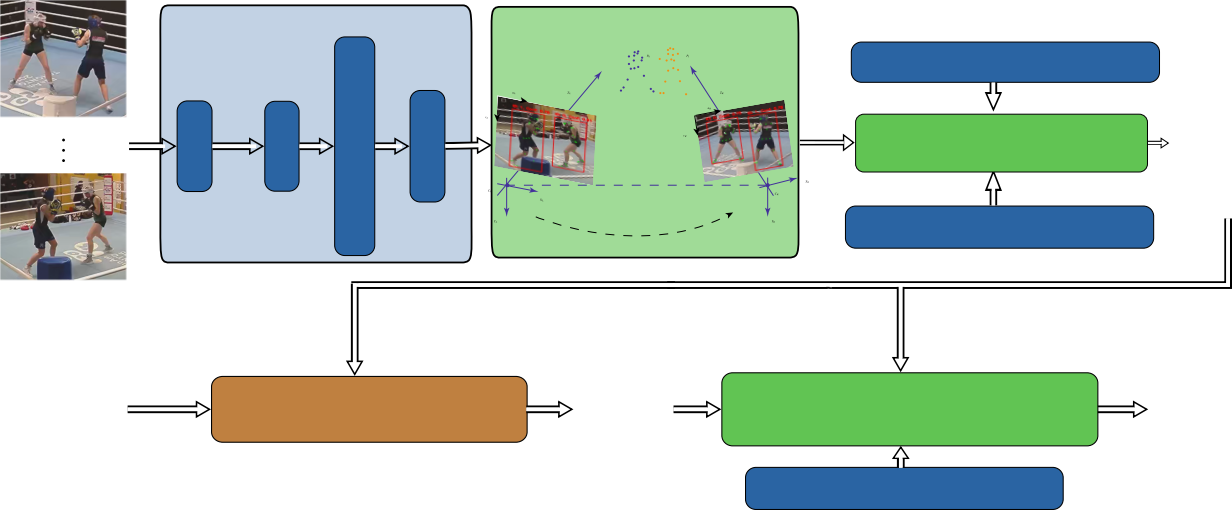


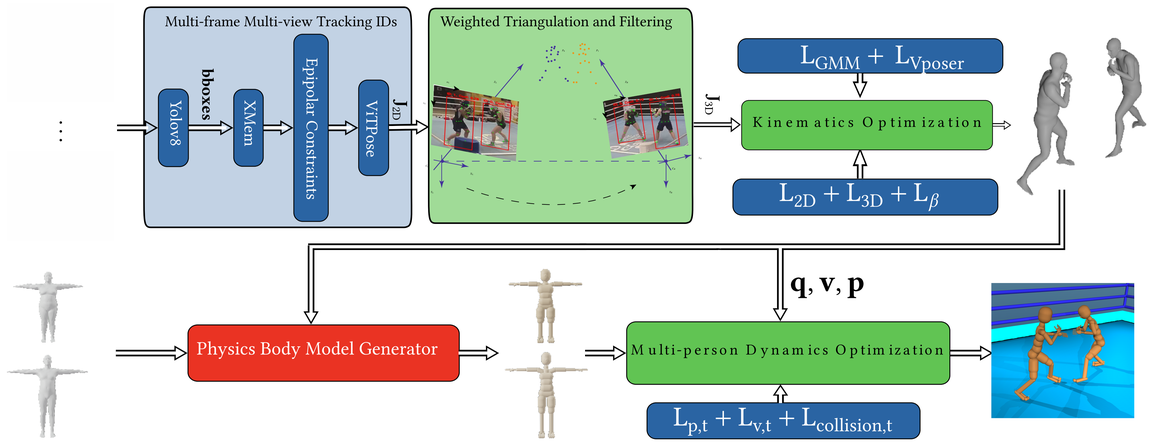
Multi-person Physics-based Pose Estimation for Combat Sports

Hossein Feiz David Labbé Sheldon Andrews
École de technologie supérieure, Montréal, Québec, Canada

Introduction

Motion capture in sports scenes like boxing presents challenges due to occlusion, crowding, and fast movements. We used a multi-view setup requiring visibility from just two cameras. Our pipeline create realistic motions using multi-person physics optimization considering the collisions between characters.

