Large-Scale 3D Pose Estimation of Professional Tennis Serves from Broadcast Video

Jason Wang[1] Robert Chen[1] Patrick Ho[1] Emmy Kim[1]
Samuel Min[1] Jaden Shim[1] Vrishak Vemuri[1] Derek Wang[1]
Natalie Kupperman[1,†] Stephen Baek[1,†]
[1]University of Virginia [†]Corresponding authors
{jyw5hw, tcq2ms, bqu3tr, tvc5eq, xxz2ku,
atx4wg, mxw4hg, ant8zr, nak5dy, mwn4yc}@virginia.edu

Abstract

Traditional biomechanical research in sports science relies on marker-based motion capture systems that, despite high precision, are constrained by laboratory settings and small sample sizes. We present an automated, large-scale framework for 3D pose estimation of professional tennis serves using publicly available broadcast video. The pipeline integrates RTMDet for player detection, RTMPose for 2D keypoint estimation, and MotionBERT for monocular 3D lifting, alongside a Dynamic Time Warping (DTW) action recognition module for serve segmentation. Critically, each extracted 3D pose sequence is linked to rich contextual metadata, including serve speed, placement, and match state, derived from automated scoreboard reading and official records, producing a unified biomechanical and performance dataset that does not exist in any current sports pose corpus. Applied to 5,966 serves from 109 professional players at the 2024 US Open, the framework reveals distinct biomechanical archetypes and gender-based divergences in kinetic strategies. A Random Forest classifier achieves 99.2% accuracy in player identification and 97.3% accuracy in gender classification from joint-angle trajectories alone, demonstrating that markerless motion capture from monocular broadcast footage captures individualized "kinematic fingerprints" at scale. These results establish a scalable foundation for vision-based biomechanical analysis that bridges the gap between laboratory precision and competitive authenticity. Our code and dataset are publicly available at https://github.com/jasnwag/tennis_serve_dataset.

1. Introduction

Optical motion capture has long been the gold standard for biomechanical research in sports science, yet its requirement for controlled laboratory conditions, expensive hardware, and manual marker placement makes studying elite athletes in real competition settings impractical [2, 8]. As a result, most biomechanical insights derive from small, controlled samples that fail to capture the variability of professional performance.

Recent advances in human pose estimation (HPE) enable markerless motion capture from ordinary video [6, 9, 12], but existing frameworks have been primarily evaluated on general-purpose datasets (*e.g.* Human3.6M, COCO) that lack domain-specific relevance to high-speed sports motions. Furthermore, large-scale biomechanical datasets for professional athletes remain virtually nonexistent, and the few sports-specific pose datasets that do exist [3, 10, 11] are limited in scale or focus on non-tennis domains. Crucially, none of these datasets pair pose sequences with the contextual performance metadata (serve speed, ball placement, match score) that is essential for connecting kinematics to competitive outcomes.

To address these challenges, there is a pressing need for a scalable, automated framework capable of extracting, analyzing, and annotating biomechanically meaningful motion data from publicly available sports broadcasts. Such a framework must integrate multiple stages of computer vision including object detection, pose estimation, action recognition, and annotation alignment, while maintaining sufficient accuracy and throughput for large-scale analysis.

This study responds to that need by developing an automated pipeline for extracting and analyzing 3D tennis serve biomechanics from broadcast video at scale. Our contributions are:

1. An automated vision pipeline, from player detection through 3D pose estimation and serve segmentation, applied to 5,966 professional serves, substantially larger than any prior sports biomechanics dataset derived from

broadcast footage.

2. Evidence that monocular markerless motion capture captures individualized "kinematic fingerprints" with near-perfect player identification accuracy (99.2%), demonstrating the fidelity of vision-based skeletal tracking for elite sports analysis.
3. An open, large-scale dataset in which every 3D pose sequence is paired with automatically extracted performance metadata (serve speed, ball placement, point outcome, and match state), enabling, for the first time, integrated analysis of elite kinematics and competitive context at scale.

