STRIKE: A Framework for Smoothing High-Impact Martial Arts Motion

1st Alexis Pereira Ferreira Pinto
Centro de Ciências Matemáticas e da Terra - CCMN
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brasil
alicepfp@labnet.nce.ufrj.br

***Abstract*—This work investigates the use of the Unscented Kalman Filter (UKF) to improve the accuracy and temporal consistency of joint detection in Muay Thai videos processed on edge devices. By applying the UKF to noisy keypoints estimated by lightweight YOLO-based pose detectors, we observe reduced mean per joint position error (MPJPE) on benchmark datasets, particularly in sequences where motion is smooth but detection jitter is significant. Our method demonstrates that fusing temporal dynamics with lightweight detection improves pose stability on resource-constrained devices, enabling practical applications such as automated scoring and referee assistance in combat sports.**

***Index Terms*—Pose Estimation, Unscented Kalman Filter, Computer Vision, Sports Analytics, Muay Thai, Edge Computing**

I. INTRODUCTION

In this paper, we introduce the STRIKE (Smoothed Tracking & Recognition of In-fight Kinetic Engagement) framework, designed to enhance motion analysis in martial arts. Real-time human pose estimation plays an increasingly important role in sports analytics, training feedback, and referee support systems [9], [11]. In martial arts disciplines like Muay Thai, characterized by rapid, high-impact strikes and complex defensive maneuvers, accurate joint tracking is essential for an objective evaluation of technique. Deploying such systems on edge devices is attractive for in-situ analysis but typically requires lightweight pose estimation models [7], [8], which often produce noisy or incomplete detections, particularly under fast motion, occlusion, or suboptimal lighting.

This paper explores how temporal filtering, specifically the Unscented Kalman Filter (UKF), can reduce detection noise while remaining computationally efficient. By leveraging the kinematic continuity of human motion, the UKF smooths predictions from YOLO-based detectors, yielding trajectories that better reflect the true underlying motion.

II. BACKGROUND

Recent years have seen significant progress in multimodal and sensor-based human action recognition [2], [3], [6]. Techniques like hybrid deep learning [1] and multi-sensor fusion [7], [8] achieve robust performance by integrating diverse data streams. Concurrently, advances in detailed action understanding have pushed the boundaries of what can be inferred from motion [12]. However, vision-only systems remain appealing for simplicity and lower deployment cost [10], [11].

In the context of combat sports, research such as [4] highlights the challenge of noisy labels and fast, complex motion, which degrades frame-wise pose detection accuracy. Prior works often address this with computationally heavy spatiotemporal networks [10], but these are unsuitable for edge devices. Our approach differs by coupling a lightweight detector with a classic, efficient Bayesian filter. The UKF extends the standard Kalman Filter (KF) and Extended Kalman Filter (EKF) by better handling the non-linearities inherent in human motion models without requiring analytic Jacobians, making it an effective lightweight solution.

III. METHODOLOGY

We propose a two-stage pipeline. First, we extract 2D joint positions frame by frame using a lightweight YOLO-based pose estimator [13]. Then, we apply a separate Unscented Kalman Filter per joint to exploit temporal coherence.

The UKF approximates the posterior distribution by transforming a deterministic set of “sigma points” through the nonlinear motion model $f(\cdot)$ and measurement model $h(\cdot)$. The filter maintains a state vector $x = [x, y, v, \theta, \omega]^\top$, where (x, y) are joint positions, v is velocity, θ is heading, and ω is turn rate. The UKF operates in a predict-update cycle, as summarized in Algorithm 1.

A. Mathematical Formulation of the UKF

The Unscented Kalman Filter (UKF) for each joint estimates the state vector

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ v \\ \theta \\ \omega \end{bmatrix} \tag{1}$$

where (x, y) is the 2D joint position, v is velocity, θ is heading, and ω is turn rate.

Given the prior mean and covariance $(\mathbf{x}_{k-1}, \mathbf{P}_{k-1})$, the UKF proceeds as:

1) Generate $2n + 1$ sigma points:

$$\chi_0 = \mathbf{x}_{k-1} \tag{2}$$

$$\chi_i = \mathbf{x}_{k-1} + [\sqrt{(n+\lambda)\mathbf{P}_{k-1}}]_i, \quad i = 1, \dots n \tag{3}$$

$$\chi_{i+n} = \mathbf{x}_{k-1} - [\sqrt{(n+\lambda)\mathbf{P}_{k-1}}]_i, \quad i = 1, \dots n \tag{4}$$

with $\lambda = \alpha^2(n+\kappa) - n$.

