation in the wild: A weakly-supervised approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 398–407.

[39] D. C. Luvizon, H. Tabia, and D. Picard, "Human pose regression by combining indirect part detection and contextual information," *Comput. Graph.*, vol. 85, pp. 15–22, Dec. 2019.

[40] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2659–2668.

[41] J. Wang, S. Huang, X. Wang, and D. Tao, "Not all parts are created equal: 3D pose estimation by modeling bi-directional dependencies of body parts," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7770–7779.

[42] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3D human pose regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3420–3430.

[43] D. Jack, F. Maire, S. Shirazi, and A. Eriksson, "IGE-Net: Inverse graphics energy networks for human pose estimation and single-view reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7068–7077.

[44] C.-Hang Chen, A. Tyagi, A. Agrawal, D. Drover, R. Mv, S. Stojanov, and J. Rehg, "Unsupervised 3D pose estimation with geometric self-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5707–5717.

[45] A. Arnab, C. Doersch, and A. Zisserman, "Exploiting temporal context for 3D human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3390–3399.

[46] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7745–7754.

[47] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng, "Cross view fusion for 3D human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4341–4350.

[48] J. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, D. Touretzky, Ed., Burlington, MA, USA: Morgan-Kaufmann, 1989, pp. 1–7.

[49] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. New York, NY, USA: Cambridge Univ. Press, 2004.

[50] R. Hartley and F. Schaffalitzky, "L_∞ minimization in geometric reconstruction problems," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2004, pp. 504–509.

[51] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham, Switzerland: Springer, 2016, pp. 501–518.

[52] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4104–4113.

[53] I. M. Hakim, H. Zakaria, K. Muslim, and S. I. Ihsani, "3D human pose estimation using blazepose and direct linear transform (DLT) for joint angle measurement," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Feb. 2023, pp. 236–241.

[54] E. D'Antonio, J. Taborri, E. Palermo, S. Rossi, and F. Patanè, "A markerless system for gait analysis based on OpenPose library," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2020, pp. 1–6.

[55] E. D'Antonio, J. Taborri, I. Mileti, S. Rossi, and F. Patané, "Validation of a 3D markerless system for gait analysis based on OpenPose and two RGB webcams," *IEEE Sensors J.*, vol. 21, no. 15, pp. 17064–17075, Aug. 2021.

[56] S. Nousias, M. Lourakis, and C. Bergeles, "Large-scale, metric structure from motion for unordered light fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3287–3296.

[57] S. H. Lee and J. Civera, "Robust uncertainty-aware multiview triangulation," 2008, *arXiv:2008.01258*.

[58] S. Schreven, P. J. Beek, and J. B. J. Smeets, "Optimising filtering parameters for a 3D motion analysis system," *J. Electromyogr. Kinesiol.*, vol. 25, no. 5, pp. 808–814, Oct. 2015.

[59] E. Martini, A. Calanca, and N. Bombieri, "Denoising and completion filters for human motion software: A survey with code," Dept. Eng. Innov. Med., Univ. Verona, Verona, Italy, 2023, doi: 10.36227/techrxiv.22956482.v2. [Online]. Available: https://www.techrxiv.org/doi/full/10.36227/techrxiv.22956482.v2

[60] C. D Mathys, E. I Lomakina, J. Daunizeau, S. Iglesias, K. H Brodersen, K. J Friston, and K. E Stephan, "Uncertainty in perception and the hierarchical Gaussian filter," *Frontiers Hum. Neurosci.*, vol. 8, p. 825, Nov. 2014.

[61] I. T. Young and L. J. van Vliet, "Recursive implementation of the Gaussian filter," *Signal Process.*, vol. 44, no. 2, pp. 139–151, Jun. 1995.

[62] S. Butterworth, "On the theory of filter amplifiers," *Wireless Eng.*, vol. 7, no. 6, pp. 536–541, 1930.

[63] F. Crenna, G. B. Rossi, and M. Berardengo, "Filtering biomechanical signals in movement analysis," *Sensors*, vol. 21, no. 13, p. 4580, Jul. 2021.