2. Methods

2.1. Data Source

Video data were obtained from publicly available broadcast footage of the 2024 US Open, focusing exclusively on serve sequences. Each match was sampled at 60 fps and processed through the proposed pipeline.

The framework comprises four sequential vision stages, illustrated in Figure 1, followed by annotation integration, biomechanical post-processing, and data validation.

2.2. Object Detection and Player Tracking

A real-time object detection stage identifies and isolates players from the broadcast feed. RTMDet [5], a one-stage convolutional network optimized for speed and accuracy, leverages a feature pyramid network and decoupled detection head to predict bounding boxes and class scores concurrently, achieving $>50\%$ Average Precision on the COCO "person" class.

Detected bounding boxes are linked across frames by the SORT algorithm [1], which combines a Kalman filter for motion prediction with Hungarian-algorithm data association to maintain player identity with minimal identity switches.

2.3. 2D Pose Estimation

Following detection and tracking, each cropped player frame is processed by RTMPose [4], a top-down pose estimator implemented through the MMPose framework. RTMPose employs a deep convolutional backbone with multi-scale feature fusion and an attention-enhanced keypoint detection head, providing resilience to occlusions and camera perspective changes. Each frame produces 17 anatomical keypoints (shoulders, elbows, wrists, hips, knees, ankles, plus head landmarks) linked into skeletal representations.

2.4. 3D Pose Lifting

2D keypoints are lifted to 3D coordinates using MotionBERT [12], which combines spatial attention with transformer-based temporal modeling to infer depth. The model enforces geometric consistency and joint connectivity constraints, enabling accurate 3D reconstruction from monocular inputs. MotionBERT achieves $\sim$35.4 mm MPJPE on Human3.6M. The resulting 3D pose sequences form the basis for all biomechanical analyses described in subsequent sections.

2.5. Serve Segmentation

Serves are automatically identified using Dynamic Time Warping (DTW) [7]. A library of $\sim$200 manually labeled serve instances defines a canonical motion template. For each candidate sequence, DTW aligns joint-angle trajectories to the template and a normalized distance metric is computed. A cross-validated distance threshold classifies positive serve instances, achieving a balanced trade-off between sensitivity and specificity to distinguish serves from smashes and between-point movements.

2.6. Annotation Integration

Each identified serve was cross-referenced with match metadata (serve speed, placement, point outcome) extracted from on-screen scoreboards via a visual-language model. Player-specific annotations (gender, height) were obtained from ATP/WTA records. The resulting dataset contains, for every serve: (1) 3D joint trajectories, (2) serve type and location, (3) serve speed, and (4) match context, enabling integrated analysis of biomechanical performance metrics and tactical decision-making.

2.7. Biomechanical Post-Processing

Joint angles for eight joints (bilateral shoulder, elbow, hip, and knee) were computed via the cosine rule applied to limb-segment vectors defined by MotionBERT keypoints. Angular velocity and acceleration were derived from smoothed joint trajectories to characterize the kinetic chain during each serve phase (ball toss, racket acceleration, impact). Variable-length serve trajectories were temporally normalized to 50 equidistant time points using linear interpolation, yielding a 400-dimensional feature vector per serve, standardized to zero mean and unit variance.

2.8. Data Validation

Data validation was conducted through a dedicated GUI in which reviewers visually verified motion capture quality and labels. Each capture was assigned a quality label on a scale of 1–3, where 3 indicates no tracking errors and 1 indicates substantial tracking artifacts. Additional tasks included verification of pipeline-generated annotations such as scorebug readings and first/second serve classification, with reviewers noting and manually correcting incorrect labels.