2) Predict each sigma point by the nonlinear motion model:

$$x_k = x_{k-1} + \frac{v_{k-1}}{\omega_{k-1}} \left(\sin(\theta_{k-1} + \omega_{k-1}\Delta t) - \sin(\theta_{k-1}) \right) \tag{5}$$

$$y_k = y_{k-1} - \frac{v_{k-1}}{\omega_{k-1}} \left(\cos(\theta_{k-1} + \omega_{k-1}\Delta t) - \cos(\theta_{k-1}) \right) \tag{6}$$

For small $|\omega|$, use the linear approximation:

$$x_k \approx x_{k-1} + v_{k-1}\Delta t \cos(\theta_{k-1}) \tag{7}$$

$$y_k \approx y_{k-1} + v_{k-1}\Delta t \sin(\theta_{k-1}) \tag{8}$$

3) Reconstruct predicted mean and covariance:

$$\hat{\mathbf{x}}^- = \sum_{i=0}^{2n} W_m^{[i]} \chi_i^- \tag{9}$$

$$\mathbf{P}^- = \sum_{i=0}^{2n} W_c^{[i]} (\chi_i^- - \hat{\mathbf{x}}^-)(\chi_i^- - \hat{\mathbf{x}}^-)^T + \mathbf{Q} \tag{10}$$

4) Update: transform predicted sigma points with the measurement model $h(\cdot)$ (extracting (x, y)):

$$\gamma_i = h(\chi_i^-) = [x, y]^T \tag{11}$$

$$\hat{\mathbf{z}} = \sum_{i=0}^{2n} W_m^{[i]} \gamma_i \tag{12}$$

$$\mathbf{S} = \sum_{i=0}^{2n} W_c^{[i]} (\gamma_i - \hat{\mathbf{z}})(\gamma_i - \hat{\mathbf{z}})^T + \mathbf{R} \tag{13}$$

The cross-covariance and Kalman gain:

$$\mathbf{T} = \sum_{i=0}^{2n} W_c^{[i]} (\chi_i^- - \hat{\mathbf{x}}^-)(\gamma_i - \hat{\mathbf{z}})^T \tag{14}$$

$$\mathbf{K} = \mathbf{T}\mathbf{S}^{-1} \tag{15}$$

Apply the update:

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}(\mathbf{z}_k - \hat{\mathbf{z}}_k) \tag{16}$$

$$\mathbf{P}_k = \mathbf{P}_k^- - \mathbf{K}\mathbf{S}_k\mathbf{K}^T \tag{17}$$

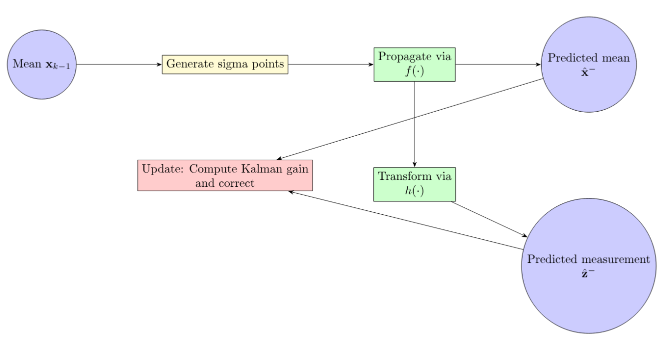

Fig. 1. The Unscented Kalman Filter Process Flow. Sigma points generated from the previous state estimate are propagated through the motion model $f(\cdot)$ to produce a predicted mean. These points are also transformed by the measurement model $h(\cdot)$ to compute the Kalman gain for the correction step.

Algorithm 1 Pose Tracking with YOLO and Unscented Kalman Filter

Require: Video Frame Sequence $V = \{v_1, v_2, ..., v_N\}$
Require: UKF Process Noise Q, Measurement Noise R
Ensure: Filtered Keypoint Sequence $\hat{\mathbf{X}} = \{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, ..., \hat{\mathbf{x}}_N\}$

```
Initialize UKF for each joint j ∈ {1, ..., J}: F = {UKF_1, ..., UKF_J}
Run YOLO on v_1 to get initial keypoints z_1
for each joint j = 1 to J do
    Initialize x̂_{1|1,j} and P_{1|1,j} with z_{1,j}
end for
Store X̂_1 ← {x̂_{1|1,1}, ..., x̂_{1|1,J}}
for k = 2 to N do
    Run YOLO on v_k to get keypoints z_k
    if person detected in v_k then
        for each joint j = 1 to J do
            Predict: x̂_{k|k-1,j}, P_{k|k-1,j} with f(·) and Q
            Update: x̂_{k|k,j}, P_{k|k,j} with z_{k,j}, h(·), R
        end for
    else
        for each joint j = 1 to J do
            Predict only
        end for
    end if
    Store X̂_k
end for
return X̂ =0
```