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

Weighted Triangulation and Filtering

Weighted triangulation estimates 3D keypoints from corresponding 2D keypoints across all N cameras, accurately determining the 3D positions by solving the linear system:

Weighted Triangulation Formulation

$$\begin{bmatrix} \mu_1(\mathbf{P}_{11} - u_1\mathbf{P}_{31}) \\ \mu_1(\mathbf{P}_{21} - v_1\mathbf{P}_{31}) \\ \vdots \\ \mu_N(\mathbf{P}_{1N} - u_N\mathbf{P}_{3N}) \\ \mu_N(\mathbf{P}_{2N} - v_N\mathbf{P}_{3N}) \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \quad (1)$$

$\mathbf{P}_{ij}$ extracts the i-th row of the projection matrix $\mathbf{P}_j$. The average confidence of a 2D point for the current frame in camera j is μ_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 triangulation fails, we apply an extended Kalman filter to estimate 3D keypoints using velocity, acceleration, and position constraints, ensuring reliable reconstructions from sparse, noisy 2D data.

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

- **2D Re-projection Loss**: Aligns the 3D joints with 2D joints using re-projection, weighted by confidence.
  $L_{2D} = \sum_j \sum_{i \in J_{2D}} c_{j,i} \rho(J_{\text{proj}\,ji} - J_{2D\,ji})$
- **3D Alignment Loss**: Minimizes the difference between predicted and actual 3D joint positions considering confidence.
  $L_{3D} = \sum_{i \in J_{3D}} c_i \| J_i(\theta, \beta) - J_{3D,i} \|^2$
- **Smoothness Loss**: Ensures temporal consistency by minimizing differences in poses and vertices over time.
  $L_{smooth} = \sum_t \| \theta^t - \theta^{t-1} \|^2 + \| \mathcal{M}(\theta^t, \beta) - \mathcal{M}(\theta^{t-1}, \beta) \|^2$
- **Prior Loss**: Guides the optimization towards natural poses using GMM and Vposer priors.
  $L_{GMM} = \frac{1}{N} \sum_{i=1}^{N} \text{GMM}(\theta_i, \beta)$, $L_{Vposer} = \frac{1}{N} \sum_{i=1}^{N} (z(\theta_i)^2)$

Physics Body Model Generator

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

Multi-person Dynamics Optimization

The iLQR algorithm [2] optimizes a control trajectory $\mathbf{u}_{0:T}$ by minimizing the cost function. 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: $\min_{\mathbf{u}_{0:T}} \sum_{t=0}^{T} w_1 L_{reg,t} + w_2 L_{p,t} + w_3 L_{v,t} + w_4 L_{collision,t}$

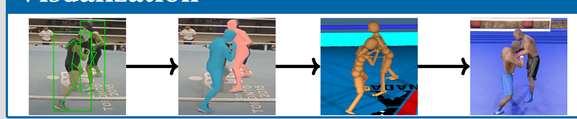
- **Position Loss**: Minimizes the difference between predicted and actual 3D joint positions in each short horizon.
  $L_{p,t} = \sum_{i \in J_{3D}} \| p_{i,t} - J_{3D,i,t} \|^2$
- **Velocity Loss**: Minimizes the difference between predicted and actual 3D joint velocities in each short horizon.
  $L_{v,t} = \sum_{i \in J_{3D}} c_i \| v_{i,t} - v_{3D,i,t} \|^2$
- **Collision Loss**: The collision term penalizes the penetration between the geoms of different humanoids.
  $L_{collision,t} = \sum_{i \in 1...M} \sum_{j \in 1...M} \max(0, \epsilon - \text{dist}(G1_i, G2_j))$
- **Regularization Loss**: Prevent sudden changes in control input. $L_{reg,t} = \| \mathbf{u}_t \|^2$

Discussion and Future Works

We created a 10-hour dataset of boxing motions to develop a 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 hosseinfeiz1