ACM SIGGRAPH / Eurographics Symposium on Computer Animation (2024)
M. Skouras and H. Wang (Editors)

Markerless Multi-view Multi-person Tracking for Combat Sports

Hossein Feiz[1], David Labbé[1], Sheldon Andrews[1]

[1]École de technologie supérieure, Montreal, Canada

Abstract

We introduce a novel framework for 3D pose estimation in combat sports. Utilizing a sparse multi-camera setup, our approach employs a computer vision-based tracker to extract 2D pose predictions from each camera view, enforcing consistent tracking targets across views with epipolar constraints and long-term video object segmentation. Through a top-down transformer-based approach, we ensure high-quality 2D pose extraction. We estimate the 3D position via weighted triangulation, spline fitting and extended Kalman filtering. By employing kinematic optimization and physics-based trajectory refinement, we achieve state-of-the-art accuracy and robustness under challenging conditions such as occlusion and rapid movements. Experimental validation on diverse datasets, including a custom dataset featuring elite boxers, underscores the effectiveness of our approach. Additionally, we contribute a valuable sparring video dataset to advance research in multi-person tracking for sports.

CCS Concepts

• ***Computing methodologies*** → *Pose Estimation; Optimization;*

1. Introduction

Combat sports present significant challenges for motion capture due to numerous close-proximity interactions and frequently crowded backgrounds. Optical marker-based tracking, while precise in controlled environments, becomes impractical due to dynamic motions and frequent collisions leading to calibration issues. Inertial measurement unit (IMU) based solutions suffer from global positional drift, affecting inter-athlete distances. Monocular vision-based approaches, though freeing athletes from tracking equipment, often lack precision due to frequent occlusions. However, these occlusions can be mitigated by incorporating data from multiple camera viewpoints, a cornerstone of our tracking pipeline.

We propose a multi-stage, multi-view tracking pipeline shown in Fig 1 designed to reconstruct high-quality 3D motion of athletes engaged in combat sports such as boxing. Our approach integrates 2D keypoints from multiple camera views through kinematic optimization, followed by physics-based trajectory refinement using model predictive control to eliminate non-physical artifacts. The contributions of our work are summarized as follows:

- A comprehensive multi-camera multi-person physics-based pose estimation framework designed for high-quality 3D pose estimation using as few as three cameras.
- A robust triangulation technique employing spline fitting and Kalman filtering to generate consistent and smooth 3D positions.
- A high-quality dataset of > 20 minutes of video footage featuring elite boxers during sparring sessions that encompasses several boxing styles, plus a multi-view dataset of motions representative of combat sports and synchronized with ground truth from an optical marker tracking system. We will release these datasets and the solved motions.

2. Pose Estimation

Our tracking pipeline is summarized in Fig. 1. We describe the main stages below.

Tracking 2D and 3D: Using epipolar constraints and long-term video object segmentation [CS22] we produce consistent ids for everyone, these ids are used to produce 2D joints positions using [XZZT22] for tracking targets. We then use linear triangulation and Kalman estimation to produce robust 3D joints positions for each individual, even in the presence of noise and outliers.

Kinematics Optimization: The kinematics optimization focuses on refining the pose estimation of athletes using 2D and 3D keypoint data. Fig 2 shows the results of the kinematics stage on different datasets. The optimization initializes shape parameters ($\beta \in \mathbb{R}^{10}$) of the SMPL model based on 3D keypoints obtained through triangulation. Subsequently, it iteratively adjusts shape and pose parameters ($\theta \in \mathbb{R}^{72}$) to refine the pose estimation based on objectives for smoothness, similarity to human motion priors, and alignment with both 2D re-projection evidence and triangulated 3D keypoints. We summarize these objective terms below:

- **2D Re-projection loss** (L_{2D}): Aligns 3D with 2D keypoints across multiple views, emphasizing high-confidence joints.
- **3D Alignment loss** (L_{3D}): Distance between predicted joint locations and triangulated keypoints, weighted by their confidence.
- **Smoothness loss** (L_{smooth}): Promotes temporal coherency in pose transitions from frame-to-frame.

