Multi-person Physics-based Pose Estimation for Combat Sports

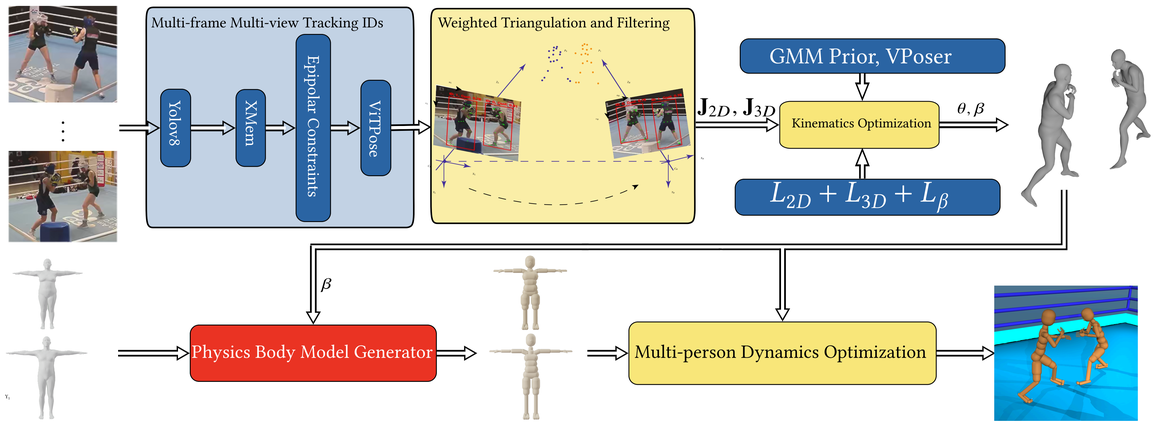
Hossein Feiz[1] David Labbé[1] Sheldon Andrews[1]

[1]École de technologie supérieure, Montréal, Québec, Canada

Introduction

Motion capture from only RGB cameras in a typical sports scene (e.g., boxing), with the presence of coaches, viewers, and close interactions between athletes, brings many challenges, such as heavy occlusion, background crowding, and fast, complicated movements. To address this problematic, a multi-view configuration was used. Our system only required the visibility of an athlete from two cameras. We also created a pipeline using machine and deep learning techniques that automated the process of creating animations

Pipeline



The pipeline begins with generating bounding boxes (**bboxes**) and robust tracking (id) for each individual in the scene. These tracking results are used to produce 2D poses ($\mathbf{J}_{2D}$) using ViTPose [5]. The triangulation process, produces smooth 3D keypoints ($\mathbf{J}_{3D}$). The kinematics optimization step incorporates the 2D and 3D keypoints, to create the SMPL parameters (θ, β). The 3D relative joint positions ($\mathbf{p}$), initial pose state ($\mathbf{q}$) and velocity state ($\mathbf{v}$) of the humanoid (created using the physics body model generator), serve as a reference for a model predictive controller, which utilizes the iLQR optimizer [2] to correct any artifacts in the motion.

Multi-frame Multi-view Tracking IDs

- **Tracking by Segmentation**: XMem segments individuals per view for consistent ID assignment across frames. [1]
- **Epipolar Constraint-based ID Matching**: across views using bounding box centroids and epipolar constraints. [4]
- **Top-Down 2D Pose Estimation**: keypoints are localized within bounding boxes for precise pose estimation. [5]

Weighted Triangulation and Filtering

The weighted triangulation estimates the 3D positions of the keypoints from their corresponding 2D keypoints from all N cameras. The 3D positions of each joint are thus determined by solving the linear system:

Weighted Triangulation Formulation

$$\begin{bmatrix} \mu_1(\mathbf{P}_{11} - \mathrm{u}_1\mathbf{P}_{31}) \\ \mu_1(\mathbf{P}_{21} - \mathrm{v}_1\mathbf{P}_{31}) \\ \vdots \\ \mu_N(\mathbf{P}_{1N} - \mathrm{u}_N\mathbf{P}_{3N}) \\ \mu_N(\mathbf{P}_{2N} - \mathrm{v}_N\mathbf{P}_{3N}) \end{bmatrix} \begin{bmatrix} \mathrm{x} \\ \mathrm{y} \\ \mathrm{z} \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (1)$$

$\mathbf{P}_{ij}$ extracts the i-th row of the projection matrix $\mathbf{P}_j$. μ_j is the average confidence of a 2D point for the current frame in camera j. Higher confidence values give more weight to certain cameras.

We use SVD to solve Eq. 1. After triangulation, outliers are handled by interpolating and smoothing using cubic spline for each joint's trajectory.

If there is no solution for triangulation, we utilize an extended Kalman filter for dynamic state estimation based on velocity, acceleration of the keypoint, and position constraint to estimate the 3D keypoints. This results in robust filtering and smoothing while preserving trajectory integrity, enabling dependable 3D reconstructions from sparse and noisy 2D keypoints.

Kinematics Optimization

Our kinematics optimization refines the SMPL model to minimize the difference between the provided 2D poses ($\mathbf{J}_{2D}$) from multiple views and 3D keypoints ($\mathbf{J}_{3D}$) obtained from triangulation. The process ensures temporal coherency and natural motion by incorporating smoothness and human motion priors. An LBFGS optimizer is used to solve the minimization problem, considering various loss terms, including 2D re-projection, 3D alignment, smoothness, and prior losses.