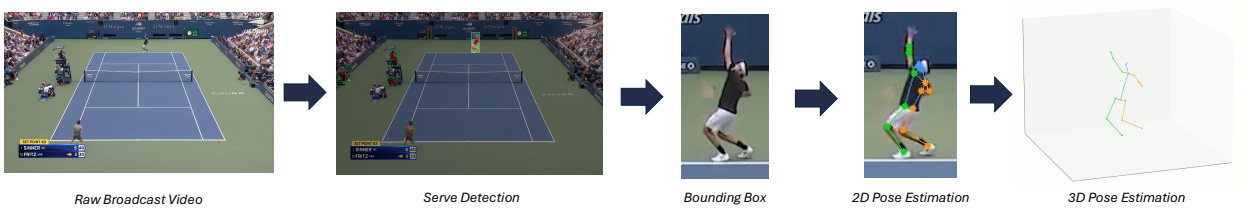

Figure 1. Flowchart showing the pipeline to obtain the 3D joint trajectories of a tennis serve.

Table 1. Classification performance using joint-angle trajectories. Accuracy and macro F1 are reported alongside the most-frequent baseline.

Target	RF Accuracy	Macro F1	Baseline	Lift
Gender	97.3%	0.972	55.0%	+42.3%
Player ID (top 14)	99.2%	0.992	10.6%	+88.6%
Serve quality	84.0%	0.757	65.0%	+19.1%

3. Results

3.1. Dataset Overview

The finalized dataset comprises 5,966 serves from 109 players (55 male, 54 female) across 113 matches. Male players contributed 3,279 serves; female players contributed 2,687. First serves accounted for 70.5% of observations ($n = 4{,}202$). Mean serve speed was 163.6 ± 24.0 km/h (range: 98–218 km/h), with male players averaging 174.1 ± 22.6 km/h versus 150.7 ± 18.8 km/h for female players (Cohen's $d = 1.12$). Serve speed was positively correlated with player height ($r = 0.490$, $p < 0.001$).

3.2. Biomechanical Joint Angle Profiles

Extracted joint-angle profiles were consistent with the kinetic chain model of the elite serve. The dominant shoulder reached a mean angle of $112.0 \pm 17.1°$ with peak angular velocities of 47.3 ± 15.8 deg/frame. Rapid elbow extension immediately preceded ball contact (peak velocity: 47.8 ± 13.6 deg/frame). Lower-extremity joints followed a flexion–extension sequence characteristic of the loading and drive phases, with the right knee exhibiting the highest mean angular acceleration (0.148 ± 0.391 deg/frame^2), quantifying the explosive leg drive essential for velocity generation.

3.3. Classification and Prediction

A Random Forest (RF) classifier (200 trees) was trained on the 400-dimensional joint-angle feature vectors. Performance was benchmarked against a most-frequent dummy classifier to determine the lift provided by kinematic features. Results are summarized in Table 1.

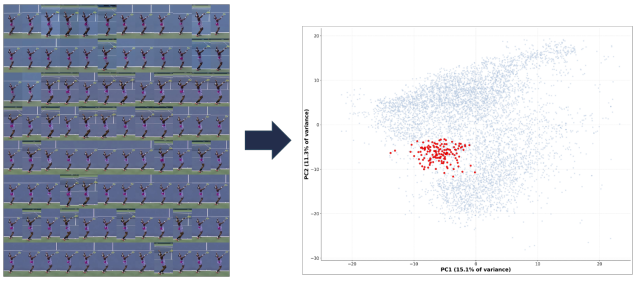

Figure 2. All of a player's serves clustered together in the first two principal components.

3.3.1. Gender Classification

The RF model achieved 97.3% accuracy (macro F1 = 0.972) in distinguishing player gender based solely on joint-angle trajectories, representing a lift of 42.3 percentage points over baseline. This high degree of separability indicates that gender-specific biomechanical strategies are deeply encoded in the spatiotemporal structure of the motion, extending beyond simple speed differentials to distinct postural and timing signatures.

3.3.2. Player Identification

Among the 14 players with $\geq$100 serves, the classifier reached 99.2% accuracy (macro F1 = 0.992) against a 10.6% chance baseline, confirming the existence of individualized kinematic fingerprints captured by the markerless pipeline. Figure 2 visualizes this by projecting all serves from one player onto the first two principal components of the full serve dataset; the distinct cluster supports the notion that elite athletes maintain highly consistent, idiosyncratic motor patterns.