IV. EXPERIMENTS AND RESULTS

A. Parameter Tuning for Filtering

For the Penn Action benchmark, we used:

- **Time step:** $dt_{\text{ukf}} = 1.0/\text{fps}$, where fps is from annotations or defaulted to 60.
- **Process noise:** $q_{\text{pos}} = 1.5$, $q_{\text{vel}} = 2.0$, $q_{\text{turn}} = 2.5$, $Q = \text{diag}(q_{\text{pos}}, q_{\text{pos}}, q_{\text{vel}}^2, 0, q_{\text{turn}}^2)$.
- **Measurement noise:** $r_{\text{val}} = 89.3$, $R = r_{\text{val}} \cdot I_2$.

These settings reflect moderate process uncertainty for sports movements as captured in a well-lit, labeled dataset. The Q parameters allow the filter to adapt to both smooth and variable velocities, and R matches the observed YOLO detection noise.

For real-world video, process noise coefficients are increased ($q_{\text{pos}} = 9.0$, $q_{\text{vel}} = 11.0$, $q_{\text{turn}} = 13.0$) to better adapt to rapid, unpredictable motion and occasional loss of detection.

B. Evaluation Metrics: MPJPE and F1-Score

MPJPE: The principal accuracy metric is mean per joint position error:

$$\text{MPJPE} = \frac{1}{NF} \sum_{f=1}^{F} \sum_{n=1}^{N} \left\| \hat{\mathbf{p}}_{f,n} - \mathbf{p}_{f,n}^{\text{GT}} \right\|_2 \tag{18}$$

F1-Score:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{19}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{20}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

where TP, FP, FN are counts of visible joint detections with respect to thresholded pixel error (25px).

C. Penn Action Benchmark Results

TABLE I
UKF Benchmarking Results on Penn Action Dataset

Metric	**Value**
Total Sequences Processed	2326
Mean Raw YOLO MPJPE	76.06 pixels
Mean Filtered MPJPE	75.37 pixels
Accuracy Improvement	0.91%
Mean F1-Score Raw YOLO	0.4280
Mean F1-Score UKF	0.4221

TABLE II
Sequences Improved by UKF

Metric	**Value**
Sequences Improved MPJPE	1135 of 2326 (48.8%)
Avg. MPJPE Improvement (in improved)	2.35%
Sequences Improved F1	747 of 2326 (32.1%)
Avg. F1 Improvement (in improved)	6.91%

TABLE III
Top 10 Sequences by MPJPE Improvement

Sequence ID	Raw MPJPE	Filtered MPJPE	MPJPE Impr. (%)
2239	170.73	13.13	92.31
2060	230.27	71.25	69.06
2001	32.84	10.77	67.20
1454	147.33	55.96	62.02
1459	189.00	84.02	55.54
1492	159.17	99.30	37.61
0434	221.19	139.25	37.05
0246	28.81	18.78	34.81
2240	163.87	110.89	32.33
0544	62.17	43.43	30.14

D. Qualitative and Real-World Analysis

Figure 1 and Table I demonstrate that the UKF provides a modest but real reduction in spatial error (MPJPE) and substantial outlier improvement in certain sequences. However, in about half the sequences, the filter did not further reduce the error—usually due to very rapid, non-linear movements that break the constant-velocity turn rate assumption.

TABLE IV
Top 10 Sequences by F1-Score Improvement

Sequence ID	Raw F1	Filtered F1	F1 Impr. (%)
1542	0.0139	0.0643	362.59
0373	0.0543	0.2255	315.29
1821	0.0057	0.0183	221.05
1934	0.0276	0.0719	160.51
0374	0.0134	0.0301	124.63
1802	0.0210	0.0449	113.81
1459	0.2313	0.4825	108.60
1461	0.0137	0.0253	84.67
1544	0.0122	0.0213	74.59
1557	0.2722	0.4691	72.34

