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

Combat sports pose significant challenges for motion capture due to dynamic interactions and crowded backgrounds. Traditional methods like optical marker tracking, IMU-based solutions are not practical due to their intrusive nature, and Monocular vision-based approaches are affected by occlusions, encouraging the development of a multi-stage, multi-view tracking pipeline. This pipeline utilizes kinematic optimization to fuse 2D keypoints from multiple cameras, followed by physics-based trajectory optimization using model predictive control to enhance realism.

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 markers tracking is precise in a laboratory setting, yet are impractical for combat sports due to the highly dynamic motions and frequent collisions that lead to calibration issues with markers. Inertial measurement unit (IMU) based solutions suffer from global positional drift, affecting inter-athlete distances. Monocular vision-based approaches leave athletes unencumbered by tracking equipment, but they lack precision due to insufficient visual data arising from frequent occlusions. Issues due to occlusion can however be mitigated by incorporating camera data from multiple viewpoints, an idea that we built on in our tracking pipeline.

We propose a multi-stage multi-view tracking pipeline that is capable of reconstructing high-quality 3D motion of athletes participating in combat sports, such as boxing. Our approach fuses 2D keypoints from multiple camera views by a kinematic optimization. This is followed by a physics-based trajectory optimization utilizing model predictive control to eliminate non-physical artifacts.

The experiments presented in the results section of this paper demonstrate that our approach is capable of reconstructing complex movements while considering various physical constraints, such as contact dynamics and collisions. We evaluate our pipeline using multi-view footage of elite-level boxers in a ring as illustrated in Figure 2. 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 we produce consistent ids for everyone, these ids are used to produce 2D joints positions 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.

Key components of the objective function include:

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

Table 1: *Comparison of PCP[% ↑] results on the Campus and Shelf datasets.*

Method	Campus				Shelf			
	Actor 1	Actor 2	Actor 3	Avg.	Actor 1	Actor 2	Actor 3	Avg.
[WZC*21]	99.3	95.1	97.8	97.4	98.2	94.1	97.4	96.6
[YOY*24]	98.2	94.6	98.2	97.0	99.5	96.0	97.7	97.7
ours (Triangulation)	**99.6**	92.2	97.6	96.5	**99.8**	95.4	**98.6**	**97.9**
ours (Kinematics)	98.5	93.5	94.4	95.5	**99.8**	**97.6**	**98.6**	**98.6**
ours (Dynamics)	97.7	93.6	94.2	95.1	97.1	**97.6**	97.2	97.3

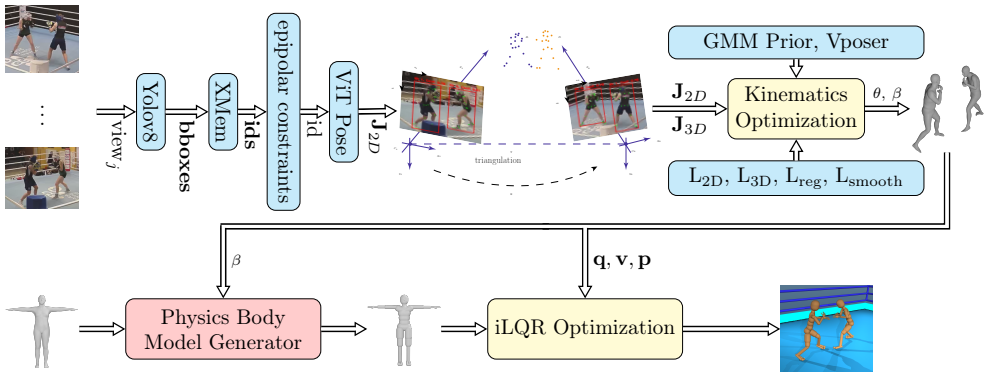

Figure 2: *The pipeline begins with generating bounding boxes and robust tracking for each individual in the scene. These tracking results are used to produce 2D poses. The triangulation process, produces smooth 3D keypoints. The kinematics optimization step incorporates the 2D and 3D keypoints, to create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

- **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.

Dynamics Optimization: Output motion of the kinematics optimization contains high-frequency artifacts. To mitigate this, a sec-

ond 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, 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

[WZC*21] WANG T., ZHANG J., CAI Y., YAN S., FENG J.: Direct multi-view multi-person 3d pose estimation. In *Advances in Neural Information Processing Systems* (2021), Ranzato M., Beygelzimer A., Dauphin Y., Liang P., Vaughan J. W., (Eds.), vol. 34, Curran Associates, Inc., pp. 13153–13164. 2

[YOY*24] YANG F., ODASHIMA S., YAMAO S., FUJIMOTO H., MASUI S., JIANG S.: A unified multi-view multi-person tracking framework. *Computational Visual Media 10*, 1 (2024), 137–160. 2