

Multi-person Physics-based Pose Estimation for Combat Sports

Hossein Feiz¹ David Labb  ¹

Sheldon Andrews¹

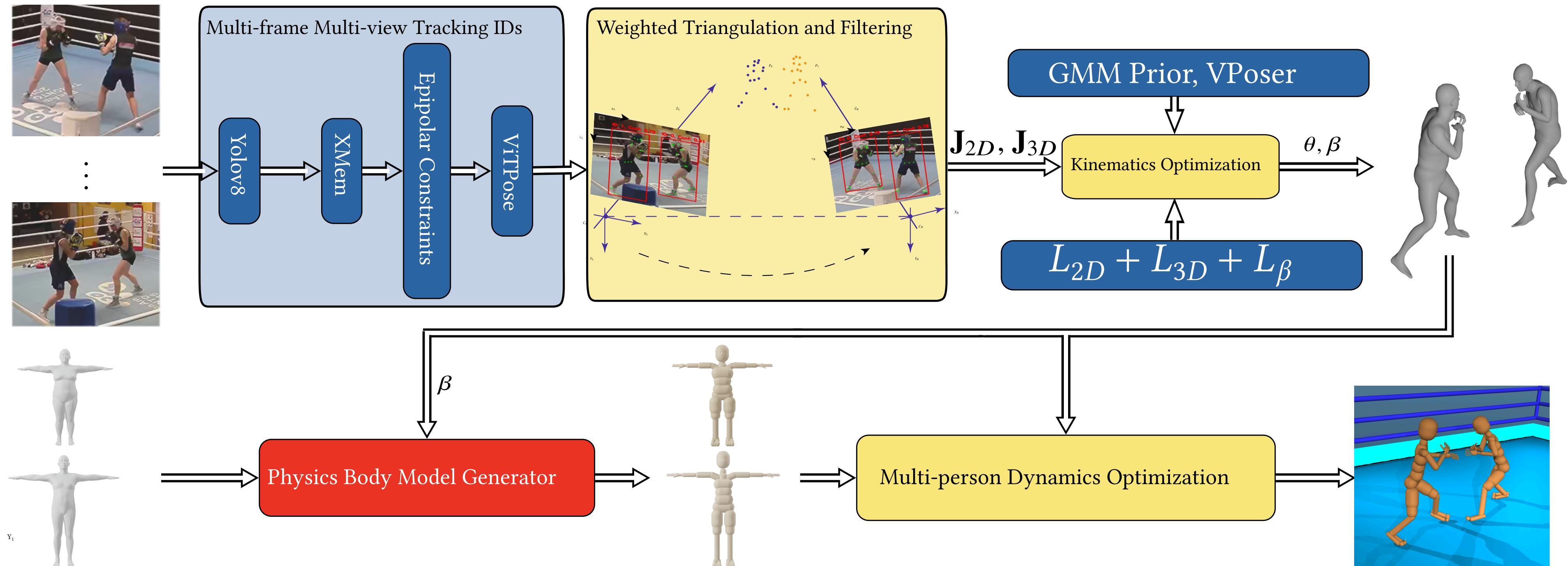
¹  cole de technologie sup  rieure, Montr  al, Qu  bec, Canada



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



Multi-frame Multi-view Tracking IDs

- **Tracking by Segmentation:** XMem is used for consistent ID assignment across frames by segmenting individuals in each view. [1]
- **Epipolar Constraint-based ID Matching:** IDs are aligned across multiple views by processing bounding box centroids with epipolar constraints. [4]
- **Top-Down 2D Pose Estimation:** This approach involves detecting individuals first and then localizing keypoints within the bounding boxes, ensuring precise and scalable pose estimation, particularly effective in challenging scenarios like low-resolution videos or fast movements. [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

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

\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

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

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

$$L_{3D} = \sum_{i \in \mathcal{J}_{3D}} c_i \|\mathbf{J}_i(\theta, \beta) - \mathbf{J}_{3D,i}\|^2, \quad (4)$$

$$L_{smooth} = \sum_t \|\theta^t - \theta^{t-1}\|^2 + \|\mathcal{M}(\theta^t, \beta) - \mathcal{M}(\theta^{t-1}, \beta)\|^2. \quad (5)$$

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

Physics Body Model Generator

Since our goal is to reconstruct the motion of K physics-based humanoids at each time step t , we define the combined state of all humanoids as $\mathbf{x}_t = (\mathbf{q}_t, \mathbf{v}_t)$, where $\mathbf{q}_t \in \mathbb{R}^{63K}$ and $\mathbf{v}_t \in \mathbb{R}^{62K}$ represent the rotation and velocity of the joints. Whereas \mathbf{q}_t is a minimal coordinate representation of position-level coordinates, we also compute $\mathbf{p}_t \in \mathbb{R}^{21K}$ which are the 3D positions of the SMPL joints relative to the humanoid's body frame of reference. This vector is used for the optimization process. we are using xml modeling of Mujoco [3] for our humanoids.

Multi-person Dynamics Optimization

The iLQR algorithm [2] optimizes a control trajectory $\mathbf{u}_{0:T}$ by minimizing the cost function:

Dynamics Optimization Formulation

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

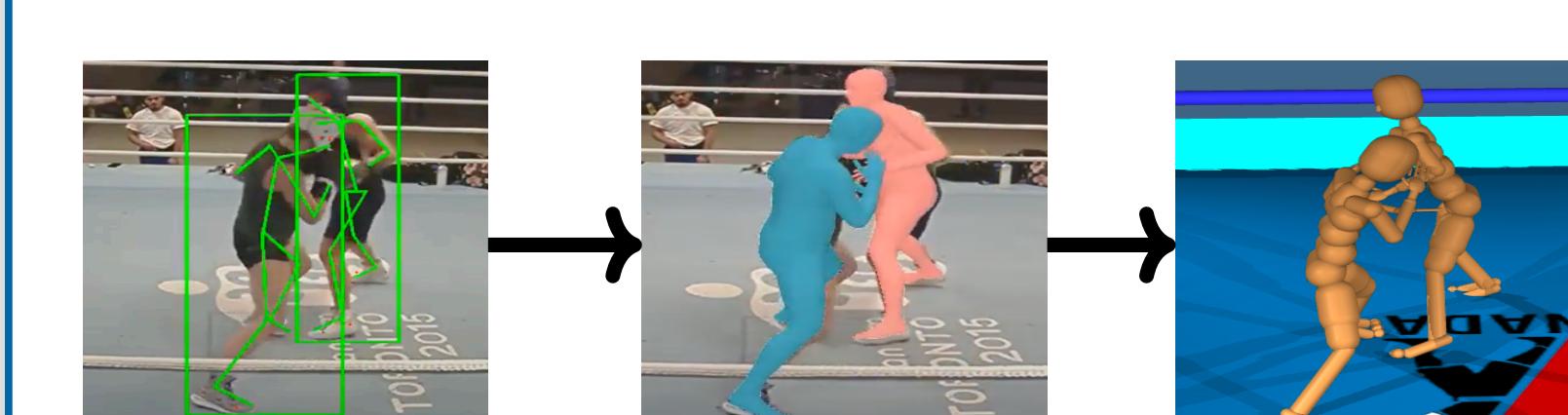
The reference positions \mathbf{p}_t and velocities \mathbf{v}_t are extracted from the kinematics optimization at each frame t , and the trajectory optimization is "warm-started" using control parameters \mathbf{u}_{t-1} from the previous frame.

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.

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.

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 W  hler. *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).