3.3.3. Serve Quality

Serve quality as described in section 2.8 was classified at 84.0% accuracy (macro F1 = 0.757), providing a +19.1 percentage point lift over the baseline. High-quality serves were associated with more synchronized kinematic patterns, particularly regarding the timing of peak shoulder and elbow velocities.

3.3.4. Speed Prediction

To evaluate the predictive relationship between gross kinematics and performance output, an RF regressor was trained to predict continuous serve speed (km/h). The model achieved $R^2 = 0.253$, MAE = 17.2 km/h, and RMSE = 20.8 km/h. Compared to the mean-prediction baseline ($R^2 \approx 0.000$; MAE = 20.0 km/h), the RF regressor captured a meaningful predictive signal. While approximately 25% of speed variance is explained by the primary joint angles, the remaining variance is likely attributable to distal factors not captured in the 8-joint model, such as wrist pronation, racket-head lag, and contact-point dynamics.

4. Discussion

4.1. Kinematic Fingerprints

The most striking finding is the near-perfect accuracy (99.2%) in player identification based solely on joint-angle trajectories. This suggests that elite tennis serves are not merely optimized toward a singular biomechanical ideal, but are instead characterized by highly consistent, idiosyncratic motor patterns. Despite the technical convergence required to compete at the professional level, the automated pipeline captured sufficient individual variance in timing, postural alignment, and segment coordination to distinguish between players with near-perfect accuracy. The 400-dimensional joint representation effectively captures the intersection of each player's unique body proportions and their learned motor programs. From a sports science perspective, this supports the principle of motor abundance, or the ability to achieve the same performance outcome through different movement configurations, as a defining feature of elite professional athletes.

4.2. Gender Differences

The high gender classification accuracy (97.3%) underscores profound differences in kinetic chain strategies employed by male and female athletes. While the speed differential (Cohen's $d = 1.12$) is a well-documented factor, the classification accuracy suggests that these differences are structural rather than merely scalar. The gender-specific signatures likely reflect differences in the transfer of momentum through the kinetic chain, potentially driven by physiological variations in shoulder mobility, pelvic rotation velocities, and trunk stiffness. The fact that the model distinguished gender even when trajectories were temporally normalized suggests that the sequencing of joint peaks, rather than motion magnitude alone, is a key differentiator. These findings indicate that coaching and injury-prevention models for professional tennis should be gender-stratified to remain biomechanically relevant.

4.3. Bridging the Laboratory–Competition Gap

This study demonstrates that markerless motion capture, powered by architectures like RTMPose and MotionBERT, can bridge the laboratory–competition gap by achieving sufficient precision for biomechanical analysis using monocular broadcast footage.

The moderate speed prediction ($R^2 = 0.253$) highlights a current frontier: distal segment dynamics (wrist pronation, racket-head lag) undergo extremely high angular velocities that are difficult to capture at 60 fps broadcast frame rates. Future iterations should incorporate blur-resistant tracking or higher-frequency sampling.

4.4. Scalability

The fully automated pipeline enables longitudinal tracking of athletic performance at unprecedented scale. By cross-referencing 3D kinematics with match metadata (speed, placement, score), the framework transforms raw broadcast video into a multimodal repository for performance analysis, injury-risk assessment, and individualized coaching, all using consumer-grade compute and publicly available footage.

4.5. Limitations

Monocular 3D lifting introduces depth ambiguities during extreme rotations; multi-camera broadcast feeds would improve reconstruction. The current 8-joint angle model omits wrist and racket dynamics critical for impact-phase analysis. Validation against marker-based ground truth in tennis-specific settings remains an important next step. While this work was done offline, the further optimization could lead to real time performance. Future research should also expand this framework to other high-dynamic sports domains where occlusions are more frequent and actions are more continuous.