[64] G. Casiez, N. Roussel, and D. Vogel, "1€ filter: A simple speed-based low-pass filter for noisy input in interactive systems," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, May 2012, pp. 2527–2530.

[65] Y. Zou, J. Yang, D. Ceylan, J. Zhang, F. Perazzi, and J.-B. Huang, "Reducing footskate in human motion reconstruction with ground contact constraints," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 448–457.

[66] G. Welch and G. Bishop, "An introduction to the Kalman filter," Univ. North Carolina Chapel Hill, Chapel Hill, NC, USA, Lect. Note, 1995.

[67] Maria Isabel Ribeiro, "Kalman and extended Kalman filters: Concept, derivation and properties," *Inst. Syst. Robot.*, vol. 43, no. 46, pp. 3736–3741, 2004.

[68] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proc. IEEE Adapt. Syst. Signal Process., Commun., Control Symp.*, Oct. 2000, pp. 153–158.

[69] F. Ahmed, A. S. M. Hossain Bari, B. Sieu, J. Sadeghi, J. Scholten, and M. L. Gavrilova, "Kalman filter-based noise reduction framework for posture estimation using depth sensor," in *Proc. IEEE 18th Int. Conf. Cognit. Informat. Cognit. Comput. (ICCI*CC)*, Jul. 2019, pp. 150–158.

[70] J. Niu, X. Wang, D. Wang, and L. Ran, "A novel method of human joint prediction in an occlusion scene by using low-cost motion capture technique," *Sensors*, vol. 20, no. 4, p. 1119, Feb. 2020.

[71] Y. R. Musunuri and O.-S. Kwon, "State estimation using a randomized unscented Kalman filter for 3D skeleton posture," *Electronics*, vol. 10, no. 8, p. 971, Apr. 2021.

[72] C. Lai. (2022). *A Markerless Multi-View 3D Human Motion Estimation System for Single Person With Modified Unscented Kalman Filter and Iterative LQR Tracking*. [Online]. Available: https://hdl.handle.net/11296/ym4z9n

[73] A. Zeng, L. Yang, X. Ju, J. Li, J. Wang, and Q. Xu, "SmoothNet: A plug-and-play network for refining human poses in videos," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 625–642.

[74] M. Véges and A. Lőrincz, "Temporal smoothing for 3D human pose estimation and localization for occluded people," in *Proc. Int. Conf. Neural Inf. Process.*, 2011, pp. 557–568.

[75] D.-Y. Kim and J. Chang, "Attention-based 3D human pose sequence refinement network," *Sensors*, vol. 21, p. 4572, Jul. 2021.

[76] T. V. Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 601–617.

[77] Z. Jiang, H. Ji, S. Menaker, and J.-N. Hwang, "GolfPose: Golf swing analyses with a monocular camera based human pose estimation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2022, pp. 1–6.

[78] R. Kurose, M. Hayashi, T. Ishii, and Y. Aoki, "Player pose analysis in tennis video based on pose estimation," in *Proc. Int. Workshop Adv. Image Technol. (IWAIT)*, Jan. 2018, pp. 1–4.

[79] C. Yan, X. Li, and G. Li, "A new action recognition framework for video highlights summarization in sporting events," in *Proc. 16th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Aug. 2021, pp. 653–666.

[80] N. F. Janbi and N. Almuaythir, "BowlingDL: A deep learning-based bowling players pose estimation and classification," in *Proc. 1st Int. Conf. Adv. Innov. Smart Cities (ICAISC)*, Jan. 2023, pp. 1–6.

[81] P. Murthy, B. Taetz, A. Lekhra, and D. Stricker, "DiveNet: Dive action localization and physical pose parameter extraction for high performance training," *IEEE Access*, vol. 11, pp. 37749–37767, 2023.

[82] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV 2014*, A. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham, Switzerland: Springer, 2014, pp. 740–755.

[83] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

[84] C. S. Tony Hii, K. B. Gan, N. Zainal, N. M. Ibrahim, S. A. M. Rani, and N. A. Shattar, "Marker free gait analysis using pose estimation model," in *Proc. IEEE 20th Student Conf. Res. Develop. (SCOReD)*, Nov. 2022, pp. 109–113.