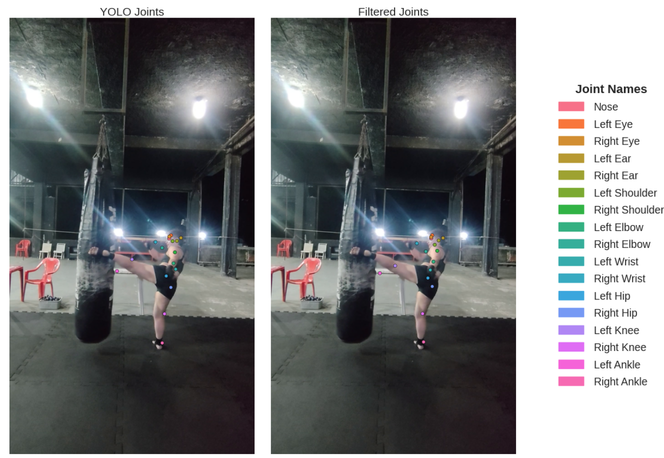

Fig. 2. Example filtering on a real-world Muay Thai training video. Left: YOLO keypoints (raw), Right: UKF-smoothed keypoints, Color: Joint type. The filter reduces temporal jitter and creates more plausible joint trajectories, especially for striking limbs and upper body.

E. Limitations

While the UKF improves performance on average, it can lag or underperform during sudden ballistic actions (e.g., kicks), demonstrating the limitation of the underlying ”constant turn rate and velocity” motion model during high-acceleration maneuvers.

V. Conclusion

This study demonstrates that lightweight temporal filtering with the UKF is a viable, efficient method for enhancing pose tracking reliability using small YOLO models on edge devices. By exploiting temporal coherence, it reduces frame-to-frame jitter and measurably improves accuracy in many sequences. Some limitations remain for athletic disciplines featuring explosive or non-linear movements. Future work will adapt richer dynamical models, extend the framework to full 3D, and integrate the method into real-time referee and coaching tools for martial arts.

References

[1] M. Tayyab et al., “A Hybrid Approach for Sports Activity Recognition Using Key Body Descriptors and Hybrid Deep Learning Classifier,” *Sensors*, vol. 25, no. 2, p. 441, Jan. 2025.

[2] X. Chao, Z. Hou, and Y. Mo, "CZU-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and 10 Wearable Inertial Sensors," *IEEE Sensors J.*, vol. 22, no. 7, pp. 7034-7042, Apr. 2022.

[3] X. Feng et al., "DAMUN: A Domain Adaptive Human Activity Recognition Network Based on Multimodal Feature Fusion," *IEEE Sensors J.*, vol. 23, no. 18, pp. 22019-22030, Sept. 2023.

[4] S. Yamanaka et al., "Evaluation System for Martial Arts Demonstration from Smartphone Sensor Data Using Deep Neural Networks on Noisy Labels," in *2023 IEEE Int. Conf. on Consumer Electronics (ICCE)*, Las Vegas, NV, USA, Jan. 2023.

[5] J. Kamminga et al., "M-MOVE-IT: Multimodal Machine Observation and Video-Enhanced Integration Tool for Data Annotation," in *Proc. of the 2024 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing*, Melbourne, Australia, Oct. 2024.

[6] S. K. Yadav et al., "MS-KARD: A Benchmark for Multimodal Karate Action Recognition," in *2022 Int. Joint Conf. on Neural Networks (IJCNN)*, Padua, Italy, Jul. 2022.

[7] B. Jayakumar and N. Govindarajan, "Multi-sensor fusion based optimized deep convolutional neural network for boxing punch activity recognition," *J. Sports Eng. Technol.*, Mar. 2024.

[8] B. Zhou, "Sanda training program based on multi-sensor data fusion in the internet of things environment," *Internet Technology Letters*, vol. 7, no. 2, p. e470, Mar. 2024.

[9] I. Ghosh et al., "Sports analytics review: Artificial intelligence applications, emerging technologies, and algorithmic perspective," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 5, Sept. 2023.

[10] Y. Sun et al., "ST-LineNet: A spatiotemporal network for real-time 3D Pose estimation in martial arts training," *Alexandria Eng. J.*, vol. 117, pp. 136-147, Apr. 2025.

[11] R. Gade et al., "The (Computer) Vision of Sports: Recent Trends in Research and Commercial Systems for Sport Analytics," in *Computer Vision*. Boca Raton: Chapman and Hall/CRC, 2024.

[12] W. Zhang, M. Zhu, and K. Derpanis, "From Actemes to Action: A Strongly-supervised Representation for Detailed Action Understanding," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013.

[13] G. Jocher and J. Qiu, "Ultralytics YOLO11," version 11.0.0, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics