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

Markerless Multi-view and Multi-person Tracking for Combat Sports

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

[1]École de technologie supérieure, Montreal, Canada

Abstract

This paper introduces 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; Motion synthesis;*

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 2 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 video footage featuring elite boxers during intense sparring sessions. The dataset encompasses various boxing styles, presenting a valuable new benchmark in 3D pose estimation. We will release a portion of the dataset (20 minutes) and motion data obtained to advance research in competitive sports.

2. Pose Estimation

Tracking 2D and 3D Data: 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. Employing the SMPL model, the optimization aims to minimize the disparity between model joints and observed data while ensuring temporal coherence and natural movement. This is achieved through a comprehensive objective function that includes terms for smoothness, similarity to human motion priors, and alignment with both 2D re-projection evidence and triangulated 3D keypoints.

Initially, 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.

Figure 1: *Poses estimated from the Campus, Shelf and supplementary datasets*

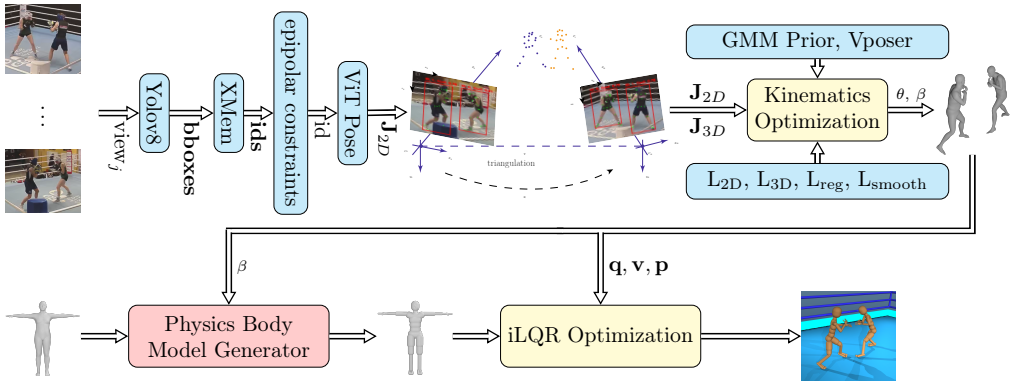

Figure 2: *The pipeline begins with generating bounding boxes and robust tracking for each individual in the scene. These tracking results are used to produce 2D poses. The triangulation process, produces smooth 3D keypoints. The kinematics optimization step incorporates the 2D and 3D keypoints, to create the 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 any artifacts in the motion.*

Key components of the objective function include:

- **2D Re-projection Loss** (L_{2D}): Aligns 3D joints with 2D joints across multiple camera views, emphasizing joints with high-confidence detections using a robust error function.
- **3D Alignment Loss** (L_{3D}): Computes the Euclidean distance between predicted 3D joint positions and triangulated 3D keypoints, weighted by their confidence scores.
- **Smoothness Loss** (L_{smooth}): Promotes consistency in pose transitions over time frames and vertices of the posed mesh.
- **Prior Losses** (L_{GMM}, L_{Vposer}): Introduces Gaussian Mixture Model (GMM) and Vposer priors to penalize unnatural poses, guiding the optimization towards more realistic and fluid motion.

This comprehensive approach ensures accurate and lifelike pose estimation suitable for applications in sports biomechanics and animation. Fig 1 shows the qualitative results of the pipeline on different datasets.

Dynamics Optimization: Output motion of the kinematics optimization contains high-frequency artifacts. To mitigate this, a second stage employs dynamics optimization using a physics-based humanoid model. This model, derived from joint positions and landmarks of an SMPL mesh, creates an articulated rigid-body structure with capsule collision geometry, featuring 56 joint-angle degrees of freedom and a 6 degree of freedom root joint.

The dynamics optimization stage aims to refine motion trajectories by considering joint torques and biomechanical constraints within a physical environment. It computes joint torques using an iterative Linear Quadratic Regulator (iLQR) algorithm [HGT*22], which optimizes control inputs to smooth and stabilize motion trajectories. 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.

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

[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