DOI: 10.2312/sca.20241162

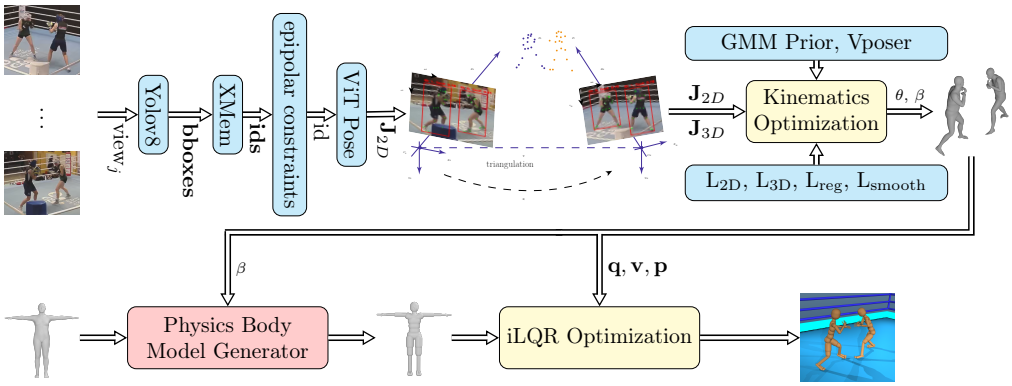

Figure 1: *Our pipeline begins with generating bounding boxes and tracking ids for each individual in the scene, which are then used to produce 2D poses for each individual for each view j. A triangulation process is then used to compute 3D keypoints. The kinematics optimization step incorporates the 2D and 3D keypoints to compute SMPL parameters (θ, β). The 3D relative joint positions, initial pose state and velocity state of the humanoid, serve as a reference for a dynamic optimizer to correct artifacts in the motion.*

Figure 2: *Poses estimated from the Campus (left), Shelf (middle) datasets, and our custom supplementary (right) dataset.*

- **Prior losses** (L_{GMM}, L_{Vposer}): Gaussian Mixture Model (GMM) and Vposer [PCG*19] priors to penalize unnatural poses.

Dynamics Optimization: Motions produced by the kinematic stage often contain high-frequency jitter and foot skating. We found that a dynamics optimization using a physics-based humanoid model helps to mitigate these artifacts. The model consists of an articulated rigid-body structure with 56 joint-angle degrees of freedom, plus 6 degrees of freedom for the root motion. Capsule collision geometry aligned with SMPL landmarks comprises the shape. The dynamics optimization refines motion trajectories from the kinematics stage by considering joint torques and biomechanical constraints within the physical environment. Joint torques are computd using an iLQR algorithm [HGT*22], This approach accounts for contact forces and body dynamics, enhancing the overall quality and naturalness of the generated motions by iteratively refining control trajectories over short time horizons. Table 1 presents quantitative metrics computed using our supplementary dataset on the motions produced by the kinematics optimizaton stage and after the dynamics optimization. The dynamics optimization clearly increases the naturalness of the solved motions.

Table 1: *MPJPE, foot skating and floating [XWI*21], and smoothness [SGXT20] metrics before and after the dynamics optimization.*

	$e_{\text{MPJPE}} \downarrow$	$e_{\text{foot},z} \downarrow$	$e_{\text{foot},v_{xy}} \downarrow$	$e_{\text{smooth}} \downarrow$
Kinematics	41.2	16.4	2.2	6.1
Dynamics	38.4	8.1	0.3	4.6

References

[CS22] CHENG H. K., SCHWING A. G.: XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *The European Conference on Computer Vision (ECCV)* (2022). doi:10.48550/arXiv.2207.07115. 1

[HGT*22] HOWELL T., GILEADI N., TUNYASUVUNAKOOL S., ZAKKA K., EREZ T., TASSA Y.: Predictive Sampling: Real-time Behaviour Synthesis with MuJoCo. arXiv:2212.00541, doi:10.48550/arXiv.2212.00541. 2

[PCG*19] PAVLAKOS G., CHOUTAS V., GHORBANI N., BOLKART T., OSMAN A. A. A., TZIONAS D., BLACK M. J.: Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2019). 2

[SGXT20] SHIMADA S., GOLYANIK V., XU W., THEOBALT C.: PhysCap. *ACM Transactions on Graphics 39*, 6 (nov 2020), 1–16. 2

[XWI*21] XIE K., WANG T., IQBAL U., GUO Y., FIDLER S., SHKURTI F.: Physics-based human motion estimation and synthesis from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2021), pp. 11532–11541. 2

[XZZT22] XU Y., ZHANG J., ZHANG Q., TAO D.: Vitpose++: Vision transformer foundation model for generic body pose estimation. *arXiv preprint arXiv:2212.04246* (2022). 1