Markerless Multi-person Pose Estimation and Motion Synthesis for Combat Sports

Hossein Feiz[1], David Labbé[2], Sheldon Andrews[3]

École de technologie supérieure (ETS), Montreal, Canada

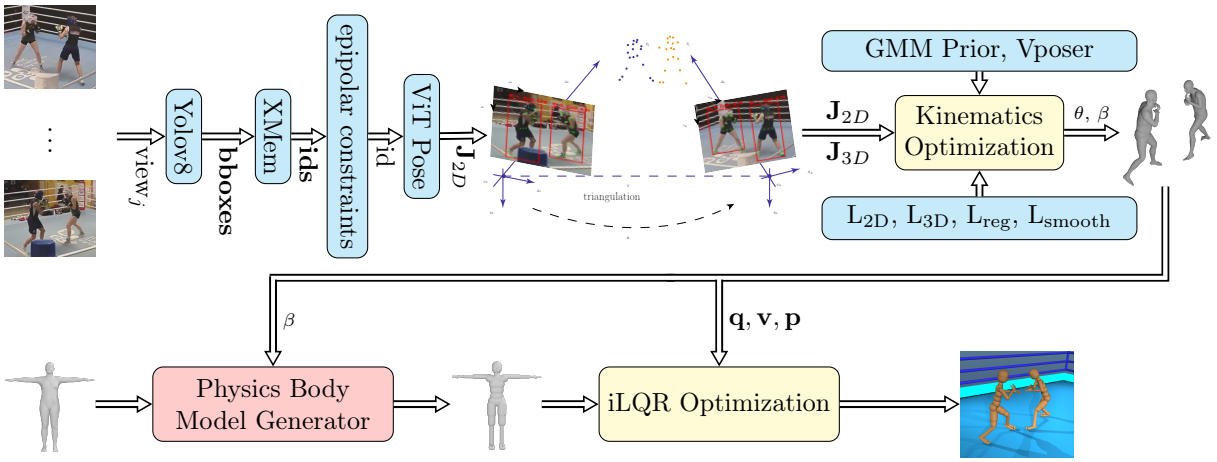

Figure 1: *The pipeline begins with generating bounding boxes and robust tracking for each individual in the scene. These tracking results are used to produce 2D poses. The triangulation process, produces smooth 3D keypoints. The kinematics optimization step incorporates the 2D and 3D keypoints, to create 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.*

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. Furthermore, leveraging interaction datasets attain from multi-view setup, the pipeline supports multi-person motion synthesis. It employs a seq2seq model to generate interactions from sparse VR headset inputs, and a diffusion prior model to generate the whole body poses. Finally, latent optimization ensures motions adhere to motion prior and satisfies high level criteria by producing controllable animations, suitable for combat sports analysis and training applications.

CCS Concepts

• ***Computing methodologies*** → *Pose Estimation; Motion synthesis;*

1. 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 fo-

cuses 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:

- **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.

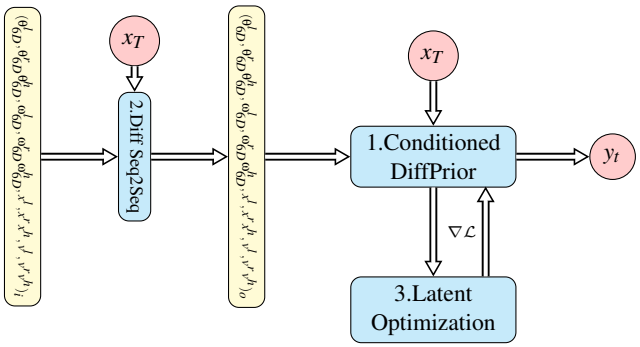

Figure 2: *The condition for the seq2seq model contains the VR sensor information of the player and the output is the corresponding VR sensor input for the opponent at each sequence. The conditional diffusion prior is conditioned on the opponent synthetic information and produces the whole body pose for the opponent. The optimization in loop make sure the output motion has the desired trajectories based on a loss function.*

Figure 2 which need the full textwidth can be typeset as Figure.

Dynamics Optimization: In the described study, a two-stage optimization approach is employed to enhance the accuracy and quality of motion tracking for athletes. The initial stage focuses on kinematic optimization, which estimates full-body poses but often introduces high-frequency artifacts known as jitter. 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, which optimizes control inputs to smooth and stabilize motion trajectories while adhering to physical realism. This approach accounts for contact forces and body dynamics, enhancing the overall quality and naturalness of the generated motions.

Furthermore, trajectory optimization in this framework minimizes a weighted sum of regularization terms and loss functions, which include penalties for velocity and control values. By iteratively refining control trajectories over short time horizons, the method achieves improved alignment with desired motion characteristics derived from the kinematic stage outputs. Overall, this integrated approach ensures robust and realistic motion tracking suitable for applications in sports biomechanics and animation.

2. Motion Synthesis

Interaction Diffusion Model:

A novel application of sequence-to-sequence (seq2seq) models in virtual reality involves generating responsive opponent motions based on sparse VR input data. In this context, the seq2seq model serves as a pivotal tool for translating sparse signals captured from VR sensors into realistic and dynamic movements for virtual opponents. Trained on a specialized multi-person boxing dataset, the model learns to map the sparse input, which could include data from wearable IMUs or HMDs, into coherent sequences of actions that simulate human-like behavior during a boxing match. By leveraging the dataset's diverse scenarios and real-world motion variations, the seq2seq model not only predicts the opponent's reactions but also adapts to different player actions, enhancing the realism and engagement of VR gaming experiences. This approach marks a significant advancement in interactive VR applications by bridging the gap between sparse sensor data and immersive virtual interactions, thereby enriching the user experience through responsive and lifelike virtual opponents. **Conditioned Diffprior:** The application of conditional diffusion models, specifically tailored for full-body motion synthesis from sparse IMU tracking signals, represents a breakthrough in virtual reality (VR) body tracking. This approach overcomes the inherent challenges of accurately predicting smooth and realistic full-body motions with minimal input data. By leveraging a lightweight MLP architecture and employing a block-wise injection scheme for embedding time step information, the conditional diffusion model effectively reduces artifacts like jittering and enhances robustness against signal loss. These advancements are underscored by its superior performance on benchmarks like the

AMASS dataset, setting new standards in motion prediction methods.

In parallel, the conditional diffusion model paradigm also finds application in frameworks designed for real-time human motion tracking in VR environments. This innovative approach integrates scalable 3DOF IMUs with head-mounted displays (HMDs), offering a flexible solution that balances tracking accuracy with user-friendliness across various VR setups. Central to its efficacy is a lightweight temporal-spatial feature learning (TSFL) network, combining LSTM for temporal feature capture and Transformer for spatial correlation learning. This hybrid architecture optimizes computational efficiency while maintaining high accuracy in estimating full-body poses, making it suitable for immersive VR experiences demanding fluid and responsive interactions.

Latent Optimization:

Optimize noisy motion inputs directly in the latent space of the diffusion model. By treating motion denoising as a black box and iteratively adjusts the diffusion noise vector using gradients computed through an ODE solver, ensuring the output motion meets user-defined criteria. This approach avoids the need for fine-tuning models for specific tasks, demonstrating superior performance in motion editing, preservation, and task fulfillment compared to existing methods.