Kinematics Optimization Formulation

$$\min_{\theta} \quad w_1 \mathrm{L}_{2D} + w_2 \mathrm{L}_{3D} + w_3 \mathrm{L}_{\mathrm{reg}} + w_4 \mathrm{L}_{\mathrm{smooth}} + w_5 \mathrm{L}_{\mathrm{GMM}} + w_6 \mathrm{L}_{\mathrm{Vposer}} \,. \quad (2)$$

$$\mathrm{L}_{2D} = \sum_{\mathrm{j}\in\mathcal{V}} \sum_{\mathrm{i}\in\mathbf{J}_{2D}} c_{\mathrm{j,i}} \rho(\mathrm{J}_{\mathrm{proj}_{j,i}} - \mathrm{J}_{2D_{j,i}}), \quad (3)$$

2D Re-projection Loss (3): Aligns the 3D joints with 2D joints using re-projection, weighted by confidence.

$$\mathrm{L}_{3D} = \sum \mathrm{i} \in \mathbf{J}_{3D} c_i \|\mathbb{J}_i(\theta, \beta) - \mathrm{J}_{3D,i}\|^2, \quad (4)$$

3D Alignment Loss (4): Minimizes the difference between predicted and actual 3D joint positions, considering joint confidence.

$$\mathrm{L}_{\mathrm{smooth}} = \sum_t \|\theta^t - \theta^{t-1}\|^2 + \|\mathscr{M}(\theta^t, \beta) - \mathscr{M}(\theta^{t-1}, \beta)\|^2 . \quad (5)$$

Smoothness Loss (5): Ensures temporal consistency by minimizing differences in poses and vertices over time.

$$\mathrm{L}_{\mathrm{GMM}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{GMM}(\theta_i, \beta), \quad \mathrm{L}_{\mathrm{Vposer}} = \frac{1}{N}\sum_{i=1}^{N} (\mathrm{z}(\theta_i)^2) . \quad (6)$$

Prior Loss (6): Guides the optimization towards natural poses using GMM and Vposer priors.

Physics Body Model Generator

Our goal is to reconstruct the motion of K physics-based humanoids at each time step t. We define the combined state as $\mathbf{x}_t = (\mathbf{q}_t, \mathbf{v}_t)$, where the joint rotations vector is $\mathbf{q}_t \in \mathbb{R}^{63K}$, and the joint velocities vector is $\mathbf{v}_t \in \mathbb{R}^{62K}$. The 3D SMPL joint positions are represented as $\mathbf{p}_t \in \mathbb{R}^{21K}$.

The vector $\mathbf{q}_t$ uses minimal coordinates for joint rotations, while $\mathbf{p}_t$ represents the 3D positions relative to the humanoid's body frame, which are used in optimization. For humanoid modeling, we used the Mujoco [3] XML format.

Multi-person Dynamics Optimization

The iLQR algorithm [2] optimizes a control trajectory $\mathbf{u}_{0:T}$ by minimizing the cost function 7. The reference positions $\mathbf{p}_t$ and velocities $\mathbf{v}_t$ are extracted from the kinematics optimization at each frame t, and the trajectory optimization is "warm-started" using control parameters $\mathbf{u}_{t-1}$ from the previous frame. The control is updated iteratively as $\Delta\mathbf{u}_t = \mathbf{K}_t\Delta\mathbf{x}_t + \alpha\mathbf{k}_t$, with $\mathbf{K}_t$ and $\mathbf{k}_t$ refined until convergence.

Dynamics Optimization Formulation

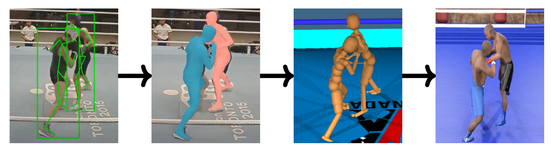
$$\min_{\mathbf{u}_{0:T}} \sum_{t=0}^{T} w_1 \mathrm{L}_{\mathrm{reg,t}} + w_2 \mathrm{L}_{\mathrm{p},t} + w_3 \mathrm{L}_{\mathrm{v},t} + w_4 \mathrm{L}_{\mathrm{collision},t} \,. \quad (7)$$

Discussion and Future Works

We created a 10-hour dataset of boxing motions showcasing various interactions, techniques, and styles. This dataset is being used to develop an interaction motion synthesis model for generating intelligent VR responses.

Video

Visualization



References

[1] Ho Kei Cheng and Alexander G. Schwing. "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.

[2] Taylor Howell et al. "Predictive Sampling: Real-time Behaviour Synthesis with MuJoCo". In: (Dec. 2022). DOI: 10.48550/arXiv.2212.00541. arXiv: 2212.00541 [cs.RO].

[3] Emanuel Todorov, Tom Erez, and Yuval Tassa. "MuJoCo: A physics engine for model-based control". In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2012, pp. 5026–5033. DOI: 10.1109/IROS.2012.6386109.

[4] Christian Wöhler. *3D computer vision: efficient methods and applications*. Springer Science & Business Media, 2009.

[5] Yufei Xu et al. "ViTPose++: Vision Transformer Foundation Model for Generic Body Pose Estimation". In: *arXiv preprint arXiv:2212.04246* (2022).

hfa.zadeh.k@gmail.com Linked in Hossein Feiz X hosseinfeiz1