Multi-person Physics-based Pose Estimation for Combat Sports

Supplementary Material

5.1. Evaluation Metrics

We evaluate our approach using two main categories of metrics: **pose metrics** and **physics metrics**.

Pose Metrics. These measure pose accuracy and smoothness:

- **WA-MPJPE** (World-Aligned MPJPE). In some works, this refers to an MPJPE calculation where the predicted and ground-truth poses are aligned in the global (or "world") coordinate system by a single rigid transformation (often estimated via a Procrustes-like method) *once* for the entire sequence. It does not allow a different transformation per frame.
- **W-MPJPE** (World PA First-MPJPE). This measures MPJPE after aligning *only the first frame* of the prediction and ground truth in the global coordinate system. The remaining frames are evaluated without further alignment, highlighting any drift that occurs over time.
- **PA-MPJPE** (Procrustes Aligned MPJPE). Each frame of the predicted pose is rigidly aligned to the corresponding ground-truth pose on a per-frame basis before computing MPJPE. This alignment factors out any frame-by-frame rigid transformations, emphasizing errors in limb positions or shape.
- **Acceleration Error (Acc. Error).** Evaluates discrepancies in acceleration between the predicted motion and ground truth, indicating whether the motion is jittery or overly smoothed.

Physics Metrics. These assess how physically plausible a motion sequence is:

- **Foot Skating (Skating).** Evaluates the extent to which feet slide on the ground when they should remain still.
- **Ground Penetration (Gnd Pen.).** Measures whether any body part intersects or penetrates the ground plane.
- **Inter-Person Penetration (Pen.).** Assesses collisions among different individuals by computing signed distance function (SDF) values. For each person, the SDF values from penetrating vertices of others are accumulated over all frames and averaged per person. This metric is measured in millimeters, whereas the other physics metrics are measured in meters.

Below are further details on the individual metrics:

5.1.1. Percentage of Correct Parts (PCP)

The PCP metric quantifies the percentage of body parts whose endpoints are correctly localized. Let the ground-truth body part endpoints be $\mathbf{s}_c$ and $\mathbf{e}_c$, and let $\mathbf{s}_n$ and $\mathbf{e}_n$ be their estimated counterparts. The part is considered correct if:

$$|\mathbf{s} - \mathbf{s}_c| + |\mathbf{e} - \mathbf{e}_c| \;\leq\; |\mathbf{s}_n - \mathbf{e}_n|\,. \tag{8}$$

5.1.2. Mean Per Joint Position Error (MPJPE)

The MPJPE metric computes the mean Euclidean distance between predicted and ground-truth 3D joint positions:

$$\mathrm{e}_{MPJPE} = \mathrm{mean}\left\|\hat{\mathbf{p}}_t - \mathbf{p}_t^{GT}\right\|. \tag{9}$$

5.1.3. Foot Floating and Sliding

Foot Floating. We compare the vertical position of the foot in the predicted motion to the ground truth:

$$\mathrm{e}_{foot,z} = \mathrm{mean}\left\|\hat{\mathbf{p}}_{foot,z} - \mathbf{p}_{foot,z}^{GT}\right\|. \tag{10}$$

Foot Sliding. We compare the foot's horizontal velocity (on the XY plane) in the predicted motion to the ground truth:

$$\mathrm{e}_{foot,v_{xy}} = \mathrm{mean}\left\|\Delta\hat{\mathbf{p}}_{foot,v_{xy_t}} - \Delta\mathbf{p}_{foot,v_{xy_t}}^{GT}\right\|. \tag{11}$$

5.1.4. Motion Smoothness

Smoothness is measured by (1) computing MPJPE on a 15-joint reduced skeleton plus the root, and (2) comparing keypoint velocities:

$$\hat{\mathbf{Jit}} = \left\|\hat{\mathbf{p}}_t - \hat{\mathbf{p}}_{t-1}\right\|, \tag{12}$$

$$\mathbf{Jit}^{GT} = \left\|\mathbf{p}_t^{GT} - \mathbf{p}_{t-1}^{GT}\right\|, \tag{13}$$

$$\mathrm{e}_{smooth} = \mathrm{mean}\left\|\mathbf{Jit}^{GT} - \hat{\mathbf{Jit}}\right\|. \tag{14}$$

5.2. Ablation Study

We conduct ablation experiments to illustrate the impact of key components in our framework.

Number of Views. Reducing camera views typically degrades 3D reconstruction due to weaker triangulation. Nevertheless, our method retains robustness in Table 6, even when views are sparse.

Triangulation Filtering. Spline smoothing and an Extended Kalman Filter (EKF) on keypoints significantly denoise the initial linear triangulation, improving subsequent kinematic optimization (Table 6).

Prior (GMM + Vposer). A Gaussian Mixture Model (GMM) prior in combination with Vposer narrows the pose parameter search space to more plausible solutions.

Table 6. Ablation Study on Shelf Dataset using PCP metric.

Method	Actor 1	Actor 2	Actor 3	Average
Baseline	99.8	97.6	98.6	98.6
Using 3 Views	98.3	98.1	98.6	98.3
Using 2 Views	97.6	48.1	96.4	80.7
No Filtering for Triangulation	99.8	94.9	97.9	97.5
No Prior	99.8	96.2	98.1	98.0
No 2D loss	91.6	76.2	81.2	83.0
No 3D loss	99.8	96.2	98.1	98.0
No pose regularization	99.8	95.7	98.5	98.0

Table 7. Configuration of the Humanoid used in physics optimization.

Joint	Force Range	Angle Range (rad)	Joint Type	No. Axes	Damping
Head	-40 to 40	-0.524 to 0.524	Hinge	3	20
Clavicle	-80 to 80	-0.349 to 0.349	Hinge	2	20
Femur	-300 to 300	-2.792 to 1.221	Hinge	3	20
Tibia	-160 to 160	0.01 to 2.967	Hinge	1	16
Foot	-120 to 120	-0.785 to 1.222	Hinge	2	12
Hand	-20 to 20	-1.57 to 1.57	Hinge	2	1
Humerus	-120 to 120	-1.571 to 1.571	Hinge	3	6
Lower Back	-300 to 300	-0.523 to 0.523	Hinge	3	24
Lower Neck	-120 to 120	-0.5236 to 0.5236	Hinge	3	40
Radius	-90 to 90	-2.967 to 0.174	Hinge	1	5
Toes	-20 to 20	1.570 to 0.349	Hinge	1	1
Wrist	-20 to 20	-3.14159 to 0	Hinge	1	1
Root	-	-	Free	-	-

Regularization. Pose regularization encourages smooth transitions between frames, reducing sudden jumps or unnatural artifacts.

5.3. CMU Humanoid Details

Table 7 shows the joint configurations for the CMU humanoid employed in physics optimization. Although the physics engine treats joints with three rotation axes as ball joints, we constrain these axes to mimic the behavior of *three consecutive hinge joints*, providing finer rotational control.

In summary, our pipeline combines noise-robust triangulation, strong pose priors, and physics-based optimization to yield smooth, physically plausible human motions under diverse scenarios, including sparse camera setups and noisy estimates.