[85] L. Bridgeman, M. Volino, J.-Y. Guillemaut, and A. Hilton, "Multi-person 3D pose estimation and tracking in sports," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2487–2496.

[86] T. Murakami and T. Nakamura, "Athlete 3D pose estimation from a monocular TV sports video using pre-trained temporal convolutional networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 2615–2620.

[87] Y. Zhang, Q. Wang, F. Tu, and Z. Wang, "Automatic moving pose grading for golf swing in sports," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 41–45.

[88] J. W. Tukey, *Exploratory Data Analysis*, vol. 2. Reading, MA, USA: Addison-Wesley, 1977.

[89] J. C. Gower and G. B. Dijksterhuis, *Procrustes Problems*. London, U.K.: Oxford Univ. Press, 2004.

[90] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," in *Readings in Computer Vision*, M. A. Fischler and O. Firschein, Eds., San Francisco, CA, USA: Morgan Kaufmann, 1987, pp. 726–740.

[91] Y.-W. Chiu. (2023). *Video Demonstration of Pose Filtering Result*. [Online]. Available: https://drive.google.com/drive/folders/1w8XfYGJEQXsaotMDu2kRDtOfSeyEKKAe?usp=sharing

YUN-WEI CHIU was born in Taipei, Taiwan, in 1998. He received the B.S. and M.S. degrees in electrical engineering from the Graduate Institute of Communication Engineering, National Taiwan University. He has joined the Machine Learning and Estimation Theory Laboratory, where he was a Research Assistant, from 2020 to 2023. His main research focuses on computer vision, machine learning, 3-D reconstruction, and signal processing.

KUEI-TING HUANG is currently pursuing the master's degree with the Graduate Institute of Photonics and Optoelectronics, National Taiwan University. As an avid baseball enthusiast, he primarily engages in reconstructing pitching motions, utilizing multi-view cameras for 2-D recognition, and 3-D reconstruction. Delving deeper into human biomechanics, he explores methods to enhance pitching motion reconstruction. Combining his passion for baseball with his research, he frequently incorporates self-testing into his studies and continuously optimize pitching motion analysis through his knowledge of the sport. He expects to develop a comprehensive pitching analysis 3-D reconstruction systems, contributing to both academia and the field of sports science. His research interests include image-based sports science, encompassing image recognition, image processing, and stereo computer vision.

YUH-RENN WU (Senior Member, IEEE) received the B.S. degree in physics and the M.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, in 2006. He is currently a Professor with the Institute of Photonics and Optoelectronic and the Department of Electrical Engineering, National Taiwan University. His area of research is in physics, the design of optoelectronic devices, and sensors. His current research interests include studies of device simulation and physical modeling, developing simulation software for semiconductor devices, optoelectronics, and image analysis.

JYH-HOW HUANG is currently a Professor with the Sport Information and Communication Department, National Taiwan University of Sport, Taiwan. His research interests include sport performance analysis, sports biomechanics, and sensor networks. He spent most of his research effort on baseball in recent years.

WEI-LI HSU is a Clinical Researcher whose research centers on gait and posture in people with movement disorders. She has been involved in numerous human movement research projects. She has formulated research studies based on clinical observations and has made valuable contributions to many research projects. She also continues her clinical practice as a Physical Therapist with the National Taiwan University Hospital. She and her team has extensive experience in characterizing movement patterns in patients with spinal disease and poor balance and analyzing the effect of rehabilitation and surgery on movement pattern. Her research aims to improve the understanding of interjoint coordination in balance control, and it provides a basis for developing tools to treat patients with sensorimotor deficits leading to balance disorders.

PEI-YUAN WU (Member, IEEE) received the B.S.E. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 2009, and the M.A. and Ph.D. degrees in electrical engineering from Princeton University, in 2012 and 2015, respectively. He was with TSMC, from 2015 to 2017. He has been an Associate Professor with the Department of Electrical Engineering, National Taiwan University, since 2017. His research interest include artificial intelligence, signal processing, estimation and prediction, and cyber-physical system modeling. He was a recipient of the Gordon Y. S. Wu Fellowship, in 2010; the Outstanding Teaching Assistant Award at Princeton University, in 2012; and the 2020 FutureTech Breakthrough Award held by MOST.