5. Conclusion

We have presented a scalable, fully automated framework for 3D biomechanical analysis of professional tennis serves from broadcast video. By processing over 5,000 serves from an elite cohort, we demonstrated that monocular markerless pose estimation captures kinematic data sufficient to identify individual athletes with near-perfect accuracy and distinguish gender-specific motion strategies. The emergence of kinematic fingerprints suggests that elite technique is characterized by idiosyncratic motor patterns rather than a uniform technical standard, opening new avenues for longitudinal performance tracking, injury-risk assessment, and individualized coaching at scale.

References

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing*, pages 3464–3468, 2016. 2

[2] Steffi L Colyer, Murray Evans, Darren P Mayberry, and Andrew L Sherwood. A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports Medicine – Open*, 4(1):24, 2018. 1

[3] Christian Kjær Ingwersen et al. SportsPose – a dynamic 3D sports pose dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5765–5774, 2023. 1

[4] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. RTMPose: Real-time multi-person pose estimation based on MMPose. *arXiv preprint arXiv:2303.07399*, 2024. 2

[5] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. RTMDet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022. 2

[6] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 1

[7] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. 2

[8] Sofia Scataglini et al. Unlocking the potential of video-based markerless motion analysis. *Journal of Sports Sciences*, 42 (5):401–415, 2024. 1

[9] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 1

[10] Tomohiro Suzuki, Ryota Tanaka, Calvin Yeung, and Keisuke Fujii. AthleticsPose: Authentic sports motion dataset on athletic field and evaluation of monocular 3D pose estimation ability. *arXiv preprint arXiv:2507.12905*, 2025. 1

[11] Callum Yeung et al. AthletePose3D: A benchmark dataset for 3D human pose estimation and kinematic validation in athletic movements. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025. 1

[12] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. MotionBERT: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023. 1, 2

A. Appendix

A.1. Pipeline Implementation Details

RTMDet was deployed with default COCO-trained weights using the MMDetection framework. The feature pyramid network and decoupled detection head predict bounding boxes and class scores concurrently, providing robust person detection across variable broadcast viewing conditions.
RTMPose was implemented through the MMPose framework using a deep convolutional backbone with multi-scale feature fusion. Each frame produced 17 anatomical keypoints (shoulders, elbows, wrists, hips, knees, ankles, plus head landmarks) linked into skeletal representations.
MotionBERT was fine-tuned on selected tennis broadcast clips to improve domain-specific generalization. The model enforces geometric consistency and joint connectivity constraints, enabling accurate 3D reconstruction from monocular inputs.
DTW-based serve segmentation used a library of $\sim$200 manually labeled instances. A normalized distance metric and cross-validated threshold achieved a balanced tradeoff between sensitivity and specificity, distinguishing serves from returns and between-point movements.

A.2. Annotation Pipeline

Scoreboard images were extracted from video frames and processed using a visual-language model to interpret onscreen information. Through prompt-engineered interactions, structured match context was extracted, validated, and aligned with serve sequences via time synchronization and scoring patterns. The resulting annotations include: (1) 3D joint trajectories, (2) serve type and location, (3) serve speed, and (4) match context (score, set, game).

A.3. Additional Dataset Statistics

The distribution of serves per player (mean 54.7 ± 48.0) reflects natural variance in match duration and tournament progression. Data validation was conducted through a dedicated GUI in which reviewers assigned quality labels (1–3 scale) and verified pipeline-generated annotations including scorebug readings and first/second serve classification. Quality label 3 indicates no tracking errors; label 1 indicates substantial tracking artifacts.

A.4. Joint Angle Computation

Joint angles for eight joints (bilateral shoulder, elbow, hip, and knee) were computed via the cosine rule applied to limb-segment vectors defined by MotionBERT keypoints. Angular velocity and acceleration were derived from smoothed trajectories to characterize the kinetic chain during each serve phase (ball toss, racket acceleration, impact). Each serve was temporally normalized to 50 equidistant time points using linear interpolation, enabling biomechanical comparison across players at equivalent motion phases.